

# **Phylogenomic analysis of energy converting enzymes**

**Dissertation**

Presented to the department of Biology/Chemistry, University of  
Osnabrück in partial fulfillment of the requirements for the degree of

*"Doctor Rerum Naturalium"*

**Daria Dibrova**

Osnabrück  
April 2013

This page is intentionally left blank

This work was performed at the University of Osnabrück.

Supervisor:  
PD Dr. A. Mulkidjanian

This page is intentionally left blank

# Table of contents

<b>Table of contents</b> .....	<b>5</b>
<b>1. Introduction: Survey of biological energy conversion</b> .....	<b>9</b>
1.1. <i>Current views on the evolution of energy conversion in biological systems</i> .....	9
1.1.1. Heterotrophic scenario on the origin of life. RNA world .....	10
1.1.2. Autotrophy hypothesis .....	16
1.1.3. First phylogenetic analyses based on rRNA sequencing and the concept of the Last Universal Cellular Ancestor (LUCA) .....	18
1.2. <i>Basics of bioenergetics. ATP as a universal energy carrier between biochemical         pathways</i> .....	25
1.2.1. Properties of ATP. High phosphate transfer potential .....	26
1.2.2. Mechanisms of enzymatic hydrolysis of ATP and GTP .....	29
1.3. <i>Evolution of ATPases and GTPases</i> .....	31
1.3.1. Evolution of the P-loop GTPases and related ATPases .....	31
1.3.2. Evolution of the P-loop kinases and related proteins.....	33
1.3.3. Evolution of the AAA+ ATPases .....	33
1.4. <i>ATP synthesis and <math>\Delta\tilde{\mu}_f</math></i> .....	34
1.4.1. Substrate-level phosphorylation.....	34
1.4.2. Oxidative phosphorylation and photophosphorylation: rotary membrane ATP synthases .....	35
1.4.2.1. General structure and mechanism of rotary membrane ATPases .....	35
1.4.2.2. Diversity of the <i>c</i> -oligomers. Sodium- and proton-dependent rotary ATPases .....	43
1.4.2.3. Hypothesis on the origin of rotary membrane ATPases from RNA translocases .....	45
1.4.2.4. Evolutionary primacy of sodium-dependent bioenergetics .....	47
1.5. <i>Review of energy-converting redox complexes</i> .....	49
1.5.1. Chemical structures of most common redox cofactors.....	50
1.5.2. Review of the variants of electron transfer chains.....	52
1.5.3. Generation of $\Delta\tilde{\mu}_{H^+}$ by respiratory chain of mitochondria.....	55
1.5.3.1. NADH:quinone oxidoreductase.....	58
1.5.3.2. Succinate::quinone oxidoreductase.....	62
1.5.3.3. Quinol:cytochrome <i>c</i> oxidoreductase (cytochrome <i>bc</i> complex) .....	65
1.5.3.4. Cytochrome <i>c</i> -oxidase .....	71
1.5.4. Photosynthesis.....	74
1.5.5. Role of reactive oxygen species and the cytochrome <i>bc</i> <sub>1</sub> -complex in apoptosis..	77
1.6. <i>Aims of the thesis</i> .....	85
<b>2. Methods</b> .....	<b>86</b>
2.1. <i>Principles of phylogenomic analysis</i> .....	86
2.1.1. Sequencing of full genomes.....	86
2.2. <i>Protein comparison via sequence alignment</i> .....	87
2.2.1. Global and local pairwise alignment.....	88
2.2.2. Position-independent amino acid substitution matrices.....	89
2.2.3. Global multiple alignment. Similarity searches.....	89

2.3. Critical analysis of multiple alignment from the biological point of view .....	95
2.4. The problem of "the same" genes and the minimal gene set concept .....	97
2.5. Phylogenomic analysis and its difference from phylogenetic analysis.....	99
2.5.1. Phylogenetic analysis. Phylogenetic tree construction methods.....	99
2.5.2. Phylogenomic analysis.....	102
2.6. Summary of classical bioinformatical algorithms, software and databases used in this study .....	105
2.7. Application of the described bioinformatical methods to the particular proteins.....	107
2.7.1. Phylogenomic analysis of GTPases and ATPases .....	108
2.7.1.1. EF-Tu and other translation factors .....	108
2.7.1.2. Recombination proteins RecA/RadA.....	109
2.7.1.3. Molecular chaperone GroEL.....	110
2.7.1.5. Membrane pyrophosphatases.....	112
2.7.2. Phylogenomic analysis of the N-ATPases .....	113
2.7.2.1. Search for cyanobacterial ATP synthases subunits .....	113
2.7.2.2. Search for the subunits of prokaryotic ATP synthases .....	113
2.7.2.3. Concatenated phylogenetic tree construction .....	114
2.7.3. Phylogenomic analysis of the cytochrome <i>b</i> .....	114
2.7.4. Multiple alignments of other components of the cytochrome <i>bc</i> <sub>1</sub> complex .....	115
2.7.4.1. Construction of multiple alignment for the cytochrome <i>c</i> <sub>1</sub> .....	115
2.7.4.2. Construction of multiple alignment for the subunit 8 (9.5 kDa subunit) of the eukaryotic <i>bc</i> complex.....	116
2.7.4.3. Conservation of positively charged residues in the sequences of cytochromes <i>c</i> <sub>2</sub> .....	116
<b>3. Evolution of the mechanisms of phosphodiester bonds hydrolysis: from K<sup>+</sup> ions to the lysine or arginine "fingers" .....</b>	<b>117</b>
3.1. Inorganic ion requirements of ubiquitous cellular systems.....	117
3.2. Potassium dependence of the enzymes catalyzing reactions of phosphate group transfer.....	121
3.2.1. Example #1: P-loop GTPases .....	122
3.2.2. Example #2: RadA and RecA.....	126
3.2.3. Example #3: chaperonine GroEL.....	131
3.2.4. Example #4: BCK (branched-chain $\alpha$ -ketoacid dehydrogenase kinase) .....	135
3.2.5. Example #5: membrane pyrophosphatases.....	138
3.3. Discussion: Activation by K <sup>+</sup> ions could be evolutionarily older than the involvement of lysine or arginine "fingers" .....	141
<b>4. N-ATPases: distinct family of rotary membrane ATPases .....</b>	<b>145</b>
4.1. Unusual operon structure of N-ATPases. Additional genes and their possible functions.....	148
4.2. Phylogenomic analysis indicates that N-ATPases are a separate family of rotary membrane ATPases.....	151
4.3. Evidences for the spreading of N-ATPases via the lateral gene transfer.....	152
4.4. Discussion: specific features of N-ATPases as possible ancient features of rotary membrane ATPases.....	154
<b>5. Phylogenomic analysis of the cytochrome <i>bc</i> complex .....</b>	<b>159</b>
5.1. Analysis of the phylogenetic tree of the cytochromes <i>b</i> .....	159

5.2. Discussion: implications from the lateral transfer of the cytochrome <i>bc</i> complexes for their evolution .....	172
5.2.1. Discrepancies caused by the hypothesis on the presence of the cytochrome <i>bc</i> complexes in the LUCA.....	172
5.2.2. Fusion of the cytochrome <i>b</i> subunits is a more probable evolutionary scenario for cytochrome <i>bc</i> complexes than their fission.....	174
5.2.3. Possible scenario of the emergence of cytochrome <i>bc</i> complexes .....	177
5.2.4. Adaptation of the cytochrome <i>bc</i> complexes to oxygenation of the atmosphere .....	181
<b>6. Evolution of apoptosis as a strategy to diminish the oxygen-caused damage to consortia of cells .....</b>	<b>184</b>
6.1. Introduction: Apoptosis as a mechanism to kill a cell with a broken mitochondrial cytochrome <i>bc</i> <sub>1</sub> -complex .....	184
6.2. Ligands for cardiolipin, a possible early detector of the ROS production, in the components of the cytochrome <i>bc</i> <sub>1</sub> complex .....	185
6.3. Evolution of the interaction between cytochrome <i>c</i> and the components of the apoptosome .....	191
6.3. Discussion: The further evolution of components of cytochrome <i>bc</i> <sub>1</sub> complex in aerobic environment could have been driven by the optimization of the apoptotic cascade.....	193
<b>7. Outlook: Evolution of biological energy conversion as inferred from phylogenomic analysis of energy-converting enzymes .....</b>	<b>196</b>
7.1. K <sup>+</sup> ions as catalysts of the primordial phosphate transfer reactions.....	196
7.2. Sodium-dependent membrane bioenergetics: The ancestral form of rotary ATPases was a sodium-dependent enzyme .....	199
7.3. Separation of bacteria and archaea and the transition to the oxygenated atmosphere.....	200
<b>8. Conclusions.....</b>	<b>203</b>
<b>9. Summary.....</b>	<b>205</b>
<b>10. References .....</b>	<b>206</b>
<b>11. Abbreviations .....</b>	<b>236</b>
<b>12. Supplementary material.....</b>	<b>237</b>
<b>13. Acknowledgements .....</b>	<b>252</b>
<b>14. Publications .....</b>	<b>253</b>
<b>15. Lebenslauf.....</b>	<b>257</b>
<b>16. Erklärung über die Eigenständigkeit.....</b>	<b>259</b>

This page is intentionally left blank



# 1. Introduction: Survey of biological energy conversion

## 1.1. Current views on the evolution of energy conversion in biological systems

Currently there is no consensus on the earliest steps of evolution of life. While some authors argue that the first living forms, being *heterotrophs*, were fully dependent on abiotically formed organic compounds (Lazcano and Miller, 1999; Miller and Cleaves, 2006; Oparin, 1924), other scholars put forward various hypotheses of primordial *autotrophy*, where already the simplest life forms are assumed to be capable of harvesting natural energy flows such as redox gradients (Wächtershäuser, 1990; Wächtershäuser, 1992; Wächtershäuser, 2007) or pH gradients (Martin *et al.*, 2008; Martin and Russell, 2003; Martin and Russell, 2007) and directly channelling the harvested energy into (bio)synthetic reactions.

Life can exist only if supported by energy flow since it reduces entropy by structuring simple chemical compounds into sophisticated polymers. Ludwig Boltzmann wrote: ... *Der allgemeine Daseinskampf der Lebewesen ist daher nicht ein Kampf um die Grundstoffe – die Grundstoffe aller Organismen sind in Luft, Wasser und Erdboden im Überflusse vorhanden – auch nicht um Energie, welche in Form von Wärme leider unverwandelbar in jedem Körper reichlich vorhanden ist, sondern ein Kampf um die Entropie, welche durch den Übergang der Energie von der heißen Sonne zur kalten Erde disponibel wird. Diesen Übergang möglichst auszunutzen, breiten die Pflanzen die unermessliche Fläche ihrer Blätter aus und zwingen die Sonnenenergie in noch unerforschter Weise, ehe sie auf das Temperaturniveau der Erdoberfläche herabsinkt, chemische Synthesen auszuführen, von denen man in unseren Laboratorien noch keine Ahnung hat. Die Produkte dieser chemischen Küche bilden das Kampfobjekt für die Tierwelt.* (cited from (Boltzmann, 1979)).

As surveyed below, different hypotheses on origin of life consider different energy sources for the first life forms. Therefore, the analysis of primeval energetics by means of bioinformatics, comparative genomics, and evolutionary biophysics might shed light on the nature of the first life forms and their habitats.

### 1.1.1. Heterotrophic scenario on the origin of life. RNA world

In the framework of hypotheses of Oparin (Oparin, 1924) and Haldane (Haldane, 1929), the first living forms could evolve from a mixture of simple organic and inorganic compounds in the ocean ("*primordial soup*"). The organic compounds were suggested to be formed from atmospheric gases ( $H_2$ ,  $NH_3$ ,  $CH_4$ , and water vapor) upon illumination of the atmosphere by the solar UV light and lightning strikes. Later, by modeling primordial syntheses in a reduced gas mixture hit by electric discharges, Miller succeeded in obtaining organic molecules, including amino acids (Miller, 1953). This experiment has shown the feasibility of obtaining organic molecules under supposedly primordial conditions.

The hypotheses of Oparin and Haldane suggest that the first organisms appeared in organics-rich media in the absence of oxygen and were performing some kind of fermentation of available organic molecules (Oparin, 1957; Wald, 1964).

However, fermentation in modern organisms is anything but simple. During fermentation energy is primarily stored in the form of chemical bonds in *cofactors*, ATP and NADH, which are produced from their dephosphorylated (ADP) or oxidized ( $NAD^+$ ) forms by proteinaceous enzymes. The enzymes, in turn, could not just "appear": a mixture of amino acids, even if they were randomly forming stable peptides (which is unlikely), has a near-zero chance of forming a protein dedicated to a specific catalytic function. In modern organisms, proteins are synthesized at ribosomes in a sophisticated, "coded" mode, by using the information stored in nucleic acids, see (Wolf and Koonin, 2007) for a review. Hence, there should have been an intermediate stage between the non-living matter and the "fermentation" step that was postulated by Oparin and Haldane. According to current views, the *RNA world* could represent this intermediate step.

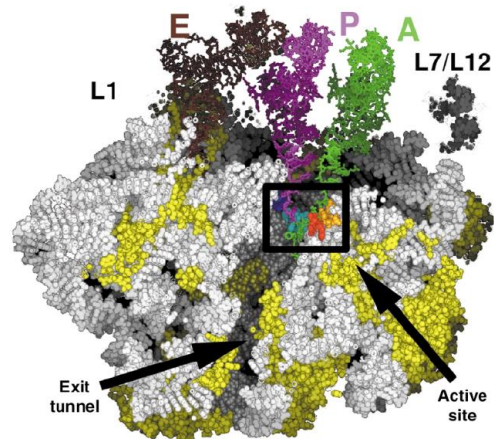
**RNA world hypothesis.** As early as in 1957, A.N. Belozersky has suggested, after discovering the non-coding RNA, that RNA molecules could be the first biological polymers: "There is no doubt that nucleic acids played an important role in the evolution of the organic world. Yet both RNA and DNA could hardly arise simultaneously in the early evolution of life. It rather seems that ribonucleotides, and then RNA, originated first. DNA came into existence far more recently, as the protoplasm became more differentiated and its functions

grew in complexity" (quoted from (Belozersky, 1957; Belozersky, 1959)). This idea gained in popularity after the first catalytic RNAs were discovered. The first two examples of RNAs performing enzymatic functions (*ribozymes*) were the self-splicing ribosomal rRNA from protists (Cech and Brehm, 1981; Grabowski *et al.*, 1981) and RNase P, the ribozyme responsible for the maturation of tRNA. In *E. coli*, the RNase P complex was shown to consist of both RNA and protein, but the RNA component was responsible for catalysis (Stark *et al.*, 1978).

Taken together, these ideas led to the hypothesis of the "RNA world", an early stage in the evolution of life, when RNA could serve both as an information carrier and a catalyst of chemical reactions, including self-replication (Gilbert, 1986). Molecules of RNA, as proteins, are capable of forming complex and ordered 3D structures (Spirin, 2005). As in proteins, hydrogen bonds play an important structural role in RNA: they can form between self-complementary regions yielding *stems* (Spirin, 1960). The concept of the RNA world has also got support from comparative virology (Koonin *et al.*, 2006). Modern cellular organisms store genetic information exclusively in the form of DNA but the very existence of RNA viruses proves that information can be transmitted between generations in the form of RNA. Therefore, viruses are now considered to be descendants from the primordial genetic pool rather than result of the degradation of full-fledge cellular life forms (Koonin *et al.*, 2006).

In the framework of the RNA world hypothesis, the emergence of protein synthesis could be reconstructed by analysis of those molecular machines that accomplish it nowadays. Modern ribosomes that perform protein synthesis (*translation*) are ubiquitous RNA-protein complexes formed of two subunits: the large subunit and the small one. The former harbors the peptidyl transferase center where amino acids are being attached to the growing peptide chain (**Figure 1.1.1**). As both the peptide and the incoming amino acid are connected to the tRNA molecules, the specific sites for tRNA binding are located in the large subunit. The A-site is occupied by a tRNA molecule with a new amino acid, the P-site binds a tRNA molecule with the peptide chain, and the E-site interacts with the deacylated tRNA molecule before its removal out of the ribosome (Steitz and Moore, 2003). The small subunit binds and "deciphers" mRNA, which means that it interacts with tRNA molecules and determines which aminoacyl-tRNA should be bound based on the current codon in mRNA. For the progress of translation, two protein *elongation factors* (EF) are required. The EF-Tu factor

delivers aminoacyl-tRNA to the ribosome and leaves only if this tRNA molecule fits the current codon; the respective change in protein conformation is induced by GTP hydrolysis. The EF-G factor takes part in the movement of the tRNA and mRNA molecules after formation of a new peptide bond. The EF-G factor is also GTP-dependent.



**Figure 1.1.1. Structure of the surroundings of the ribosome active site.**

The active site is boxed. The large ribosome subunit is shown with the three tRNA molecules added in positions which they should occupy in the A, P and E-sites of the ribosome, respectively. The large ribosome subunit is cut along the plane of the polypeptide exit tunnel. The rRNA molecules are shown in white, proteins are shown in yellow (taken from (Steitz and Moore, 2003)).

The crystal structure of the large ribosomal subunit has shown that the key activities of the ribosome, namely the decoding of mRNA and the formation of peptide bond, are performed by rRNA molecules rather than by protein subunits (Steitz and Moore, 2003). This discovery solved "the chicken or the egg" dilemma as concerning the relation between proteins and nucleic acids. Thus, production of simple peptides is possible by RNA molecules themselves, and this finding gives a strong support to the RNA world hypothesis.

Structural studies of the ribosome have shown that RNA molecules can form stable, highly complicated and asymmetrical structures. However, the peptidyl transferase center appears to be symmetrical. The tRNA molecule with the peptide and the tRNA molecule that is acylated with an amino acid are positioned on each side of this center in an optimal conformation for forming a peptide bond (Gregory and Dahlberg, 2004; Youngman *et al.*, 2004). rRNA in this region has the same structural stem-elbow-stem motif, typical for various RNA molecules including ribozymes and tRNAs (see (Belousoff *et al.*, 2010; Davidovich *et al.*, 2009; Fox,

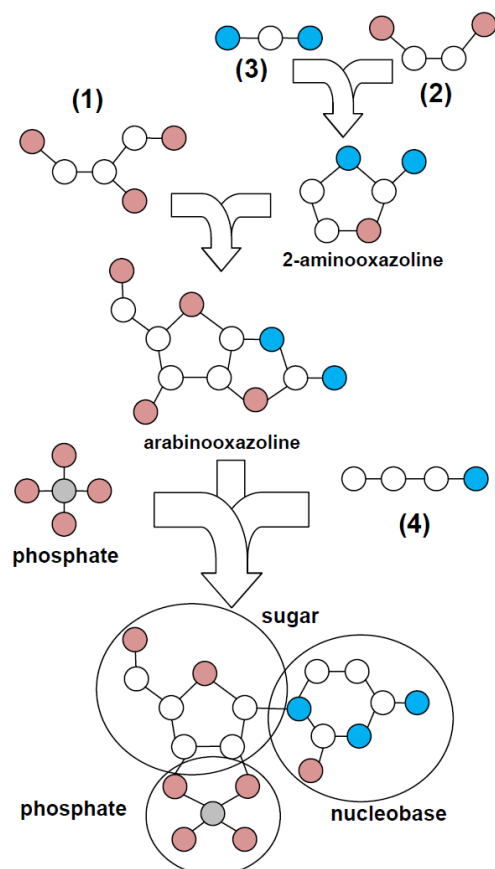
2010) and references therein). Artificially designed sequences of RNA resembling sequences observed in ribosomes were capable of forming stem-elbow-stem structural motif and of forming dimers (Davidovich *et al.*, 2009). Thus emergence of at least non-specific protein synthesis on the basis of such simple prototypes of peptidyl transferase center could be envisioned. The origin of the coded protein synthesis is, however, still enigmatic.

**Abiotic formation of the components of RNA.** The RNA world concept requires pre-existing monomers for RNA synthesis and a driving force for their polymerization. At the first glance, nucleobases that form nucleotides of RNA polymers are complex molecules. Still, nucleobases were shown to form spontaneously from formamide in response to heating (Saladino *et al.*, 2012a; Saladino *et al.*, 2012b) or UV-illumination (Barks *et al.*, 2010) of their solutions. Earlier it was shown that heating of a mixture of phosphate and nucleosides in the presence of formamide leads to the formation of nucleotides (Schoffstall, 1976).

The common property of native nucleobases, which discriminates them from other compounds of comparable complexity, is their exceptional photostability (Mulkiđjanian *et al.*, 2003; Serrano-Andrés and Merchán, 2009; Sobolewski and Domcke, 2006). It has been argued previously that nucleotides could accumulate at primordial earth due to this property, being photoselected by solar UV radiation – which should have been strong in the absence of ozone shield – from a plethora of abiotically (photo)synthesized organic compounds (Mulkiđjanian *et al.*, 2003).

Recently, the abiogenic synthesis of whole pyrimidine nucleotides from a mixture of simple compounds was reported (Powner *et al.*, 2009). The methodology of this work implied mixing of simple compounds, as shown in **Figure 1.1.2**, rather than pre-formed ribose, nucleobase and phosphate components. Furthermore, it has been shown that after a prolonged UV-illumination of the complex product mixture of ribonucleotides and diverse byproducts of nucleotide synthesis, only natural 2',3'-cyclic pyrimidine nucleotides remained in the solution as the most photostable of the molecules produced, whereas other products broke down and, apparently, reshuffled to yield 2',3'-cyclic nucleotides (Powner *et al.*, 2009). The 2',3'-cyclic ribonucleotides can polymerize into oligomers even in the absence of templates (Verlander *et al.*, 1973); the polymerization is driven by the cleavage of one of the two phosphoester bonds (transesterification). Hence, photostable cyclic nucleotides, which could

form abiotically at high concentrations of formamide and phosphate (Costanzo *et al.*, 2011; Costanzo *et al.*, 2007; Saladino *et al.*, 2012a), could serve as both monomers and the energy source for the abiotic formation of RNA replicators and ribozymes.



**Figure 1.1.2. Scheme of the reactions of synthesis of nucleotides as performed in** (Powner *et al.*, 2009).

Figure adapted from (Szostak and Ricardo, 2009). Abbreviations of compounds: (1) – glyceraldehyde, (2) – glycoaldehyde, (3) – cyanamide, (4) – cyanoacetylene.

The photostability of nucleobases increases upon their stacking and formation of Watson-Crick pairs (Sobolewski and Domcke, 2006). Polymers that contained complementary nucleobases could fold into photostable, double-stranded structures and also increase their photostability by pairing with other polymers (Mulkiđjanian *et al.*, 2003). As argued by Saladino *et al.* (Saladino *et al.*, 2006), at low water content and/or at high levels of simple amides the polymerization could be even thermodynamically favorable. In principle, such interactions could ultimately yield replicating entities.

Molecules of ATP are synthesized from ADP and  $P_i$  by sophisticated molecular machines – rotary membrane ATP synthases, which could have evolved only at a relatively late stage of evolution. However, the ubiquitous usage of ATP, as well as other nucleotides, in a plethora of cellular processes, from energy storage to the coding of genetic information, suggests a very early selection of nucleotides in evolution. It is noteworthy that the exceptional photostability of natural nucleotides is not directly related to their functions in modern organisms. To summarize, activated, energy-rich cyclic nucleotides, due to their exceptional photostability could selectively accumulate on primordial earth. The first replicator molecules from the ancient "RNA world" could have used photostable abiotically formed nucleoside phosphates, including nucleoside triphosphates, both as monomers for construction of new replicators and as an energy source.

**Possible sources of simple organic compounds.** As mentioned above, in the presence of inorganic phosphates, synthesis of nucleotides could be possible from simple carbon- and nitrogen-containing organic compounds, but the latter still needed to be formed in substantial amounts in abiotic reactions. Several abiotic processes could have supported the formation of organic molecules on the primordial earth.

One such process is called *hydrothermal alteration* and is known to occur in both continental and deep-sea hydrothermal systems (Sleep *et al.*, 2004; Taran *et al.*, 2010). Hydrothermal alteration occurs when iron-containing rocks interact with water at temperatures of approximately 300°C, which is typical for geothermal systems. Under these conditions, part of the  $Fe^{2+}$  in the rock is oxidized to  $Fe^{3+}$ , yielding magnetite ( $Fe_3O_4$ ). The electrons released in this reaction are accepted by protons of water yielding  $H_2$ ; in the presence of water-dissolved  $CO_2$ , diverse hydrocarbons are ultimately produced (Sleep *et al.*, 2004). Similar reactions could lead to the ammonia formation (Brandes *et al.*, 1998), which might account for the high ammonia content in the exhalations of geothermal systems (Bortnikova *et al.*, 2009).

In addition, organic compounds could have formed upon irradiation of photocatalytic minerals such as ZnS, MnS and  $TiO_2$ , by the solar UV-light (which was much stronger before the emergence of the ozone shield) (Guzman and Martin, 2009; Schoonen *et al.*,

2004). Among different studied photocatalysts, ZnS is the most effective one, being able to catalyze formation of formic acid from CO<sub>2</sub> with the yield of up to 80% (Henglein, 1984).

### 1.1.2. Autotrophy hypothesis

In the same year when Belozersky suggested that RNA molecules were the first biological polymers, a quite different view on the early evolution of life was introduced by Granick (Granick, 1957). The hypothesis of Granick implied that redox processes and photosynthesis could be involved already at the earliest steps of life evolution, but, perhaps, were performed by less complex structures. Iron-containing minerals were proposed to be highly important for such processes. With iron as an example, Granick suggested that the catalytic activity of a single metal ion could be increased hundred-fold upon creation of ferric hydroxides, and much more after its incorporation into heme and further incorporation of heme into cytochromes. The same property was proposed for other metal ions: wherever they are used in enzymes, they could have been used before the enzymes, but less efficiently. Discussing the possible origin of photosynthesis, Granick supposed that minerals containing a mixture of FeS and different iron oxides could have catalyzed water decomposition into H<sub>2</sub> and O<sub>2</sub>, producing a reductant and an oxidant as sources of chemical energy. The modern-type, (bacterio)chlorophyll-based photosynthesis is no longer considered to evolve very early (reviewed in (Mulkidjanian *et al.*, 2006), see also Section 1.5.4), so that the modern proponents of the autotrophy hypothesis suggest chemosynthesis as a means for utilizing natural energy fluxes in the first organisms, as discussed e.g. in refs. (Lane *et al.*, 2010; Lane and Martin, 2012; Nealson, 2003; Schoepp-Cothenet *et al.*, 2013; Wächtershäuser, 1988; Wächtershäuser, 1990).

One of the core processes in the framework of autotrophy hypothesis is the carbon fixation (creating carbon-carbon bonds) by the first living organisms. This issue was addressed by Wächtershäuser (Wächtershäuser, 1990). He suggested that a cycle similar to the reverse (reductive) Krebs cycle could have been performed on mineral surfaces with participation of FeS and other iron sulfides. The role of surface here is to bind the negatively charged intermediates and bring them together, as well as to drive the reactions by oxidation of FeS to



FeS<sub>2</sub> (pyrite). This reaction is indeed possible as its  $\Delta G$  is negative, but, as pointed out by Orgel, the negative  $\Delta G$  value still does not guarantee any reasonable yield (Orgel, 2008).

The most recent autotrophic scenario suggests that life emerged within porous structures formed outside of the alkaline hydrothermal vents at the ocean bed (Lane *et al.*, 2010; Martin and Russell, 2003). These vents could deliver simple organic compounds and hydrogen resulting from hydrothermal alteration process happening deeper inside the rock, as observed in the recently discovered Lost City hydrothermal fields (Kelley *et al.*, 2005; Von Damm, 2001). It is hypothesized that the first life forms fixed carbon dioxide, by using hydrogen, directly via acetyl-CoA (Wood-Ljungdahl) pathway, obtaining ATP at the same time.

Oxidation of hydrogen as a possible energy source is hypothesized to be coupled to the proton translocation across the membranes of porous hydrothermal precipitates – mineral bubbles. The FeS/FeS<sub>2</sub>-containing mineral membranes of such bubbles at the alkaline hydrothermal vents are suggested to functionally precede biological membranes in maintaining the proton potential (Lane *et al.*, 2010; Martin and Russell, 2003). The rotary membrane ATPase, in the framework of this hypothesis, could have been embedded into the walls of mineral bubbles.

This scenario is based on the fact that ATP and other nucleoside triphosphates are ubiquitous molecules which "transport" energy to be spent in numerous cellular processes. The synthesis of ATP is achieved in several ways, with the largest amounts produced by such ubiquitous enzymes as rotary ATP synthases. Chemiosmotic coupling mechanism, which can drive ATP synthesis or other cellular activities, is also widespread (see Section 1.5 for a detailed survey). Finally, redox reactions are used by almost all modern organisms to produce transmembrane potential. This could seem like a solid base for the idea that the rotary ATP synthase, the chemiosmotic coupling and redox chemistry (based on utilization of FeS clusters, quinones, and hemes) were among the first inventions of life and were likely present in the first living systems.

The weak point of the autotrophy hypotheses is that they do not offer plausible explanations for the appearance of biopolymer catalysts – either proteins or RNA molecules. Formation of simple organic compounds, as described above, can proceed abiotically both in geothermal and marine systems under anoxic conditions. However, these reactions do not provide a driving force for further synthesis of more complex biopolymers capable of Darwinian

evolution. Even if the hypothetical capability of the mineral bubbles to hold ion potential (Lane *et al.*, 2010; Lane and Martin, 2012) were proven, spontaneous emergence of an incredibly complex rotary ATP synthase to harvest this energy source does not seem likely.

### **1.1.3. First phylogenetic analyses based on rRNA sequencing and the concept of the Last Universal Cellular Ancestor (LUCA)**

The appearance of molecular data on sequences of biological polymers (DNA, RNA and proteins) allowed switching from sheer speculations about the early life to the reconstruction of the first life forms.

**The division between bacteria and archaea is evolutionarily the deepest.** On a shorter time scale, the reconstructions of ancestral forms were implicitly and explicitly performed starting from the works of Charles Darwin. Finding features that are common to all organisms in some group (characteristic features) and are sufficiently complex to originate in evolution only once could be used as a way of constructing the taxonomy of organisms. For example, the exclusive presence of nucleus in some organisms and its absence in other organisms was considered as a major distinguishing feature and allowed dividing the all life forms in the two groups: eukaryotes and prokaryotes (Stanier and Van Niel, 1962).

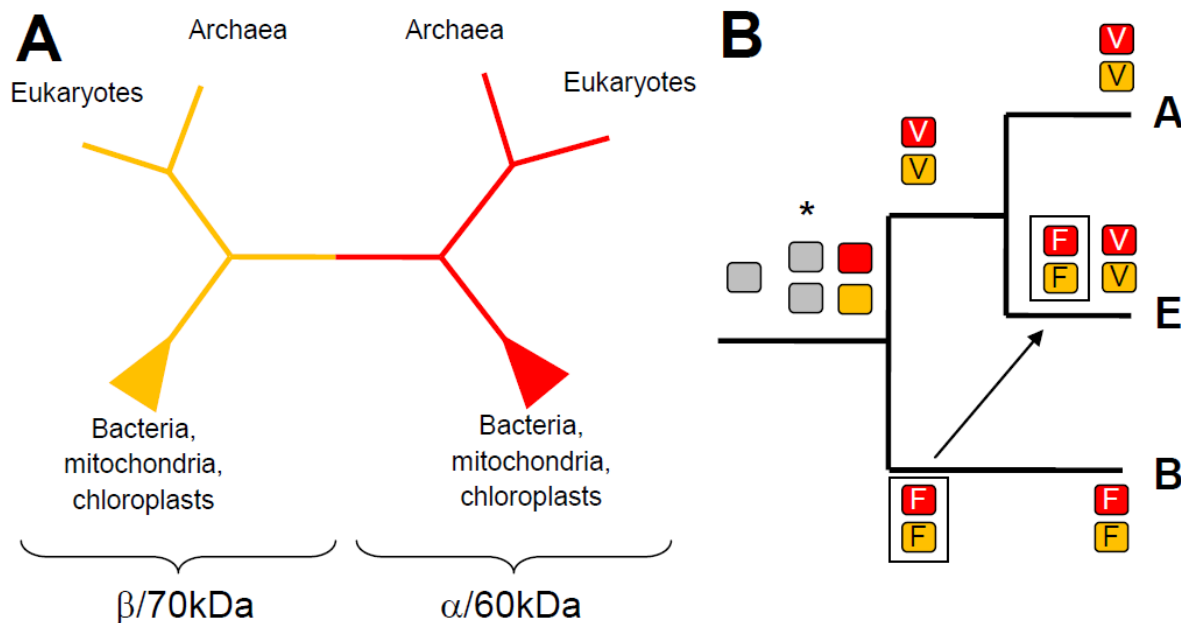
With the onset of sequencing of biopolymers, investigators could switch the focus from the "features" or "attributes" of biological polymers to their sequences (Zuckerandl and Pauling, 1965). Particularly important was the work of Woese and Fox (Woese and Fox, 1977), in which the authors have analyzed the sequences of the ubiquitous ribosomal 16S rRNA. They have digested rRNAs of different organisms by RNase, separated them by 2D electrophoresis and compared the oligonucleotide fingerprints belonging to different species. The similarity between the fingerprints allowed clear clustering of all species into three *domains*: bacteria, archaea and eukaryotes.

The same approach helped to clarify the origin of mitochondria and chloroplasts. Endosymbiotic origins of chloroplasts (Mereschkowsky, 1910) and mitochondria (Wallin, 1923) were proposed in the beginning of the 20th century, but the popularity came to these ideas after 1967 (Sagan, 1967), particularly, after the sequences of 16S rRNA from mitochondrial and chloroplast ribosomes were shown to group with prokaryotic 16S rRNA

sequences (Gray, 1989). Endosymbiosis is believed to have played an important role in the evolution of eukaryotes, although still much is unclear about their origin (see (Koonin, 2010) for a review). Specifically, the relation of eukaryotes to archaea and bacteria was addressed by Gogarten *et al.* upon the phylogenetic analysis of the subunits of rotary membrane ATP synthases (Gogarten *et al.*, 1989). These enzymes are ubiquitous in all organisms and all share an interesting feature: the hexamer that performs the hydrolysis/synthesis of ATP contains subunits of two different types. Both subunits share strong sequence similarity far beyond random and thus their genes are considered *homologs* as sharing a common ancestral sequence and *paralogs* as they arose from a gene duplication event. As modern organisms from all three domains have two types of subunits in their catalytical hexamers, this duplication is assumed to have occurred prior to the domain separation. The phylogenetic tree of the available sequences showed that  $\alpha$ - and  $\beta$ -subunits of the ATP synthase from *Sulfolobus* (an archaeon) grouped with the corresponding 60 kDa and 70 kDa subunits of eukaryotic vacuolar V-ATPase (**Figure 1.1.3A**).

The tree of  $\beta$ - and 70 kDa subunits was shown to be in correspondence with the 16S rRNA tree. Under the assumption that both types of subunits were present before the domain separation, this tree could be rooted between the  $\alpha$ - and  $\beta$ -subunits. Such rooting suggested that archaea with eukaryotes form a branch that is separate from the bacterial branch. The analysis of ubiquitous ribosomal proteins also suggests the grouping of archaea with eukaryotes (Iwabe *et al.*, 1989; Yutin *et al.*, 2012).

It is now generally accepted that deepest division of life is between archaea and bacteria, while eukaryotes, although related to archaea, were largely influenced by inter-domain lateral gene transfer (including endosymbiosis) after their separation and thus were shaped as a separate domain after the separation of bacteria and archaea.



**Figure 1.1.3. Separation of major domains.**

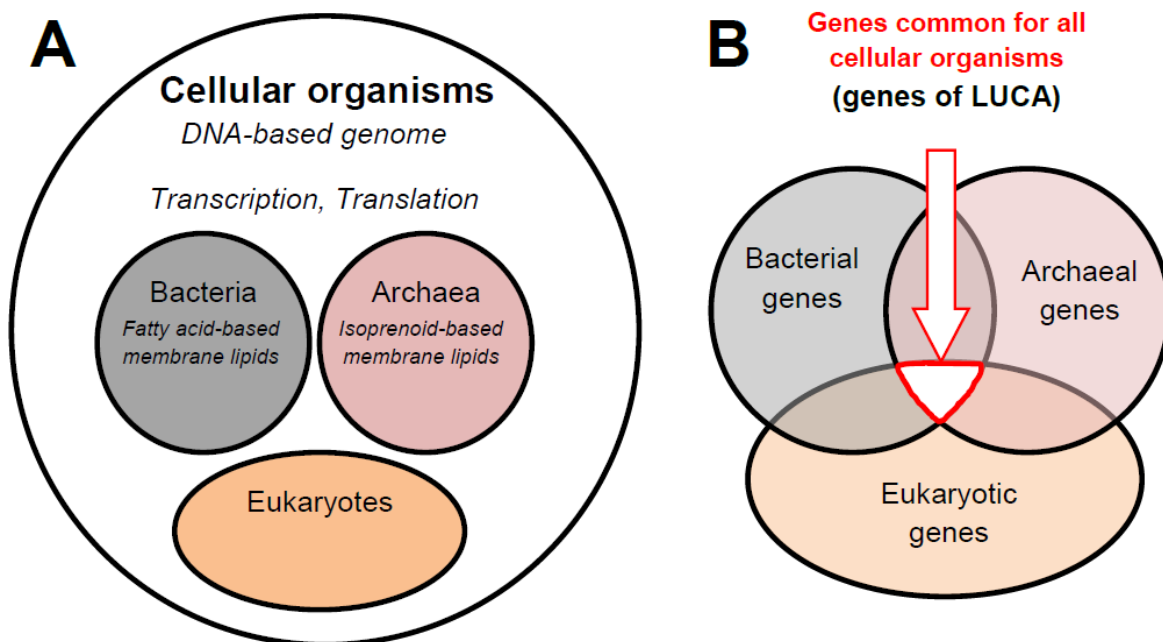
(A) Phylogenetic tree of the hexamer subunits of rotary membrane ATPase; (B) the relationship between three domains as inferred from the tree.

The figure is adapted from (Gogarten *et al.*, 1989). The catalytic subunit  $\beta$  (A, 70 kDa) is colored orange, the non-catalytic subunit  $\alpha$  (B, 60 kDa) is colored red. The ancestral gene (colored grey) duplication event is marked by an asterisk. The endosymbiosis between the ancestor of mitochondria and the eukaryotic host is shown by an arrow.

**Early benchmarks of the life evolution.** In the framework of the RNA world hypothesis, Benner and co-workers suggested that the modern cellular biochemistry can be considered as a "palimpsest", i.e. a parchment that was re-written several times while previous text wasn't removed completely and can still be seen through (Benner *et al.*, 1989). They have pointed out that metabolism even of the first RNA-based organisms appears to be complex as a number of cofactors can be dated back to this stage. Many cofactors, such as  $\text{NAD}^+/\text{NADP}^+$ ,  $\text{FAD}/\text{FMN}$ ,  $\text{CoA}$ ,  $\text{ATP}/\text{ADP}$ , and  $\text{S-adenosyl methionine}$  contain RNA-related moieties that have nothing to do with the current biochemical function of these cofactors, but could serve as "handles" in the RNA world. At the next stage, denoted as a "*progenitor*" stage, the biochemical functions could shift towards proteins, which were, as considered from the universal structure of ribosomes and genetic code in all extant life, coded by nucleic acids. Combining the emerging knowledge of gene sequences with biochemical reasoning, Benner and co-authors have suggested a number of metabolic pathways which could have been

established already at the stage of the very first, "breakthrough" organisms; they hypothesized that the breakthrough organisms could synthesize DNA from RNA and possessed pathways for tetrapyrrole biosynthesis and biosynthesis of terpenes (compounds formed by polymerization of isoprenoids).

**LUCA – the Last Universal Cellular Ancestor and its possible properties.** The full genomes of cellular organism became available starting from 1995 and were immediately used for comparisons. A rough scheme of gene distribution is shown in *Figure 1.1.4B*: some genes are restricted to one domain, some are shared among two of them and some genes are common for all three domains (it should be stressed that rigorous definition for the term "the same gene" is still absent; this problem is discussed in Section 2.4 in more detail).



*Figure 1.1.4.* Evidence in favor of the LUCA (Last Universal Cellular Ancestor) as demonstrated by common features of cellular organisms (A) and by gene comparisons (B).

Emergence of the term "Last Universal Common Ancestor (LUCA)" can be dated back to 1999 (Lazcano and Forterre, 1999). It is attributed to the life form which existed immediately before the separation of modern cellular life into bacterial and archaeal domains; this term differs from a less strict term "Last Common Ancestor" (Mushegian and Koonin, 1996). In fact, the LUCA is related very closely to the "progenitor" that was reconstructed by Benner *et*

*al.* from their analysis of common traits of modern organisms; however the properties of the LUCA are being retraced from the analyses of complete genomes (Koonin, 2003; Kyrpides *et al.*, 1999).

**The choice between the models of simple and complex LUCA, respectively, depends on the extent of the Lateral Gene Transfer (LGT).** It is not yet clear if the LUCA was something similar to a separate "species" (i.e. a group of organisms with restricted genetic information exchange with other groups of organisms), or if it was rather a consortium of genomes with a free gene drift.

Before the end of the 20th century it was generally assumed that gene evolution is shaped by only three types of events: (1) the emergence of a gene in evolution, (2) the transmission of the gene to the descendants (this type of inheritance is called *vertical inheritance*) and (3) the loss of the gene in some of the descendants. In this framework, if a certain gene is found in at least some bacteria and some archaea, it should have been in their common ancestor, the LUCA. A lot of genes were dated back to the LUCA based on this criterion, see e.g. (Castresana and Moreira, 1999), making LUCA a highly complex organism which would have possessed numerous variants of metabolic pathways (Glansdorff *et al.*, 2008). But, in the meantime, also a fourth type of genetic evolutionary event was identified and called Lateral Gene Transfer (LGT) (Berg *et al.*, 1975; Guarente *et al.*, 1980; Kleckner, 1977). This term emphasizes that a gene can be transferred to a cell not from its progenitor but from a contemporary cell (Krawiec and Riley, 1990; Smith *et al.*, 1991).

LGT was considered to be a minor event until the paper by Koonin *et al.* which suggested the LGT to be a very important mechanism for explaining gene occurrence, at least in prokaryotes (Koonin *et al.*, 2001). This change of perspective, which was impossible before sequencing of full genomes, heavily influenced the debates on the nature of the LUCA. If LGT was intensive, then the LUCA contained – for sure – only those genes that are common for all archaea and all bacteria. It could contain also other genes common for some archaea and some bacteria, but for each gene it should be then determined whether it has been involved in the LGT between archaea and bacteria or not. A mere presence of a certain gene both in bacteria and archaea could no longer be considered an evidence for its presence in the LUCA.

**Universal genes as a core of the LUCA gene set.** The number of protein-coding genes that are common to all cellular life forms initially decreased with more genomes being sequenced, but stopped at the mark of about 60 (Koonin, 2000; Koonin, 2003). The functional distribution of these genes is remarkably different from uniform and provides some information about the LUCA. The major functional group of ubiquitous proteins are the proteins involved in translation (Koonin, 2003; Mushegian, 2005). These include proteins of large and small ribosome subunits, translation factors, and several aminoacyl-tRNA-synthetases (Charlebois and Doolittle, 2004; Koonin, 2000). Hence, it appears that the translation machinery of the LUCA was pretty complex. As deduced from the common gene set, it already contained specific ribosomal proteins and even translation factors, mostly regulatory GTPases. Thus, the LUCA would have already had a proteins coded by nucleic acids (this is deduced from the diverged groups of aminoacyl-tRNA-synthetases) and synthesized through translation similar to the modern one (even with the mechanism of removal of the N-terminal methionine with the methionine aminopeptidase). Their structure could have already been rather sophisticated as a specific molecular chaperon GroEL/Hsp60 is also in the ubiquitous list.

Although modern cellular organisms have DNA-based genomes, the components of DNA replication machinery in bacteria and archaea are quite different (eukaryotes have DNA replication similar to archaea) (Edgell and Doolittle, 1997). This observation is supported by the fact that the whole core of the DNA replication genes is missing from the ubiquitous genes list. It has been hypothesized that the double-strand DNA replication had evolved after the separation of bacterial and archaeo-eukaryotic ancestors, whereas the LUCA may have used DNA that was produced by reverse transcription (Leipe *et al.*, 1999).

In fact, the functions of most other genes in the ubiquitous set are somehow connected with processing nucleic acids (RecA/RadA recombinase, transcription antitermination factor NusG, transcription pausing factor NusA, 5'-3' exonuclease, topoisomerase IA, clamp loader ATPase) or their monomers, nucleotide phosphates (thymidylate kinase, subunits of the  $\alpha_3\beta_3$ -hexamer of rotary membrane ATPase, pseudouridylate synthase). This observation implies that the LUCA could have used proteins to operate with its genetic information. The redox enzymes are limited to the enzymes of the cysteine thiol-disulfide exchange, namely

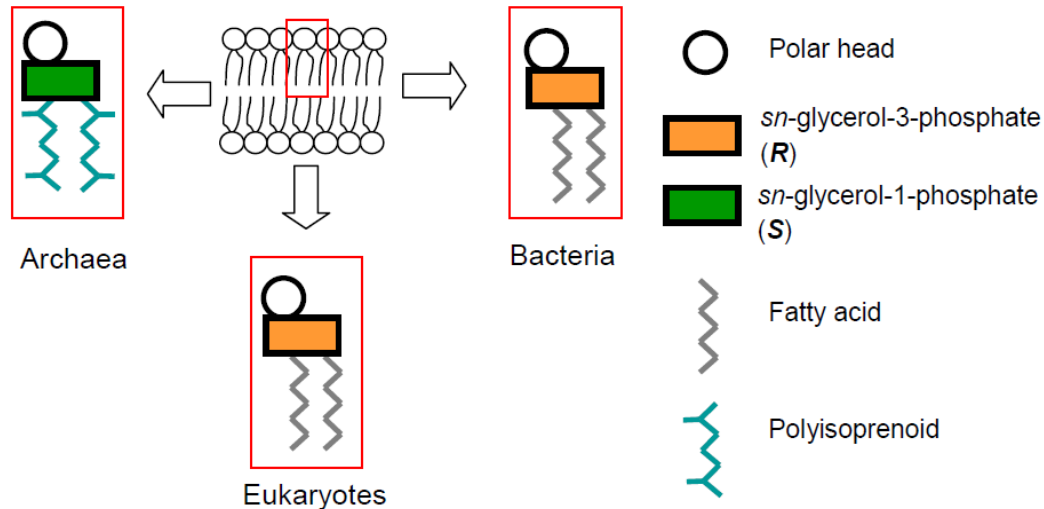
thioredoxin and its reductase. Hence, it is unlikely that the LUCA could utilize redox gradients.

Finally, as all known cellular organisms have bilayer lipid membranes, one would suggest that the LUCA should have also been encapsulated by such membranes. The main difficulty here is a dramatic difference in the chemical structure of two-tailed phospholipids in bacteria/eukaryotes and archaea. This is a topic for a detailed review and separate analysis, but some points deserve at least a brief mention. Bacterial and archaeal lipids differ in (*Figure 1.1.5*):

- the stereoisomer of glycerol phosphate (*R*-isomer is incorporated into bacterial lipids while *S*-isomer is used in archaeal lipids);
- the nature the hydrophobic tails connected to the glycerol phosphate moiety (in bacteria these are fatty acids, while archaea contain isoprenoids);
- the type of the linker bond (in bacterial lipids the aforementioned parts are connected with ester bonds, while components of archaeal lipids are connected with the ether bonds).

Different chemical nature of the archaeal and bacterial phospholipids first has brought scientists to the conclusion that the LUCA was unlikely to have any membrane (Martin and Russell, 2003), but was a virus-like entity inhabiting mineral porous structures. However, later it was pointed out that membrane proteins, in particular the *c*-subunits of rotary membrane ATPase and subunits of Sec translocation system are ubiquitous, that implies the presence of some lipid-type barrier between the inside and the outside of the LUCA cell (Jekely, 2006).





**Figure 1.1.5. Schematic representation of the structure of membrane phospholipids of modern archaea, bacteria and eukaryotes.**

Another membrane-associated ubiquitous protein is the CDP-diglyceride phosphate synthase. This enzyme is responsible for the first step of attachment of polar lipid heads to both fatty-acid based lipids in bacteria and eukaryotes and isoprenoid-based lipids of archaea (Morii *et al.*, 2000). In short, all these observations suggest the presence of at least some (perhaps primitive or different from modern) membranes in the LUCA (Jekely, 2006; Mulkidjanian and Galperin, 2010).

## 1.2. Basics of bioenergetics. ATP as a universal energy carrier between biochemical pathways

Molecules of adenosine triphosphate (ATP) serve as the "energy currency" of cells. Other nucleoside triphosphates (CTP, GTP and UTP) are characterized by free energies of hydrolysis that are similar to the free energy of ATP hydrolysis ( $\Delta G^0 = -31,8 \text{ kJ/mol}$  or  $-7,6 \text{ kcal/mol}$  at  $[\text{Mg}^{2+}] = 10^{-3} \text{ M}$  (Alberty and Goldberg, 1992; Guynn and Veech, 1973)). These nucleoside triphosphates could be either produced from ATP by transphosphorylation or from respective nucleoside diphosphate by substrate phosphorylation. They are preferred by certain enzymes in particular cellular reactions (Cramer and Knaff, 1991) (e.g. translation factors perform GTP hydrolysis, glucose-1-phosphate can react with UTP yielding important

metabolite UDP-glucose, and CTP is involved in the final steps of the phospholipid biosynthesis). A vast majority of cellular ATP is synthesized at cellular membranes upon *oxidative phosphorylation* and *photophosphorylation*. These two cellular processes are based on the same principles of chemiosmotic coupling (Cramer and Knaff, 1991; Mitchell, 1961; Mitchell, 1966; Skulachev, 1988; Skulachev, 1972).

Briefly, chemiosmotic coupling implies that energetically favorable reactions within biological membranes drive generation of transmembrane difference in electrochemical potentials of the so-called "coupling" ions ( $H^+$  or  $Na^+$ ). The transmembrane difference in electrochemical potentials of hydrogen ions (proton potential) is designated as  $\Delta\tilde{\mu}_{H^+}$ . The energetics of many prokaryotic organisms is based on utilization of the transmembrane difference in electrochemical potentials of sodium ions ( $\Delta\tilde{\mu}_{Na^+}$ ) (Skulachev, 1988). Generally, the transmembrane difference in electrochemical potentials of an ion can be designated as  $\Delta\tilde{\mu}_i$ .

Mitchell has defined the proton-motive force (PMF) by the equation:

$$PMF = \Delta\psi - 2.3 \frac{RT}{F} (pH_{in} - pH_{out}) = \Delta\psi - 2.3 \frac{RT}{F} \Delta pH,$$

where  $\Delta\psi$  is difference in electric potential,  $\Delta pH$  is difference in chemical potential and  $pH_{in}$  corresponds to pH from the *n*-side of the membrane (after "negative", or "inside" for bacterial cell) while  $pH_{out}$  gives pH from the *p*-side (after "positive", for bacterial cell it is "outside"). The same equation can be used to calculate sodium motive force (SMF) if  $Na^+$  is used as a coupling ion (Chernyak *et al.*, 1983; Mitchell, 1984).

Although PMF or SMF can be used to synthesize ATP from ADP and inorganic phosphate, the energetic functions of the membrane are not restricted to ATP synthesis. Membrane potential could also be used for doing other kinds of work, e.g. for rotating prokaryotic flagella.

### 1.2.1. Properties of ATP. High phosphate transfer potential

The reaction of ATP hydrolysis can be described as a transfer of a phosphoryl group to a water molecule. The highly negative value of the standard free energy of ATP hydrolysis

( $\Delta G^0 = -31,8$  kJ/mol or  $-7,6$  kcal/mol at  $[\text{Mg}^{2+}] = 10^{-3}\text{M}$  (Alberty and Goldberg, 1992; Guynn and Veech, 1973)) arises from both the positive value of  $\Delta S^0$  and the negative value of  $\Delta H^0$  (as  $\Delta G = \Delta H - T \cdot \Delta S$ ). The actual free energy of ATP hydrolysis depends on concentrations of ATP, ADP and phosphate in the cell via the equation:

$$\Delta G = \Delta G^0 + RT \ln \frac{[\text{ADP}][P_i]}{[\text{ATP}]} = \Delta G^0 + 2.303RT \log \frac{[\text{ADP}][P_i]}{[\text{ATP}]}$$

and can be even lower than the standard free energy values: for the rat hepatocytes the  $\Delta G$  value was estimated as  $-48$  kJ/mol (Moran *et al.*, 2011).

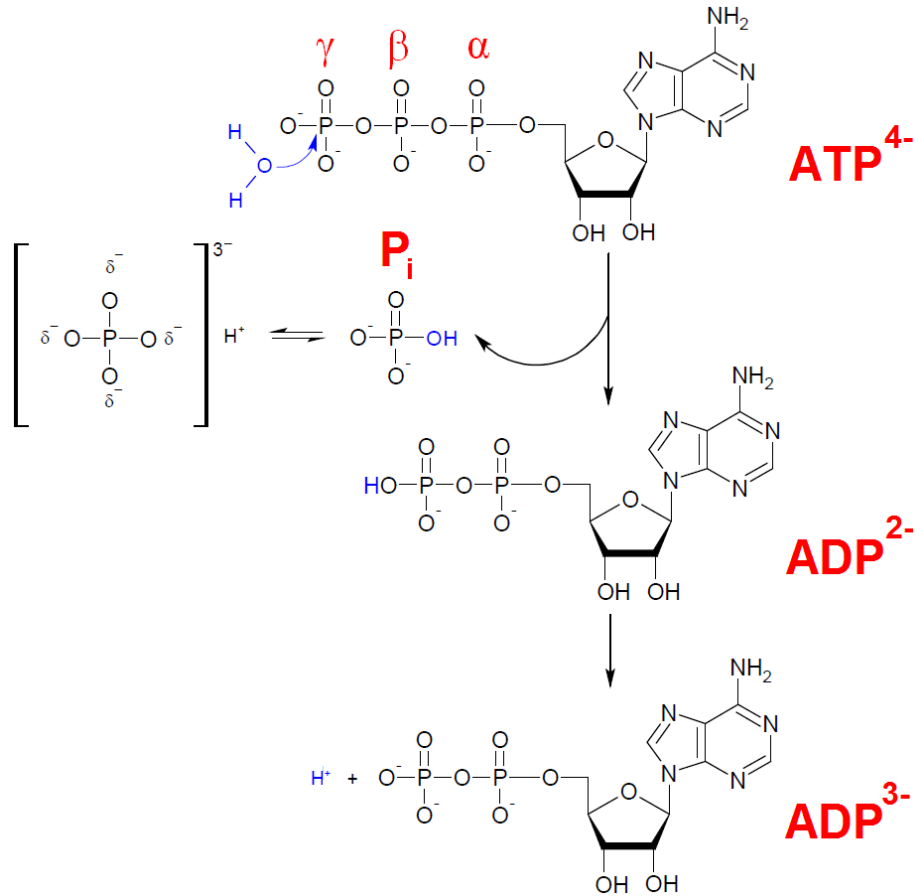
Four major factors contribute to the high standard free energy of ATP hydrolysis. These factors are summarized in **Figure 1.2.1**.

The large positive value of  $\Delta S^0$  of ATP hydrolysis has several causes. First, the number of states available for a free phosphate ion is larger than for a bound phosphate group of ATP because of the resonance (Oesper, 1950). Second, the solvation by water is different for the products and for the reactants (de Meis *et al.*, 1985; George *et al.*, 1970), generally, large positive values of  $\Delta S^0$  are typical for hydrolysis of charged compounds with notable solvation energies, such as pyrophosphate or polyphosphates, including ATP.

The large negative value of  $\Delta H^0$  of ATP hydrolysis arises from the three negative charges of the phosphate groups (in the absence of  $\text{Mg}^{2+}$ ) which cause their repulsion. As the charge depends on pH, the value of  $\Delta G^0$  is pH dependent.

The hydrolysis of ATP is often coupled with a motion or a protein conformation change. Examples of such processes are muscle contraction (Bagshaw and Trentham, 1973; Kamm and Stull, 1985), movement of protein complexes relative to a DNA molecule (Connolly and West, 1990; Putnam *et al.*, 2001; West, 1996; Yamada *et al.*, 2001), unwinding of a nucleic acid by helicases (Erzberger and Berger, 2006; Gorbalenya and Koonin, 1993; Singleton *et al.*, 2007; Skordalakes and Berger, 2003), rotation of molecular machines such as rotary ATP synthases (Noji *et al.*, 1997), switching between "on" and "off" states for signal proteins (Fukata *et al.*, 2001; Nobes and Hall, 1994; Scheffzek *et al.*, 1997) and so on. Another widespread reaction is the transfer of a phosphoryl group either directly to a substrate or to a side chain of a protein amino acid residue (Krebs and Beavo, 1979). Most metabolic reactions that require energy are realized in two steps: phosphate or pyrophosphate group is

transferred to a substrate or a protein and then is cleaved in such a way that the released energy is used to drive the respective endergonic reaction.



**Figure 1.2.1. Reaction of ATP hydrolysis and the factors that contribute to the free energy  $\Delta G^{\circ}$  for this reaction.**

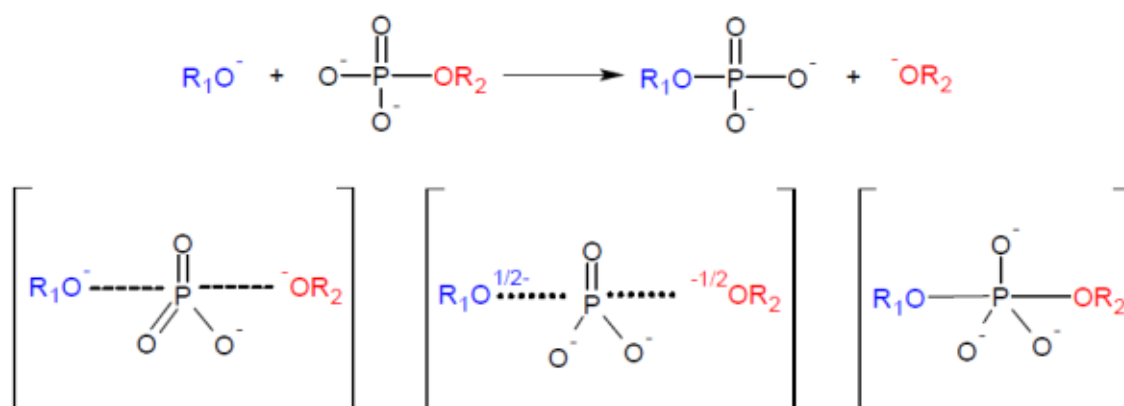
- (1) The release of the phosphate group is thermodynamically favorable since negatively charged phosphate groups repel each other;
- (2) The free inorganic phosphate ion P<sub>i</sub> is stabilized through the resonance structure which is impossible within an ATP molecule;
- (3) The direct product of hydrolysis, ADP<sup>2-</sup>, rapidly ionizes. This process is favorable since the proton concentration in the media is very low ( $10^{-7}$  at pH = 7);
- (4) The extent of hydration is lower for products of hydrolysis than for an ATP molecule itself.

Another extremely important property of ATP molecules is their kinetic stability. Phosphate groups are negatively charged and the resulting charge-charge repulsion with the attacking nucleophile contributes to the very high barrier for hydrolysis (Kamerlin *et al.*, 2013;

Westheimer, 1987), thus ATP and other nucleotides can persist in an aqueous environment (Miller and Westheimer, 1965).

### 1.2.2. Mechanisms of enzymatic hydrolysis of ATP and GTP

Hydrolysis of ATP implies a nucleophilic attack on the bond between two phosphate groups. In the course of this reaction, two possible extreme transition states are considered: dissociative (**Figure 1.2.2**, left) and associative (**Figure 1.2.2**, right). Continuum of intermediate states lies between them. Non-enzymatic reactions can proceed through all these states (Sträter *et al.*, 1996).

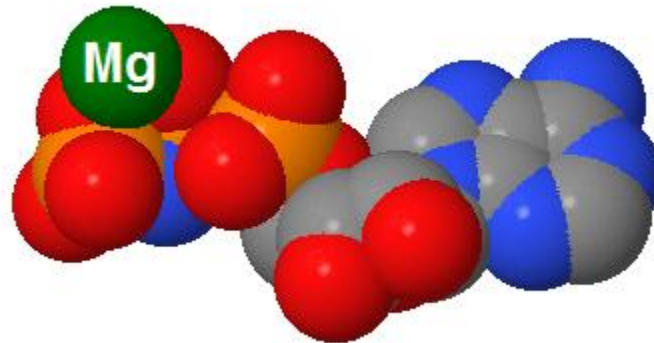


**Figure 1.2.2. General mechanism of phosphoryl group transfer.**

In the ATP hydrolysis the role of  $\text{R}_1\text{O}^-$  is played by a hydroxyl group and the role of  $\text{OR}_2^-$  is played by ADP. Dissociative, intermediate and associative transition states are depicted from left to right.

Hydrolysis of ATP and other nucleotides can be catalyzed by facilitating the nucleophilic attack. Bivalent cations can do this because of their positive charges: by neutralizing the negative charge of phosphate groups they destabilize the hydration shell. Thus, addition of a bivalent cation (usually  $\text{Mg}^{2+}$  or  $\text{Mn}^{2+}$ ) is a well-known way to stimulate the hydrolysis of ATP (Lehninger, 1950; Melchior and Melchior, 1958; Tetas and Lowenstein, 1963). Although numerous possible conformations of an ATP- $\text{Mg}^{2+}$  complex exist, only few of them are realized within enzymes, as insightfully envisioned by Melchior (Melchior, 1954). Within protein structures  $\text{Mg}^{2+}$  and ATP are usually found as a complex that is depicted in **Figure 1.2.3**.

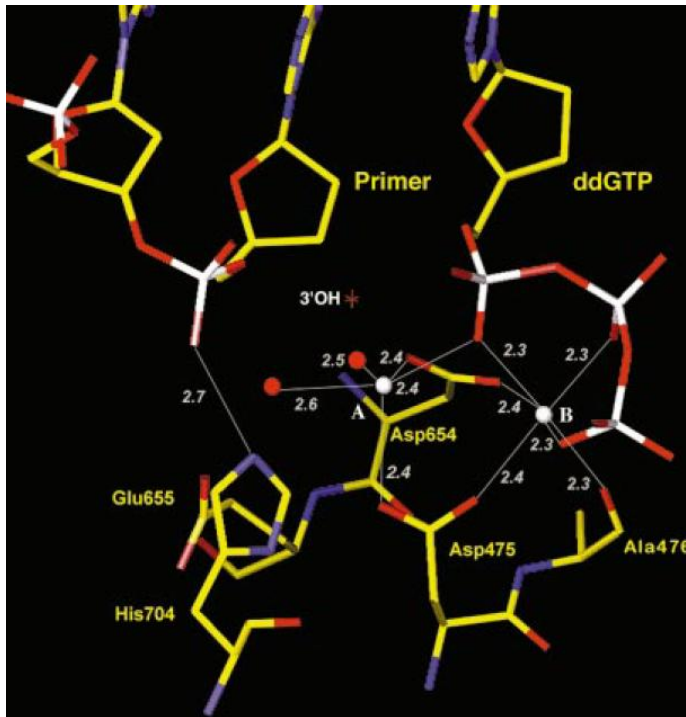
Many enzymes additionally make use of positively charged side chains of lysine or arginine residues which enter the active sites during hydrolysis. Mutations of these residues are mostly critical for the enzyme activity: mutants usually don't hydrolyze nucleotides. Typical examples of such proteins are rotary membrane ATPases (Komoriya *et al.*, 2012; Stock *et al.*, 1999) and oncogene Ras (Bourne, 1997; Scheffzek *et al.*, 1997).



**Figure 1.2.3. Relative position of the  $Mg^{2+}$  ion and the nucleotide in the yeast rotary membrane ATP synthase.**

PDB ID 2XOK (Stock *et al.*, 1999).

Other inorganic electrophiles can be used in addition to a single  $Mg^{2+}$  ion. One example of an enzyme that utilizes several cations is the DNA replication complex of bacteriophage T7 (Doublet *et al.*, 1998). Similarly to DNA polymerase from *E. coli* (Klenow fragment (Burgers and Eckstein, 1979)), this complex has binding sites for two bivalent cations:  $Zn^{2+}$  and  $Mg^{2+}$  or  $Mn^{2+}$ . Bivalent cations engage the bridging phosphate of a dinucleotide substrate and assist in phosphodiester bond cleavage. **Figure 1.2.4** shows the arrangement of bivalent metals in the active site of the DNA polymerase of bacteriophage T7 (PDB ID 1T7P). It is worth mentioning that the reactions are not metal-specific; for most phosphoryl transfer enzymes activity is seen with several divalent metal ions (Sträter *et al.*, 1996). Many more examples of the enzymes that are activated by divalent metal ions are given in a detailed review of Sträter *et al.* (Sträter *et al.*, 1996).



**Figure 1.2.4.** Two metals (white spheres marked "A" and "B") in the T7 polymerase active site.

Metal ions ligate the unesterified oxygens of all three phosphate groups of the incoming nucleotide (figure taken from (Doublet *et al.*, 1998), PDB ID 1T7P). "A" is supposed by authors to be a  $Zn^{2+}$  ion, while "B" is supposed to be a  $Mn^{2+}$  ion.

### 1.3. Evolution of ATPases and GTPases

Enzymes that bind nucleoside phosphates and, in particular, perform hydrolysis of nucleoside triphosphates belong to many protein families and are responsible for the whole spectrum of important functions. Such enzymes have been extensively studied (a detailed review on these enzyme families can be found e.g. in ref. (Vetter and Wittinghofer, 1999)). Around a half of the aforementioned ubiquitous proteins which should have been present in the LUCA (Koonin, 2000) interact with nucleotides.

#### 1.3.1. Evolution of the P-loop GTPases and related ATPases

One of the largest superfamilies of proteins that hydrolyze nucleoside phosphates is the superfamily of P-loop containing ATPases/GTPases. These enzymes differ significantly by sequence but all contain a highly conserved region denoted as a Walker A motif (or *P-loop*) and also usually have a similar structural organization of the nucleotide-binding domain.

This superfamily can be divided into two classes, namely TRAFAC (proteins related to the translation factors) and SIMIBI (typical members of this class are proteins of the signal

recognition particle, protein MinD responsible for the proper localization of a division place in *E. coli* and BioD protein of the biotin biosynthesis pathway). Each class can be further separated into several families (35 and 23, respectively) of individual proteins. Phylogenomic analysis of each group allowed identification of 13 groups that were possibly present in the LUCA (Leipe *et al.*, 2002).

The ancient groups from the TRAFAC class contain translation factors (IF2, EF-G/EF2, EF-Tu/EF1, SelB/eIF2g) and other proteins possibly operating in the protein synthesis. These are related proteins obg/DRG and YyaF/Ygr210. Functions of another group of evolutionarily old proteins HflX are not exactly known despite the fact that the protein has been purified (Dutta *et al.*, 2009) and shown to bind the 50S-subunit of the ribosome (Jain *et al.*, 2009). The last group of TRAFAC proteins which can be putatively traced to the LUCA is composed of YawG/YlqF proteins. They are suggested to play a role in the assembly of the large ribosome subunit (Im *et al.*, 2011).

Analysis of the SIMIBI class has indicated the following 5 groups of proteins that could be attributed to the LUCA. PyrG protein is a CTP-synthase (Weng *et al.*, 1986) that catalyzes the ATP-dependent amidation of the UTP. PurA is responsible for the adenylosuccinate synthesis from aspartate and inosinic acid (hypoxantine ribonucleotide phosphate). Group Mrp/NBP35 is present in all domains of life. These enzymes are highly similar to the well-studied MinD protein (which regulates the position of the separating septum made of polymerizing FtsZ protein (de Boer *et al.*, 1989; Ivanov and Mizuuchi, 2010)) and thus is suggested to take part in the cell division. SR/FtsY is the  $\alpha$ -subunit of the receptor and SRP54/Ffh is a GTPase. These proteins function as part of the protein-RNA complex which transports proteins through the membrane during the translation process (Walter *et al.*, 1981). It is noteworthy that two proteins of the signal recognition particle (SRP) are also suggested to be among the LUCA enzymes.

The major division between the two classes of P-loop GTPases has been suggested to correlate with a major event in the evolution of life, the appearance of the cell membrane (Leipe *et al.*, 2002). The ancient proteins of the SIMIBI class are functionally coupled with membranes, they are involved in the transmembrane protein translocation and cell division. The early function(s) of the TRAFAC class could have been in turn connected with the translation machinery.



### 1.3.2. Evolution of the P-loop kinases and related proteins

P-loop kinases make another large family of P-loop-containing proteins. Kinases are ubiquitous enzymes that transfer the  $\gamma$ -phosphate of ATP to a wide range of substrates, ranging from nucleotides and other small molecules to nucleic acids and proteins (Cheek *et al.*, 2002). Phylogenomic analysis of this family allowed ordering of its members into 40 groups. The representatives of two to four of these groups could have been present in the LUCA (Leipe *et al.*, 2003). A separate family of kinases called DxTN by its characteristic sequence motif is likely to have evolved in viruses but was then recruited by eukaryotes for several activities (Leipe *et al.*, 2003).

The Tmk kinase (thymidine monophosphate kinase or thymidylate kinase) performs phosphorylation of the thymidine monophosphate to thymidine diphosphate at the expense of ATP (Nelson and Carter, 1969). This protein is ubiquitous and could be present already in the LUCA. Other nucleotide kinases are widely spread among all organisms, but their evolutionary history is more complex.

### 1.3.3. Evolution of the AAA+ ATPases

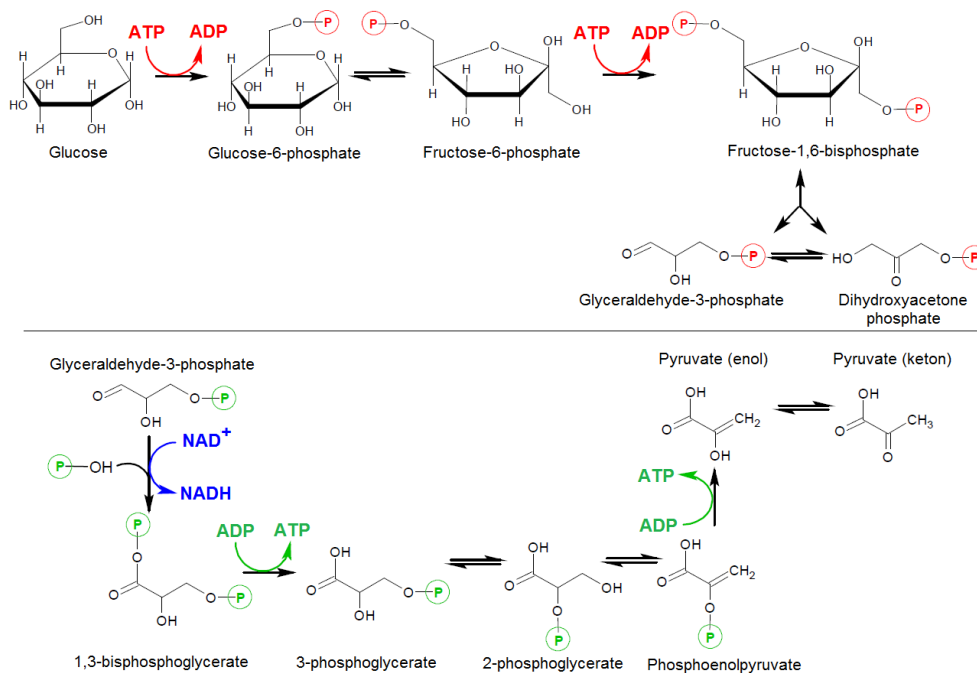
Proteins mentioned in Sections 1.3.1 and 1.3.2 belong to the Kinase–GTPase superfamily of P-loop-containing proteins. Another superfamily is denoted as ASCE (after Additional Strand, Catalytic E). ATPases from the latter superfamily contain an additional  $\beta$ -strand between the Walker A motif (P-loop) and the  $Mg^{2+}$ -binding Walker B motif. Also these ATPases have a conservative glutamate residue known to participate in the hydrolysis of phosphoester bond (Leipe *et al.*, 2003). ASCE and Kinase–GTPase superfamilies, most likely, have separated long before the LUCA. Most known chaperone-like enzymes (denoted as AAA after ATPases Associated with a variety of cellular Activities) were unified with a large number of new members into a large AAA+ superfamily (Neuwald *et al.*, 1999).

Phylogenomic analysis has suggested that 5 or 6 protein groups from this superfamily could have existed in the LUCA (Neuwald *et al.*, 1999). They are assumed to perform two important functions: (1) the driving conformational change in various proteins at the expense of ATP and (2) the unwinding of nucleic acids (helicase activity) (Iyer *et al.*, 2004).

## 1.4. ATP synthesis and $\Delta\tilde{\mu}_i$

### 1.4.1. Substrate-level phosphorylation

Since not only ATP has a high phosphoryl transfer potential, other phosphorylated compounds can also donate a phosphate group to ADP yielding ATP. Such reactions are denoted as substrate-level phosphorylation reactions. Two steps of glycolysis are classical examples of such a mechanism: here, phosphate groups are transferred from intermediate products to ADP. Glycolysis is not directly related to the scope of this thesis; we consider it only briefly, as an alternative (to oxidative phosphorylation) way of ATP synthesis (see *Figure 1.4.1*).



**Figure 1.4.1. Substrate-level phosphorylation during glycolytic digestion of glucose to pyruvate.**

Initial part of the pathway requires spending of two ATP molecules per one glucose molecule (top), while further steps (bottom) allow synthesizing of four ATP molecules, resulting in total synthesis of two ATP molecules per one oxidized glucose molecule. **Red arrows** correspond to the reactions where an ATP molecule donates a phosphate group, and **green arrows** show reactions where an ADP molecule accepts a phosphate group.

Green arrows depict two different reactions of substrate-level phosphorylation: (1) a phosphate group transfer to ADP from 1,3-bisphosphoglycerate, i.e. acylphosphate and (2) a phosphate group transfer to ADP from phosphoenolpyruvate. Further details can be found in (Nelson and Cox, 2005).

The substrate-level phosphorylation is believed to be an ancient mechanism of ATP synthesis, and its effectiveness in terms of the ATP yield, as compared to oxidative phosphorylation, is low: only two molecules of ATP are obtained per one molecule of oxidized substrate. Still, some anaerobic organisms (for instance, lactic acid bacteria) are able to survive by using only fermentation.

#### **1.4.2. Oxidative phosphorylation and photophosphorylation: rotary membrane ATP synthases**

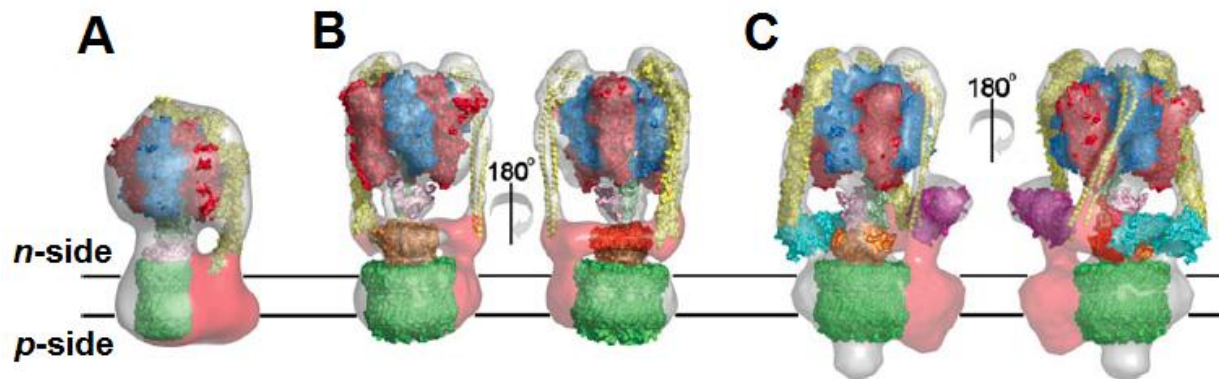
Membrane rotary ATPases/ATP-synthases are sophisticated molecular machines that couple hydrolysis or synthesis of ATP to the translocation of  $\text{Na}^+$  or  $\text{H}^+$  ions across the membrane. Two major classes of rotary ATPases, which differ in subunit composition, are (1) F-type ATPases (F-ATPases), commonly found in bacteria and eukaryotic organelles of bacterial origin (mitochondria, chloroplasts), and (2) A/V-type ATPases (A/V-ATPases) of eukaryotic intracellular membranes (V-ATPases) and archaeal cells (A-ATPases), respectively (Muller and Gruber, 2003).

##### **1.4.2.1. General structure and mechanism of rotary membrane ATPases**

The structures of rotary ATPases are depicted in **Figure 1.4.2** and **Figure 1.4.3**, where the available crystal structures of particular subunits are fitted into low-resolution density maps obtained by cryo-electron microscopy (Baker *et al.*, 2012; Benlekbir *et al.*, 2012; Muench *et al.*, 2009; Muench *et al.*, 2011). **Figure 1.4.2** shows comparative structures of rotary F-, A- and V- ATPases, whereas **Figure 1.4.3** depicts the arrangement of subunits of mitochondrial F-ATPases and proposed organization of their dimers in mitochondrial membranes. The F- and V-ATPases have different subunit composition and nomenclature of the subunits. The

main structural components of a typical bacterial F-type ATPase, taken here as an example, are:

- a hexamer of alternating three  $\alpha$ - and three  $\beta$ -subunits with a hole in the middle. The  $\alpha$  and  $\beta$  subunits are homologous, but distinct;
- membrane subunit *a* connected with the hexamer through the components of the peripheral stalk;
- components of the peripheral stalk (a dimer of b-subunits and a  $\delta$ -subunit);
- a ring of *c*-subunits in the membrane, the central cavity of which is partially occupied with the  $\gamma$ -subunit that interacts with the inner surface of the hexamer with its other end;
- an  $\epsilon$ -subunit that seems to play a regulatory role.

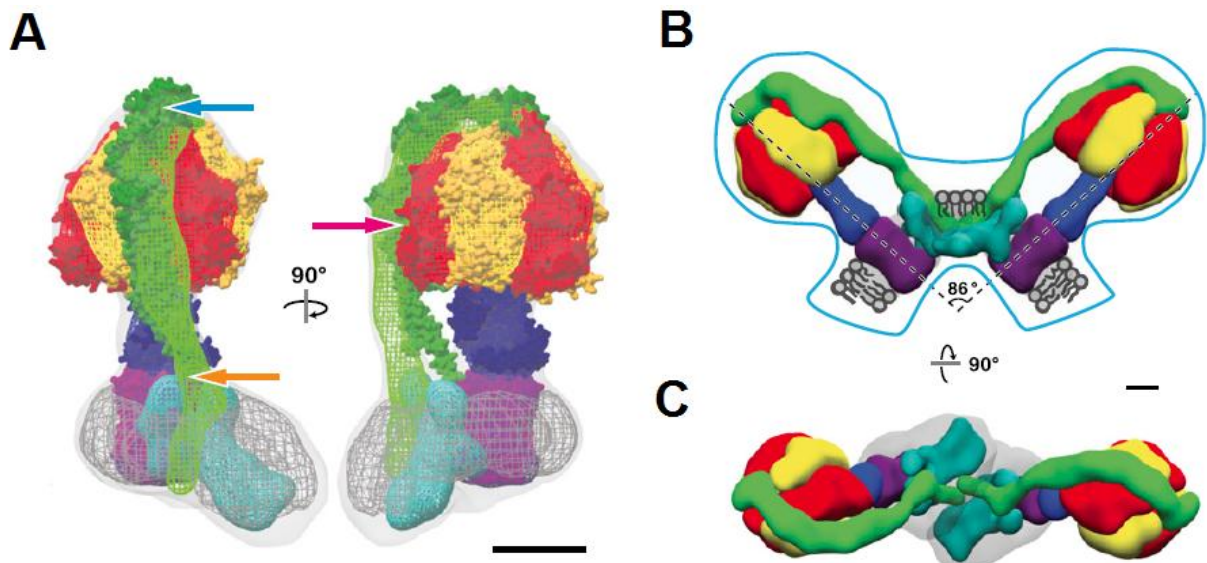


**Figure 1.4.2.** Comparative structures of F-, A- and V-ATP synthase complexes (figure from (Muench *et al.*, 2011)).

(A) Bovine mitochondrial F-ATPase.  $\alpha$ -subunits are shown in blue,  $\beta$ -subunits are shown in red, the  $\gamma$ -subunit is shown in pale green, *c*-subunits are shown in dark green, the complex of OSCP subunits (corresponding to *b* and  $\delta$  in bacterial enzyme) is shown in yellow, while the light-red moiety is showing the *a* subunit for which no structural data is available.

(B) A-ATPase and (C) V-ATPase. A-subunits are shown in red, B-subunits are shown in blue, the D-subunit is shown in pale green, the F-subunit is shown in pink. Heterodimers of subunits E and G are shown in yellow, subunit *d* is shown in orange, while *c*-subunit ring is shown in dark green. The peripheral subunits C (cyan) and H (purple) are shown for V-ATPase. The light-red moieties in the membrane correspond to subunits for which no structural data are available.

The components of rotary ATPases are assembled into two subcomplexes which can detach from each other under certain conditions, such as low ionic strength or sonication, and still preserve the ability to hydrolyze ATP and to translocate ions across the membrane, respectively (Cramer and Knaff, 1991). The  $F_1$  part (cytoplasmic, water-soluble) is composed of an  $\alpha_3\beta_3$ -hexamer, and  $\gamma$ ,  $\delta$  and  $\epsilon$  subunits. It is capable of ATP hydrolysis. The  $F_0$  membrane part consists of a  $c$ -subunit oligomer, a single subunit  $a$  and two  $b$ -subunits. It can perform ion transfer (the type of ions that can be transferred is defined by a  $c$ -subunit sequence, see Section 1.4.2.2) across the membrane. Similar structural modules are present in the A/V-ATPases which are less well studied (see (Muench *et al.*, 2011) for a review).



**Figure 1.4.3. Model of the mitochondrial F-type ATP synthase (figures from (Baker *et al.*, 2012)).**

(A) Overall structure of the complex. The  $\alpha$ -subunits are shown in red,  $\beta$ -subunits are shown in orange, the  $\gamma$ -subunit together with the  $\delta$ -subunit ( $\epsilon$  in bacterial complex) and the  $\epsilon$ -subunit (missing in bacterial complex) are shown in purple, the  $c$ -subunit oligomer is shown in violet, the peripheral stalk is shown in green. The  $a$ -subunits together with subunits  $e$  and  $g$  are colored cyan. The bar represents the scale of 50Å.

The blue and pink arrows show possible flexible regions in which the curvature of the peripheral stalk in the electron density map differs from the one observed in crystal structure. Orange arrow shows another place where the fit is not perfect: it is a location where the stalk enters the lipid bilayer.

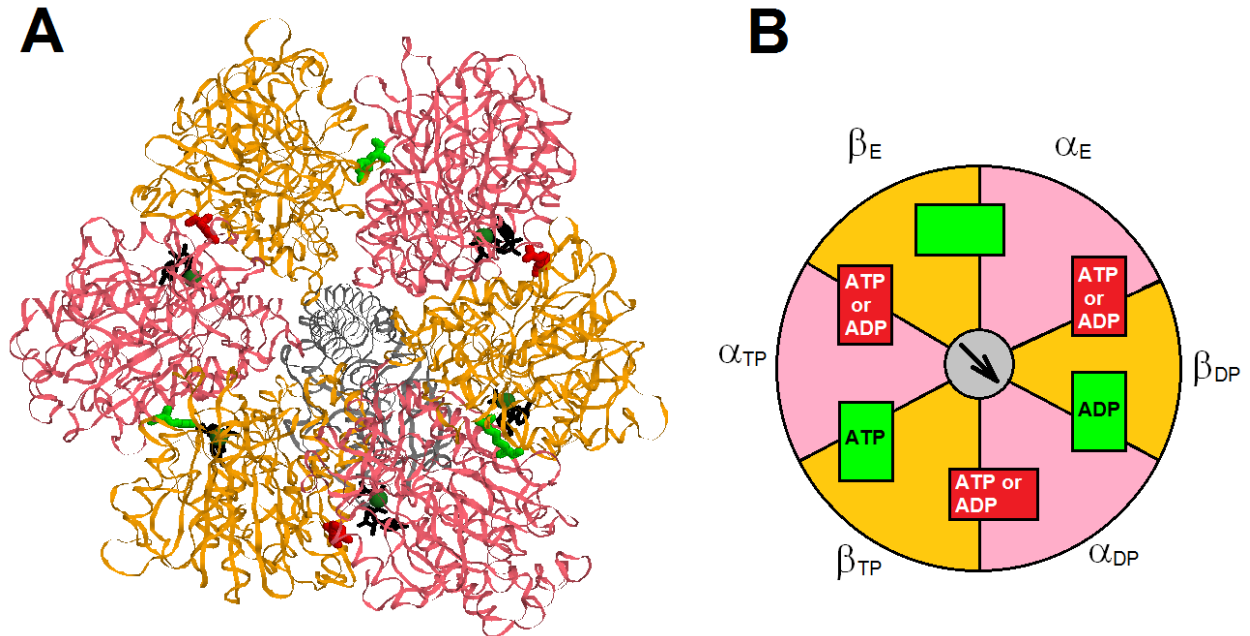
(B) and (C) Side and top views of the proposed model of dimers of ATP synthase. The bar represents the scale of 25Å.

One of the differences between F- and V-ATPases is the stability of F<sub>0</sub>F<sub>1</sub>-complex *in vivo*, as compared to the ability of the V<sub>0</sub>V<sub>1</sub> ATPases to reversibly separate into two non-functional parts *in vivo*, under conditions when a futile hydrolysis of ATP should be prevented (Beyenbach and Wieczorek, 2006; Drory and Nelson, 2006). Another difference is the number of transmembrane  $\alpha$ -helices in the subunits that make the oligomeric ring. In the F<sub>0</sub>-type rings, the *c*-subunits contain two helices, while corresponding subunits of the A<sub>0</sub>/V<sub>0</sub>-rings are typically composed of tandem homologous repeats of the sequences of F-type *c*-subunits (Mandel *et al.*, 1988; Nishi and Forgac, 2002). A four-helical proteolipid subunit was observed in the structure of V<sub>0</sub> of *Enterococcus hirae* (Murata *et al.*, 2005). The subunits of V<sub>0</sub>-oligomer of *Thermus thermophilus* analyzed with cryo-electron micrograph images of 2D crystals were suggested to contain only two helices (Toei *et al.*, 2007). The genome of *Methanopyrus kandleri* contains a gene encoding a single polypeptide that is a 13-mer repeat of a helical hairpin polypeptide (Slesarev *et al.*, 2002).

Rotary ATPases couple transmembrane cation translocation with the rotation of the *c*-subunit oligomer, together with the  $\gamma$ - and  $\epsilon$  subunits, relative to the hexamer of  $\alpha$ - and  $\beta$ -subunits. The rotation of the  $\gamma$ -subunit causes conformational changes of the nucleotide binding sites in the hexamer, which result in the synthesis or hydrolysis of ATP, depending on the direction of rotation (Nakanishi-Matsui *et al.*, 2010). A recent paper by Adachi *et al.* describes a controlled rotation in the F<sub>1</sub>-ATPase. The authors argue that the rates of binding and release of nucleotides do not depend on the rotation direction and conclude that the ATP synthesis can proceed through direct inversion of the rotation occurring during hydrolysis (Adachi *et al.*, 2012).

**Figure 1.4.4** shows the hexamer of  $\alpha$ - and  $\beta$ -subunits of the bovine mitochondrial ATPase (PDB entry 1E79, see also (Gibbons *et al.*, 2000)). **Figure 1.4.4A** shows the catalytic hexamer from the top. Subunit  $\gamma$  can be seen in the middle (colored grey). Structures of nucleotides (ADP and ATP) are shown in black bold lines; the binding sites for them are located at the interface between the subunits. However, only three sites out of six are used during synthesis or hydrolysis of ATP: they are located mostly within  $\beta$ -subunits. Green bold lines show positions of catalytically active (Le *et al.*, 2000; Senior *et al.*, 2000) arginine residues (Arg373 in the  $\alpha$ -subunits) while the red lines depict corresponding arginine

residues in catalytically inactive nucleotide binding sites (Arg356 in the  $\beta$ -subunits). In **Figure 1.4.4B** the described structure is presented schematically.



**Figure 1.4.4.** The top view on the hexamer of  $\alpha$ - and  $\beta$ -subunits in the 3D structure of bovine mitochondrial ATPase.

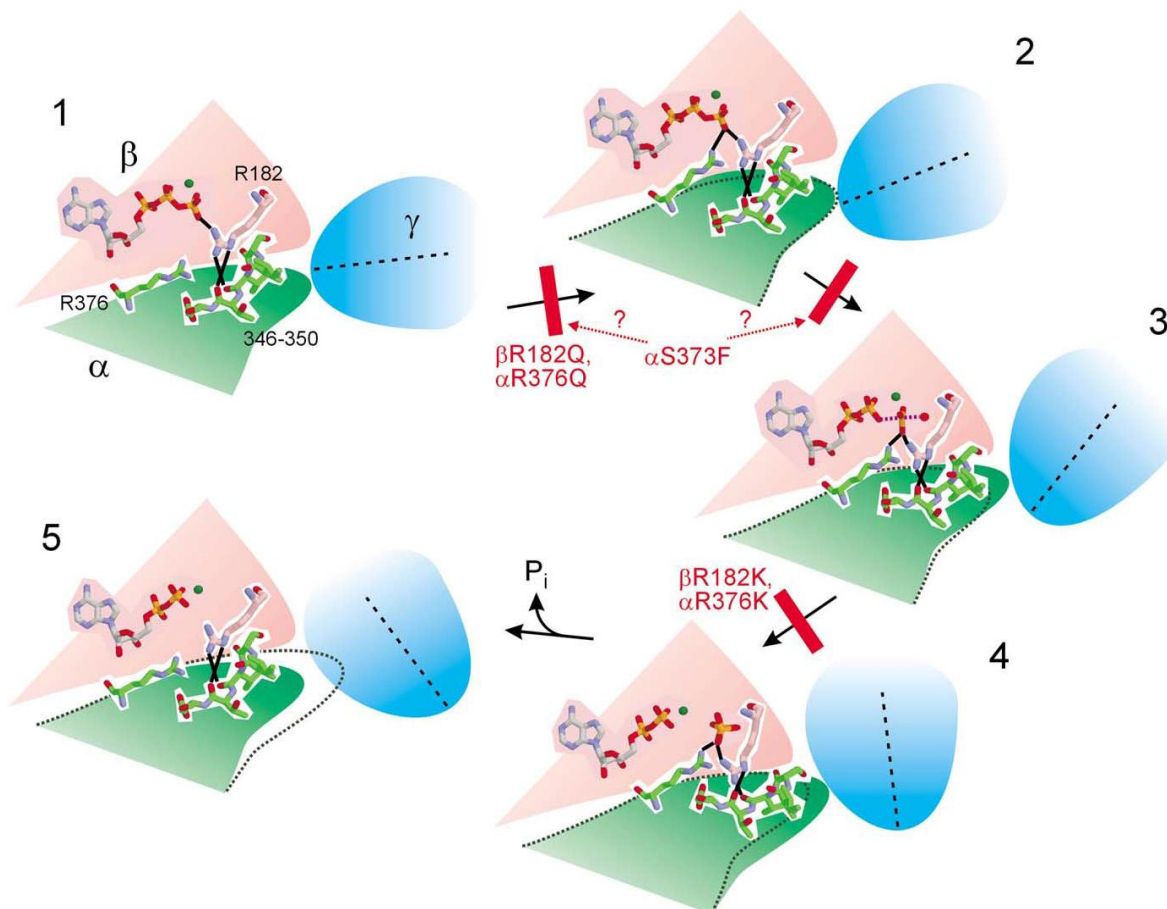
(A) Crystal structure (PDB ID 1E79); (B) schematic representation. The  $\gamma$ -subunit is shown in grey,  $\alpha$ -subunits are shown in pink and  $\beta$ -subunits are shown in orange. The binding sites for the nucleotides are shown on the scheme with green (catalytically active sites) or red (catalytically inactive sites). The same colors are used to show the "arginine finger" residues ( $\alpha$ Arg373) in the corresponding sites on the structure. Nucleotides are shown on the structure with black.

Subscripts are as in (Menz *et al.*, 2001): E for the empty conformation, DP for the half-closed conformation and TP for closed conformation.

Thus, the hexamer contains three catalytically active sites with different affinities to ATP and ADP at any step of the catalytic cycle (Boyer, 1993), with each site actually going through three states (Menz *et al.*, 2001). The particular state of each site is controlled by the relative position of the  $\gamma$ -subunit.

The rotation within the enzyme complex was demonstrated when the  $F_1$ -ATP complex was fixed with the histidine tag on the glass surface covered with Ni-NTP agarose while the  $\gamma$ -subunit was connected to a long fluorescent actin tag which allowed direct monitoring of the

rotation in a light microscope (Noji *et al.*, 1997). The mechanism of rotary ATPase includes sequential change of relative positions of the hexamer and the  $\gamma$ -subunit. Upon each catalytic step, the  $\gamma$ -subunit is rotated by  $120^\circ$ , but this is performed in two consecutive movements: the first movement (rotation by  $90^\circ$ ) is coupled with the ion translocation, while upon the second movement (rotation by  $30^\circ$ ) the substrates are being released (Yasuda *et al.*, 2001). According to (Weber and Senior, 2003), the catalytic cycle seems to be coupled to the rearrangement of the nucleotide binding site during the  $\gamma$ -subunit rotation (**Figure 1.4.5**).



**Figure 1.4.5. Mechanoenzymatic machinery of ATP hydrolysis (taken from (Weber and Senior, 2003)).**

The  $\beta$ -subunit is colored pink, the  $\alpha$ -subunit is colored green and the  $\gamma$ -subunit is colored blue. Effects of mutations which prevent transitions are shown in red. The mechanism is described further in the text.

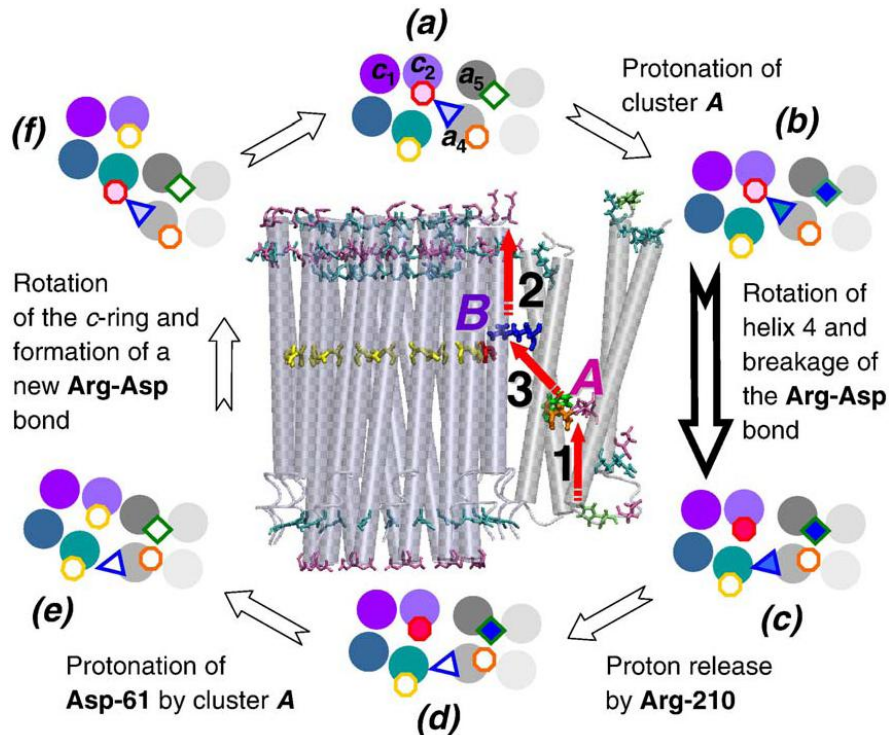
Upon ATP hydrolysis, taken here as an example, the proposed mechanism implies insertion of a positively charged residue (an "arginine finger",  $\alpha$ Arg376 in *E. coli*) into the catalytic



site upon partial rotation of the  $\gamma$ -subunit (transition from stage 1 to stage 2 in **Figure 1.4.5**). In such a position, this residue facilitates the formation of a transition state (stage 3) where the terminal phosphate of ATP is bound by  $\alpha$ Arg373 and  $\beta$ Arg182. The collapse of the transition state (stage 4) implies detachment of the phosphate group while arginine residues move away from the ADP. At this stage the  $\gamma$ -subunit would move by  $90^\circ$ . During the last  $30^\circ$  rotation (stage 4  $\rightarrow$  stage 5) the inorganic phosphate ion is released.

The mechanism of coupling between the transmembrane transfer of  $H^+$  or  $Na^+$  ions and the rotation of  $c$ -subunit oligomer is not fully clarified yet. The current models are based on the mechanism that has been initially suggested for bacterial flagella (Glagolev and Skulachev, 1978) and later adapted to the ATP synthase (Cherepanov *et al.*, 1999; Skulachev, 1988; Vik and Antonio, 1994). According to this mechanism, two non-collinear half-channels connect the  $H^+$ - or  $Na^+$ -binding sites in the middle of the membrane with the two water phases, so that translocation of ions along these channels is mechanistically coupled with a stepwise rotation of the  $c$ -oligomer relative to the  $a$  subunit. The strictly conserved arginine residue in the  $a$  subunit has been suggested to play a key role in creating an electrostatic barrier that prevents uncontrolled ion leakage (Elston *et al.*, 1998). However, the proton transfer rate via  $F_O$ , as measured with membrane vesicles (chromatophores) of *Rhodobacter capsulatus* treated by EDTA (which thus contained the  $F_O$  parts but were depleted of the  $F_1$  parts of the rotary ATPases), was independent of the  $\Delta\psi$  component of PMF (Feniouk *et al.*, 2004), which excludes the existence of long intramembrane half-channels. Alternatively it has been proposed that two membrane-embedded protonable groups denoted  $A$  and  $B$  with different pK values could be involved in the binding and release of protons, respectively, and that the rate limiting step of the overall transition is a rotative conformational change coupled to the proton transfer between the groups  $A$  and  $B$  (Feniouk *et al.*, 2004). The group  $A$  from the  $p$ -side of the membrane was inferred to have pK  $\sim 6$  and could correspond to a cluster of conserved ionizable residues of helices 4 and 5 of the  $a$ -subunit (Feniouk *et al.*, 2004; Mulkidjanian, 2006). The strictly conserved arginine residue in the  $a$  subunit (Arg210 in *E. coli* and Arg164 in *R. capsulatus*) would be a candidate for the group  $B$  with an inferred pK value of around 10 and in equilibrium with the  $n$ -side of the membrane (Feniouk *et al.*, 2004; Mulkidjanian, 2006). This hypothetical mechanism is summarized in the **Figure 1.4.6**. A recent cysteine mapping study confirmed the existence of two envisioned protein cavities

reaching the ionizable cluster within subunit *a* and the conserved arginine residue, respectively (Dong and Fillingame, 2010).



**Figure 1.4.6. Proposed mechanism of the ion translocation through  $F_0$  (taken from (Mulkidjanian, 2006)).**

In the middle, an oligomer of *c*-subunits from *Ilyobacter tartaricus*  $\text{Na}^+$ -ATP synthase (PDB ID 1YCE (Meier *et al.*, 2005)) is combined with a model for the *a*-subunit of *E. coli* complex (PDB ID 1C17 (Rastogi and Girvin, 1999)). Uncharged carboxyl groups in the *c*-ring (Glu65 in *I. tartaricus* corresponds to Asp61 in *E. coli*) are colored yellow, the only charged carboxyl group (shown in red) interacts with the Arg210 of subunit *a* (shown in blue). The residues in subunit *a* which form the protein-binding cluster A are shown in green (His245) and in orange (Glu219). Other arginine and lysine residues are colored cyan, other glutamate and aspartate residues are colored pink and other histidine residues are colored lime. The red arrows show the pathway of proton transfer; the *p*-side of the membrane is at the bottom of the figure.

The ring around the central figure shows the catalytic cycle in the synthesis direction, as envisioned from the *n*-side of the membrane. Residues are shown in the same colors as in the central figure. Intensity of the filling color follows the charge strength: uncharged residues are colored white, negatively charged residues have filling from red to pink and positively charged residues have filling from blue to light-blue. The bold arrow shows the rate-limiting step. States (a) and (f) are equivalent.

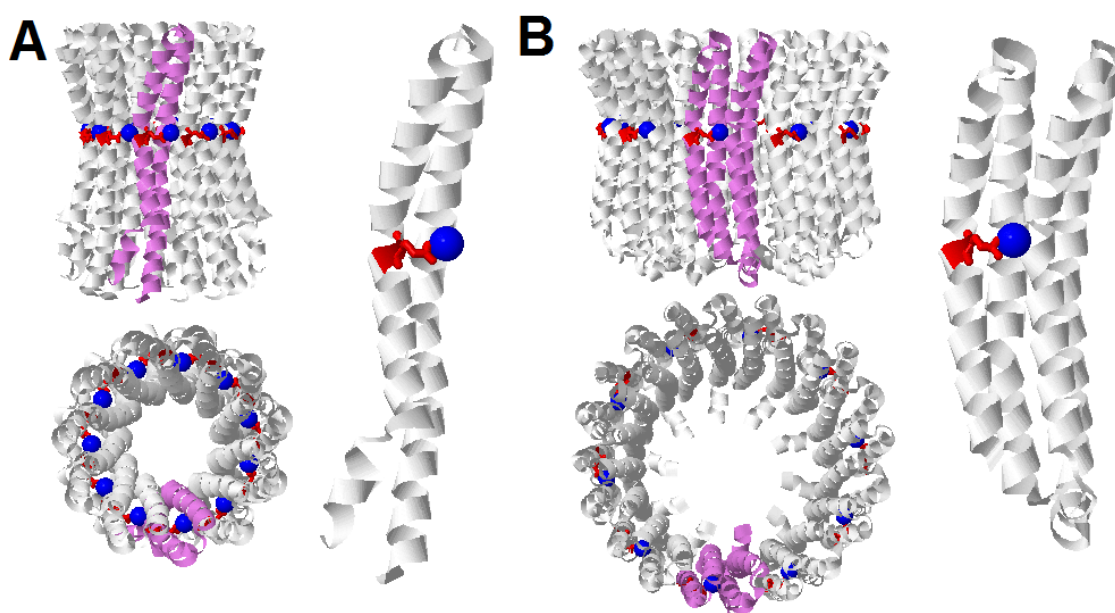
The cation binding sites in an oligomer of *c*-subunits should possess different properties depending on their relative position to the *a*-subunit. In case of proton-translocating enzymes, the sites that are facing the membrane should remain protonated in a broad pH range (up to pH of about 10) (Pogoryelov *et al.*, 2010). However if the microenvironment is being changed (for instance, through positioning of the cation site in front of a positively charged arginine residue of the *a*-subunit), the ion affinity of the site could change, enabling the release of a proton or a sodium ion.

#### 1.4.2.2. Diversity of the *c*-oligomers. Sodium- and proton-dependent rotary ATPases

The number of *c*-subunits in the ring-like oligomers can vary in a broad range; in case of F-ATPases it varies from 8 to 15. For instance, in the Na<sup>+</sup>-translocating F-ATPase from *Ilyobacter tartaricus* there are 11 *c*-subunits (Meier *et al.*, 2005; Meier *et al.*, 2005), the cyanobacterial ATPases have from 13 to 15 *c*-subunits (Pogoryelov *et al.*, 2007), whereas the animal enzymes have only 8 subunits (Watt *et al.*, 2010). This number of subunits equals the number of cations that are to be transferred across the membrane during one full rotation of an oligomer. A full rotation, in turn, allows the synthesis of three ATP molecules. In other words, a large number of *c*-subunits allows the enzyme to synthesize ATP under conditions of lower  $\Delta\tilde{\mu}_{H^+}$  or  $\Delta\tilde{\mu}_{Na^+}$  (Muench *et al.*, 2011; von Ballmoos *et al.*, 2008). Thus, the enzymes of cyanobacteria can actively synthesize ATP even at PMF values around 150 mV (Busch *et al.*, 2012). In a number of studies, the stoichiometry of *c*-subunits in ATP synthase of a particular organism was shown to be constant despite the differences in the growth media. This was checked for *E. coli* (Ballhausen *et al.*, 2009), chloroplasts (Meyer Zu Tittingdorf *et al.*, 2004), two different *Bacillus* strains (Ivey *et al.*, 1994; Meier *et al.*, 2007) and even for the *Acetobacterium woodii* which has an atypical *c*-subunit ring formed by 9 F<sub>O</sub>-type 8 kDa subunits and 1 V<sub>O</sub>-type 18 kDa subunit (Fritz *et al.*, 2008).

Usually *c*-subunits are short proteins consisting of two transmembrane  $\alpha$ -helices connected with a short loop. Mutations in the loop region can disturb the proper binding between F<sub>1</sub> and F<sub>O</sub> (Deckers-Hebestreit and Altendorf, 1996). The typical structure of the *c*-subunit oligomer is shown in **Figure 1.4.7A** with the *c*-ring of the sodium-dependent F-ATPase from *Ilyobacter tartaricus* as an example (Meier *et al.*, 2005). The residue Glu65 (colored red)

corresponds to the Asp61 of *E. coli* protein which binds a proton. Interestingly, in the *E. coli* F<sub>1</sub>F<sub>0</sub>-ATP synthase, this essential for the ion translocation residue could be moved to the adjacent helix without loss of the function (Miller *et al.*, 1990). Only mutants that had this residue in the middle of the membrane were able to produce ATP by oxidative phosphorylation (Miller *et al.*, 1990). However, not all *c*-subunits have the same arrangement. The V-ATPase from *Enterococcus hirae* (PDB ID 2BL2 (Murata *et al.*, 2005)) contains 10 *c*-subunits each formed by 4 helices (**Figure 1.4.7B**). The glutamic acid residue that corresponds to Glu65 in the *I. tartaricus* *c*-subunit and Asp61 in the *E. coli* *c*-subunit is located in the fourth helix (Glu139).



**Figure 1.4.7. Oligomers of *c*-subunits of the sodium-dependent F-ATPase from *Ilyobacter tartaricus* (PDB 1YCE) (A) and the sodium-dependent V-ATPase of *Enterococcus hirae* (PDB 2BL2) (B).**

The side view (*p*-side - bottom, *n*-side - top), the top view from the *n*-side and one individual *c*-subunit are shown.

Rotary ATPases of both F- and V-type can be either H<sup>+</sup>-translocating or Na<sup>+</sup>-translocating. Interestingly, sodium ATPases may be able to transfer protons in the absence of sodium (Dimroth, 1997; von Ballmoos and Dimroth, 2007), whereas proton ATPases appear unable to translocate sodium ions (Zhang and Fillingame, 1995). The possible reason for that is the requirement for 4-6 ligands to bind a sodium ion within the hydrophobic membrane, whereas

a proton can be bound by a single glutamic or aspartic acid residue. Indeed, a conserved set of residues in the Na<sup>+</sup>-translocating ATPases was detected, which was absent in H<sup>+</sup>-translocating ATPases (Dzioba *et al.*, 2003; Rahlfs and Muller, 1997). Proton-translocating ATPases miss one or more residues from this set; the *E. coli* *c*-subunit does not have any of the sodium ligands except for Asp61.

#### 1.4.2.3. Hypothesis on the origin of rotary membrane ATPases from RNA translocases

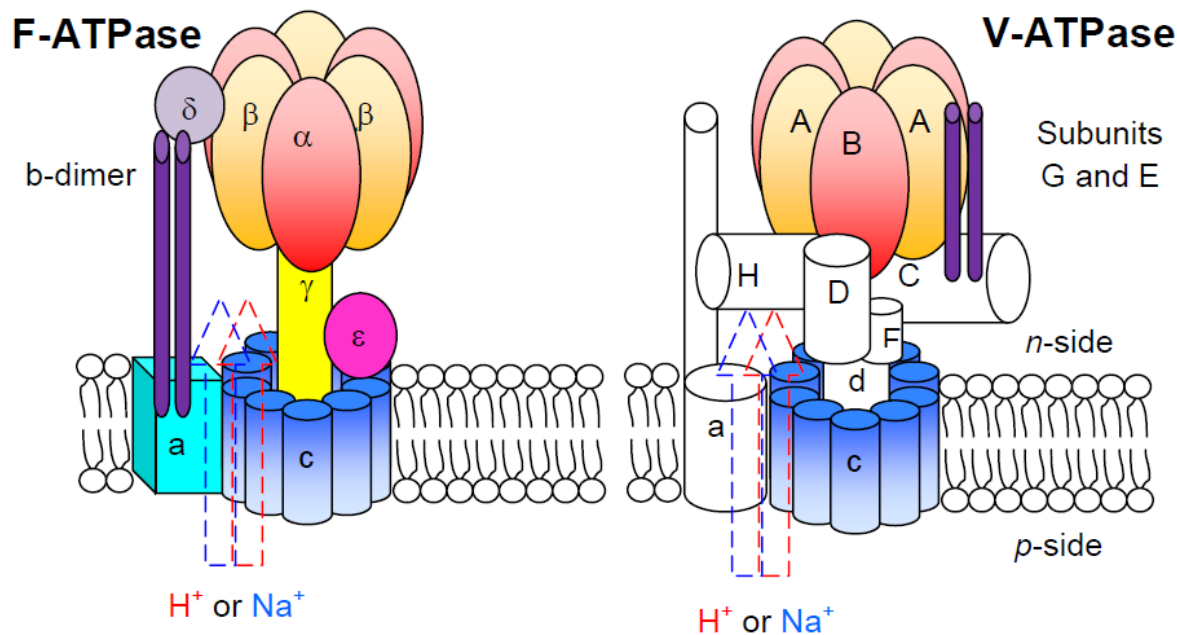
Apparent structural and functional similarities between the F- and V-type ATPases indicate a common origin of these enzymes. Both enzymes have the same mushroom-like structure and use similar catalytic mechanisms. As early as in 1997, John Walker has proposed in his Nobel lecture that the rotary membrane ATPase originated from a combination of a proton channel (ancestor of the *c*-subunits oligomer) and a RNA- or DNA-helicase (ancestor of the  $\alpha$ - and  $\beta$ -subunits hexamer) (Walker, 1998).

Indeed, the catalytic hexamers of F- and V-type ATPases are homologous to hexameric helicases (Walker, 1998). More precisely, they belong to the P-loop ATPase superfamily (see Section 1.3) and share the same subfamily with the bacterial RNA helicases Rho. These helicases have a hexameric structure with two RNA binding sites located in the inner side of the ring. This organization is common for the RNA- and DNA-helicases involved in the nucleic acid transfer (Patel and Picha, 2000). Interestingly, several of such helicases show a rotational movement during the unwinding of the RNA (Laskey and Madine, 2003; Lee and Yang, 2006; Skordalakes and Berger, 2006).

F- and V-type ATPases have homologous subunits in both membrane and cytoplasmic parts; those are the membrane *c*-subunits and the subunits of the  $\alpha_3\beta_3$  and  $A_3B_3$  hexamers, respectively (**Figure 1.4.8**) (Beyenbach and Wiczorek, 2006; Drory and Nelson, 2006; Mulkidjanian *et al.*, 2007). The homology between the  $\delta$ -subunit of the F-type ATPase, the portion of the *b*-subunit of the F-type ATPase and the E- and G-subunits of the V-ATPase, was also inferred from weak sequence similarity (Pallen *et al.*, 2006).

Homology can be also traced between the catalytic subunits and components of the peripheral stalk of rotary membrane ATPases and several subunits of bacterial flagella (responsible for the ATP-dependent export of flagellin) and the type III secretion system

(exporting bacterial toxins). However, other subunits, including components of the central stalk, are not homologous in the F- and V-ATPases. Based on this homology pattern, it has been proposed that, initially, an interaction of a helicase and a membrane pore could yield an ATP-dependent membrane translocase of RNA. Next, a switch to protein translocase could have occurred, being preceded or accompanied by the recruitment of the components of peripheral stalk. The latter could have provided better contact between the membrane and cytoplasmic parts of the complex. Mutations in the pore subunits, the ancestors of *c*-subunits of rotary ATPases, could occasionally cause sticking of the translocated protein in the pore. Then, ATP hydrolysis would have led to rotation of one part of the protein in relation to another. Charged residues which hold the membrane subunits together via salt bridges could have been recruited to translocate cations in the favorable direction. This could explain why F- and V-type ATPases have unrelated central stalk proteins: at this stage the exact nature of the central protein could have been unimportant, and different proteins could have been used at different times or in different lineages (Mulkidjanian *et al.*, 2007).



**Figure 1.4.8. Scheme of rotary membrane ATPases (figure adapted from (Mulkidjanian *et al.*, 2007)).**

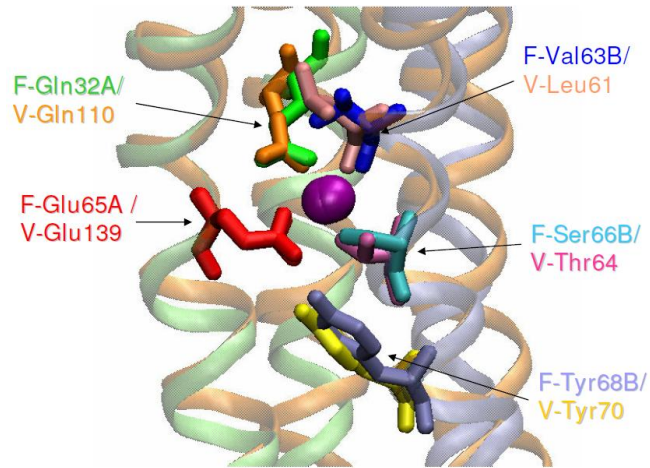
Orthologous subunits in both types of rotary ATPases are colored in the same way. The direction of cation translocation is shown for the ATP synthesis.

#### 1.4.2.4. Evolutionary primacy of sodium-dependent bioenergetics

The ion specificity of rotary membrane ATP synthase determines the type of transmembrane ion gradient which the cell can harvest and thus determines the type of bioenergetics the cell possesses. Until recently the proton-dependent rotary membrane ATP synthase and thus proton energetics in general were considered to be evolutionarily primal, while the cases of sodium-dependent energetics were believed to be secondary results of adaptation to extreme environments (Berry, 2002; Deamer, 1997; von Ballmoos and Dimroth, 2007). An alternative view (Dibrov, 1991; Hase *et al.*, 2001; Skulachev, 1988) did not draw much attention. However, the hypothesis of initial utilization of protons has at least one severe flaw: it assumes that the early membranes were tight enough to hold proton potential, which is not quite likely.

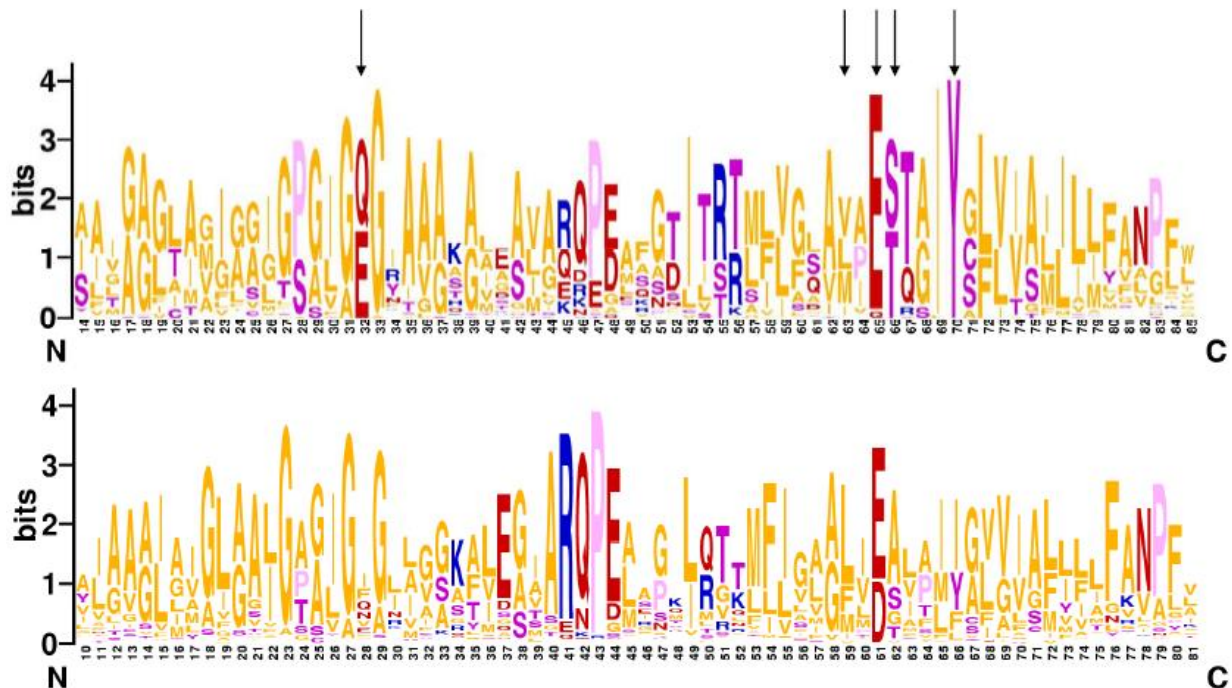
As mentioned above (Section 1.4.2.2), the sodium-translocating rotary membrane ATPases can be easily distinguished from their proton-translocating counterparts based on a specific  $\text{Na}^+$ -binding motif. This motif (**Figure 1.4.9**) has been shown to be conserved in all sodium-translocating membrane ATPases independently of their position on the phylogenetic tree (Mulkidjanian *et al.*, 2008b). Thus either exactly the same set of 5-6 residues appeared in evolution multiple times (which is unlikely) or the original form of the enzyme already had a sodium-binding site. This, in turn, suggests that the common ancestor of rotary ATP synthases could translocate  $\text{Na}^+$  ions, and existence of bioenergetics based on the SMF generation and harvesting (sodium bioenergetics) could be primal.

Evolution of membrane bioenergetics, as argued in ref. (Mulkidjanian *et al.*, 2009) can be traced through evolution of the ATP synthase (**Figure 1.4.11**). Its ancestral form could have been functioning as a translocase for macromolecules (Section 1.4.2.3), and thus could have been useful even if the membranes were leaky to low-molecular-weight compounds. Already at this stage of a RNA or protein translocase, sodium ions could have stabilized the *c*-subunit oligomers. The stability of *c*-oligomers in the modern  $\text{Na}^+$ -translocating membrane ATPases drops dramatically upon the depletion of  $\text{Na}^+$  (Meier and Dimroth, 2002).



**Figure 1.4.9.** Structural superposition of sodium-binding sites of the *c*-subunits from F-type ATPase of *Ilyobacter tartaricus* (PDB 1YCE) and K-subunits from V-ATPase of *Enterococcus hirae* (PDB 2BL2).

Figure taken from (Mulkidjanian *et al.*, 2008b). The most important coordinating bonds to the Na<sup>+</sup> ion are provided by the key ligand Glu65/Glu139 (the first residue is from *I. tartaricus* and the second is from *E. hirae*). Other residues are conserved in known sodium ATPases but are not preserved in proton ATPases.

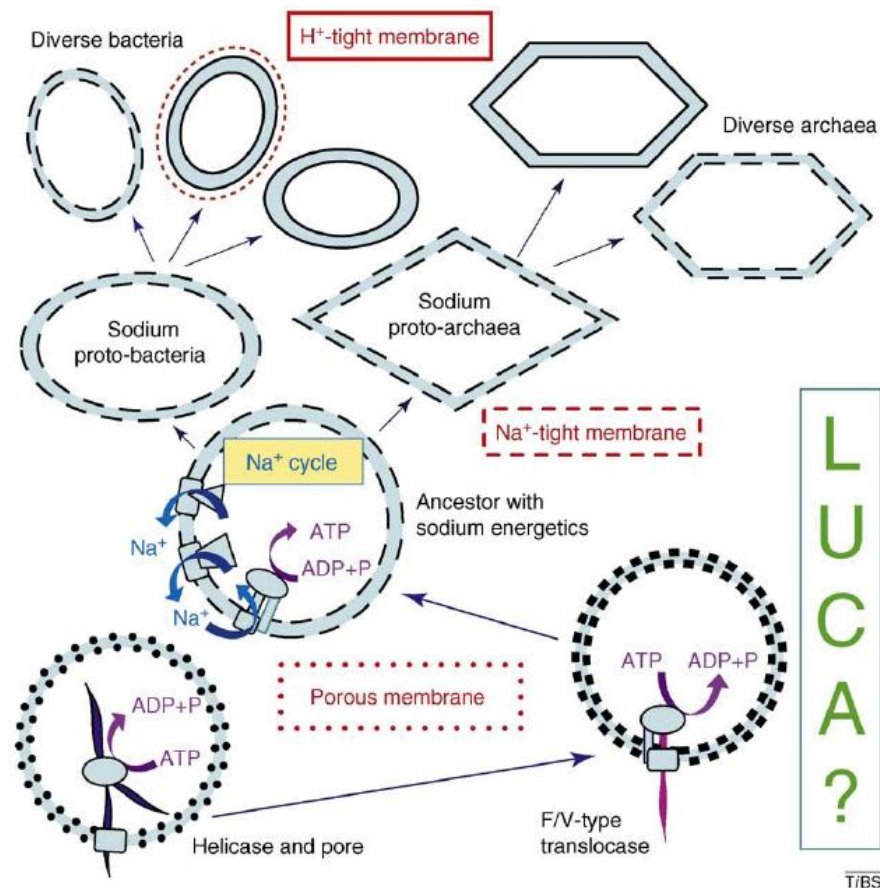


**Figure 1.4.10.** Sequence logos for the transmembrane segments of Na<sup>+</sup>-binding (top) and H<sup>+</sup>-binding (bottom) *c*-subunits.

Residues which bind sodium (also depicted in **Figure 1.4.9**) are shown with arrows. Complete Na<sup>+</sup>-binding pattern is absent in H<sup>+</sup>-binding *c*-subunits, but is partly preserved in different *c*-subunits of both F- and V-type ATPases. The figure is taken from (Mulkidjanian *et al.*, 2008b).



With the membranes becoming tight for  $\text{Na}^+$ , primordial cells could become able to control the concentrations of ions inside; an ancient  $\text{Na}^+$ -translocating ATPase, together with other  $\text{Na}^+$  pumps, could be used for outward pumping of  $\text{Na}^+$  ions. Later, the  $\text{Na}^+$ -translocating ATPase could have switched to harvesting the sodium gradient and become a rotary ATP synthase, completing the first, sodium-dependent energetic cycle. In the framework of this scenario, further evolution towards  $\text{H}^+$ -tight membranes would allow independent switching to  $\text{H}^+$ -based bioenergetics in bacteria and archaea, driven by the emergence of specific generators of PMF and a consecutive loss of  $\text{Na}^+$ -binding ligands in rotary ATP synthases.



**Figure 1.4.11.** The proposed scenario for the evolution of membranes and membrane enzymes from separate RNA helicases and primitive membrane pores, via membrane RNA and protein translocases, to the F- and V-type ATPases.

The figure is taken from (Mulikidjanian *et al.*, 2009). The scheme shows the proposed transition from primitive, porous membranes that were leaky both to  $\text{Na}^+$  and  $\text{H}^+$  (dotted lines) via membranes that were  $\text{Na}^+$  tight but  $\text{H}^+$  leaky (dashed lines) to the modern-type membranes that are impermeable to both  $\text{H}^+$  and  $\text{Na}^+$  (solid lines).

## 1.5. Review of energy-converting redox complexes

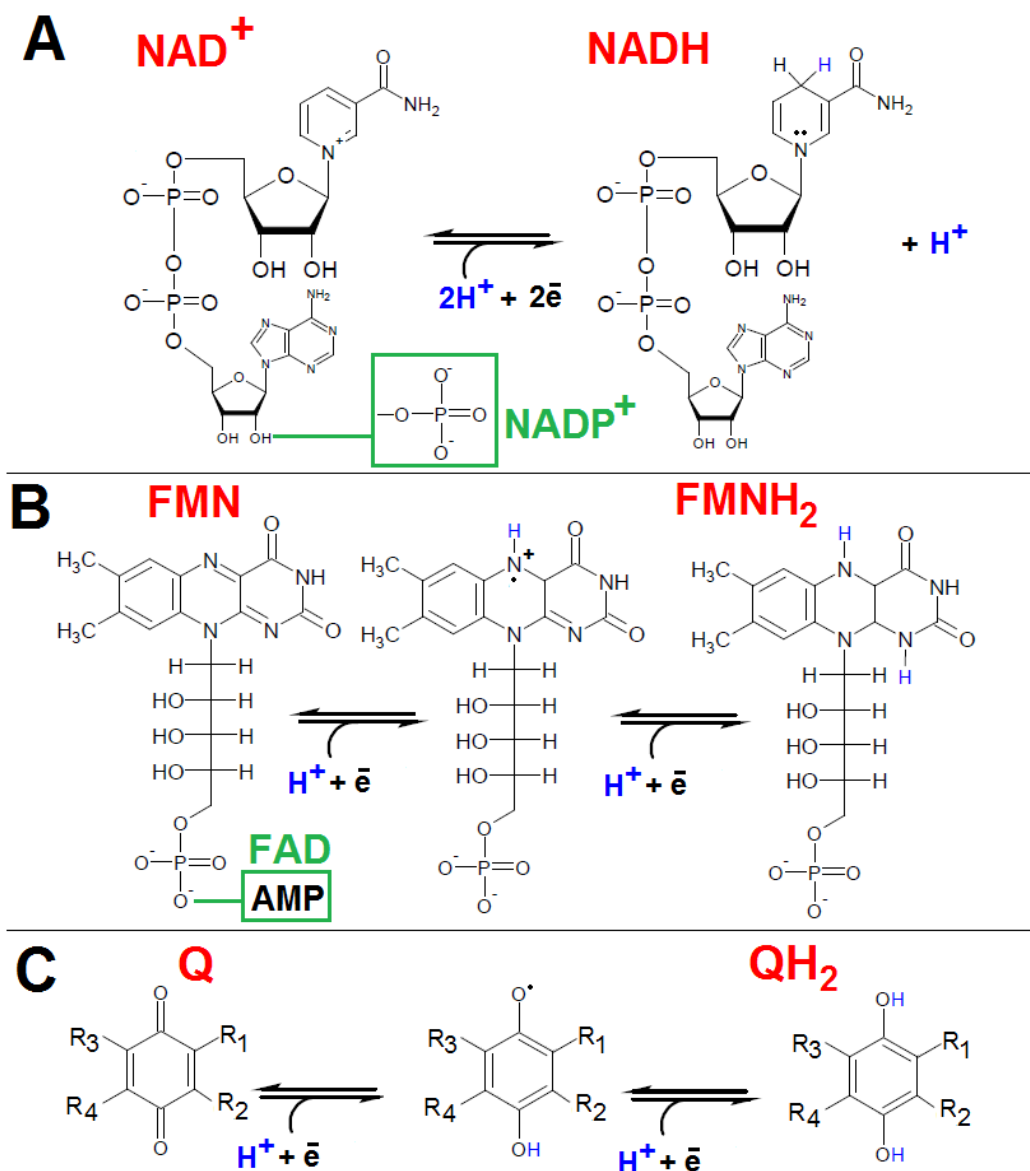
Organisms are capable of converting external energy sources, such as sunlight or reduced compounds, into internal energy sources; usually, upon such conversion, electron transfer from external low-potential electron donors to high-potential electron acceptors drives the translocation of coupling ions ( $H^+$  or  $Na^+$ ) across the membrane "against" their electrochemical potential yielding PMF or SMF.

Chlorophyll-based photosynthesis (see Section 1.5.4) is based on the same principle, but the low-potential electron donor and the high-potential electron acceptor are directly generated by the energy of absorbed light quanta. An electron flow from a donor to an acceptor, however, is not obligatory for generation of chemiosmotic potential as shown by *rhodopsins*: in bacteriorhodopsin, for example, the absorbed energy of a light quantum leads to isomerisation of the retinal cofactor, which causes a conformation change of the protein and pumping of protons across the membrane, in the absence of any redox reaction (Oesterhelt and Stoeckenius, 1973). Rhodopsins demonstrate different ion specificity: in addition to proton-pumping bacteriorhodopsin, halorhodopsin, a light-driven chloride pump (Kolbe *et al.*, 2000) and newly discovered sodium-pumping rhodopsins (Kwon *et al.*, 2013) have been described.

### 1.5.1. Chemical structures of most common redox cofactors

The electrons (*reducing equivalents*) from external electron donors are first accepted by specific carriers of reducing equivalents, *redox cofactors*, which then interact with core energy converting enzymes. Some of redox cofactors are ubiquitous ( $NAD^+/NADP^+$ , FMN/FAD), whereas others form classes of compounds which differ by the substitutions in particular positions. Structures and redox reactions of those redox cofactors that transduce two reducing equivalents are shown in *Figure 1.5.1*.

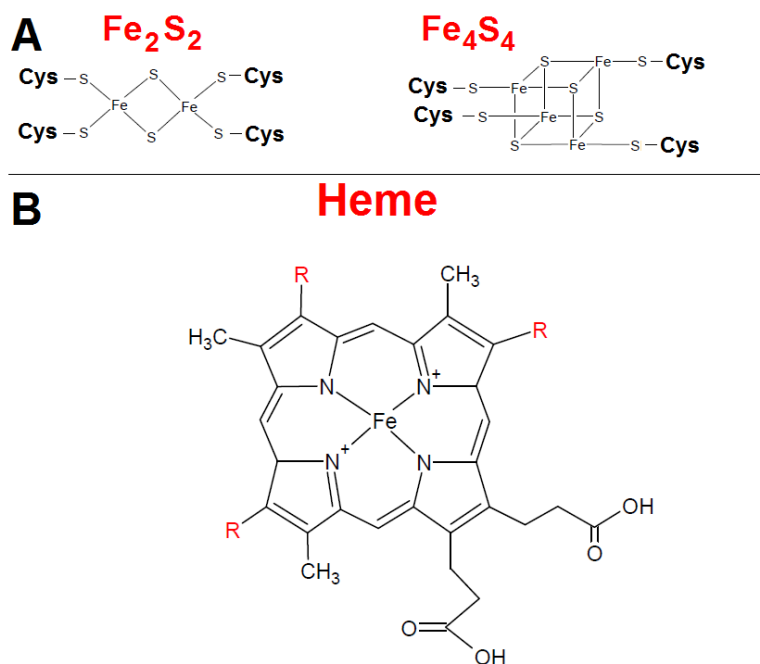
$NAD^+/NADP^+$  are obligatory *two-electron carriers*, whereas FMN, FAD and quinones can receive electrons sequentially, one by one. Quinones are usually hydrophobic, because they contain long-chain isoprenoid moieties attached to the phenyl rings.



**Figure 1.5.1. Structures and redox reactions of the cofactors which can carry two reducing equivalents.**

Protons that are bound concomitantly with electron transfers are colored blue. Green boxes show additional moieties which are present in NADP<sup>+</sup> and FAD as compared to NAD<sup>+</sup> and FMN.

Major one-electron carriers are depicted in **Figure 1.5.2**. They can transfer electrons strictly one by one. FeS-clusters consist of 2-4 iron atoms connected to sulfur and cysteinyl groups of proteins. Hemes have a ring structure typical for porphyrins, and coordinate an iron ion in the middle of the ring. All the aforementioned electron carriers are widely used in energy converting enzymes and will be referred to in the further text.



**Figure 1.5.2. One-electron carriers commonly used in energy conversion.**

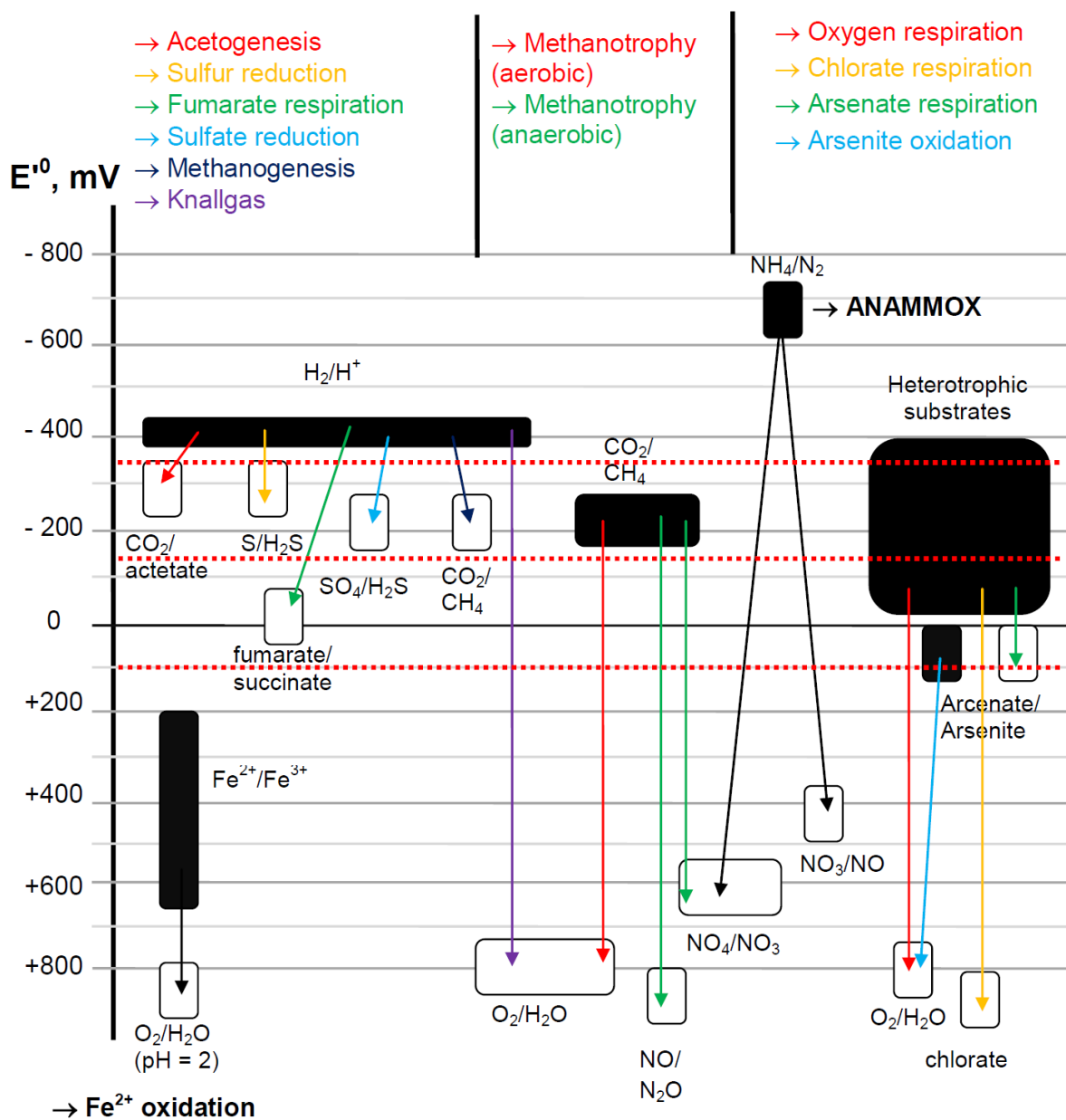
In contrast to the two-electron carriers shown in *Figure 1.5.1*, at least two types of which ( $\text{NAD}^+/\text{NADP}^+$  and quinones) are individual compounds able to diffuse within cells, FeS-clusters and hemes are tightly packed within proteins to prevent their spontaneous oxidation.

### 1.5.2. Review of the variants of electron transfer chains

The *Figure 1.5.3* presents an overview of some biologically relevant donors and acceptors of electrons. The energy gap between an electron donor used by an organism and the corresponding electron acceptor defines how much free energy could be gained and how many protons could be pumped across the membrane during reduction of an equivalent amount of electron acceptor.

To increase effectiveness of respiratory chains, organisms utilize several *coupling sites*. Usually one coupling site corresponds to an enzyme complex that couples a particular redox reaction to the transmembrane transfer of protons or sodium ions. Mobile compounds, such as hydrophobic quinols and cytochromes carry electrons between coupling sites.

The number of coupling sites per electron transfer chain (ETC) could vary. We will further show examples of such chains with one, two or three coupling enzymes.

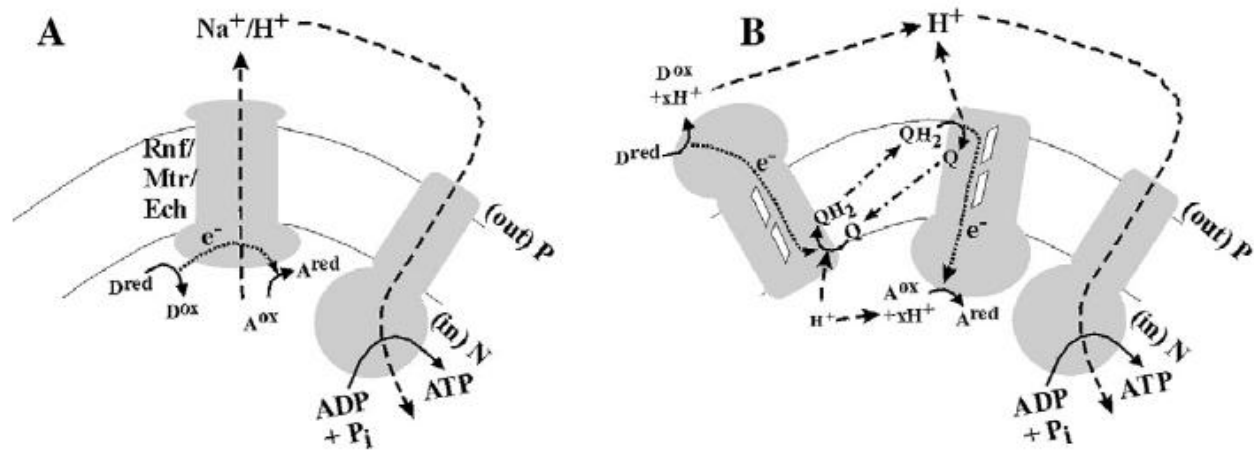


**Figure 1.5.3. Standard redox midpoint potentials for compounds that are utilized in different energy conversion processes.**

Figure adapted from (Schoepp-Cothenet *et al.*, 2013). Empty boxes show electron acceptors while black boxes denote electron donors. The names of the processes are indicated after the  $\rightarrow$  sign. The red dotted lines show oxidation potentials for NADH (-320 mV), menaquinol (-70 mV) and ubiquinol (+100 mV), respectively.

**Electron transfer chains with one coupling enzyme complex.** Acetogens and methanogens do not use quinone carriers and thus have only a single coupling enzyme complex (Rnf-complex in acetogens (Biegel *et al.*, 2011; Kumagai *et al.*, 1997; Schmehl *et al.*, 1993) and

Mtr-complex in methanogens (Gottschalk and Thauer, 2001)). In these enzymes a redox reaction in the protruding hydrophilic part of an enzyme is directly coupled to the transmembrane ion translocation. One more example of such coupling site is the electrogenic [NiFe] hydrogenase Ech (Deppenmeier, 2002; Hedderich and Forzi, 2005; Welte *et al.*, 2010) (*Figure 1.5.4A*).



*Figure 1.5.4. Schemes of different electron transfer chains.*

Figure taken from (Schoepp-Cothenet *et al.*, 2013). Dashed lines mark ion transfers, dotted lines stand for electron transfers. Diffusion of the quinones (or other membrane-soluble carriers) is shown with the dash-dotted lines. (A) Shortest scheme with a single coupling enzyme (methanogens, acetogens). (B) A more complex electron transfer chain with two enzyme complexes connected by a quinone loop.

**Two energy-converting complexes.** In the cases when quinones are involved in electron transfer, their reduction sites are usually located on the *n*-side of the membrane whereas their oxidation occurs on the *p*-side (Schoepp-Cothenet *et al.*, 2013). Then the redox reactions of quinone are accompanied by proton binding from the *n*-side of the membrane and their release to the *p*-side, yielding PMF (Jormakka *et al.*, 2003). For electron translocation across the membrane, "wires" that are formed by two collinearly placed hemes are often used (see *Figure 1.5.4B*).

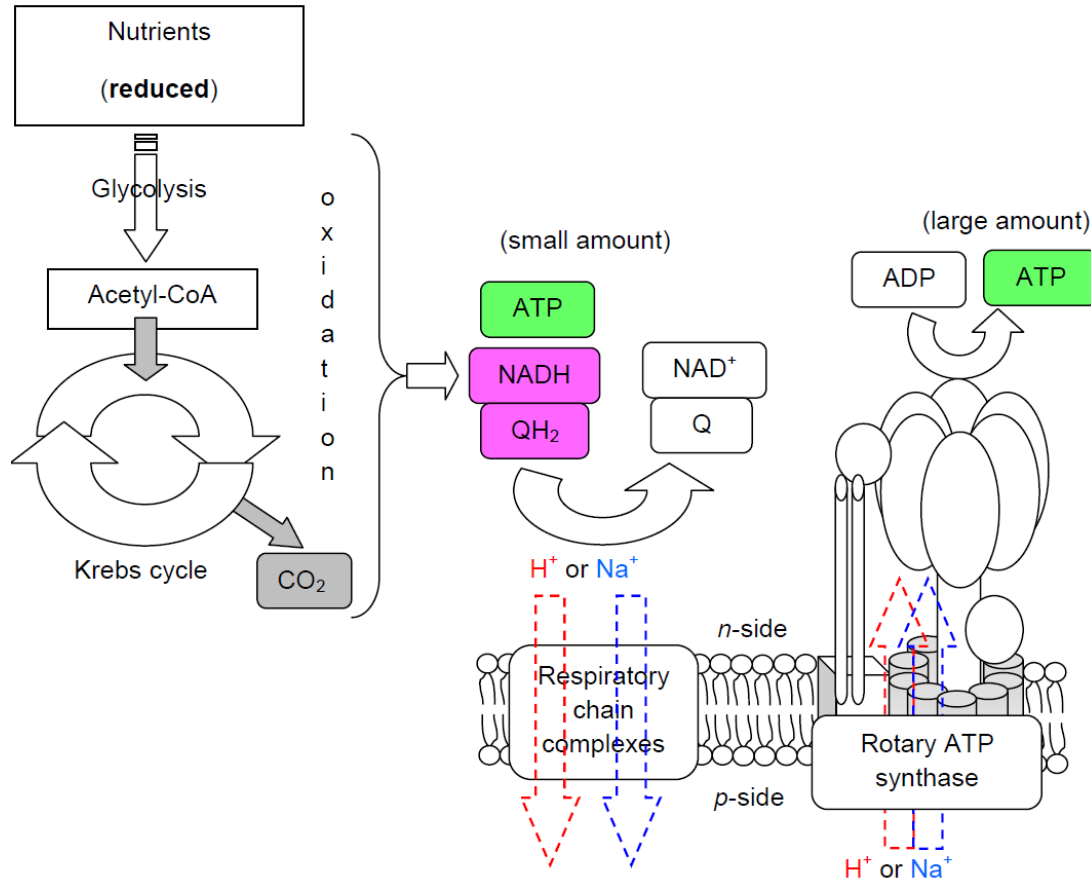
It appears that not only quinones can be used as lipid-soluble carriers. Archaea from the order *Methanosarcinales* utilize methanophenazine, a phenazine derivative with the standard transformed (at 25°C and pH 7) reduction potential  $E^0 = -165$  mV (Tietze *et al.*, 2003).

**Three coupling sites in mitochondria.** As mitochondria are descendants of  $\alpha$ -proteobacteria (Yang *et al.*, 1985), mitochondrial respiratory chain is closely related to the respiratory chains typical for  $\alpha$ -proteobacteria. It contains three coupling sites and thus allows energy conversion to be very efficient. While the catalytic cores of mitochondrial energy-converting complexes are similar to those from their bacterial counterparts, the mitochondrial energy-converting complexes have many more subunits than their prokaryotic homologs; the functions of these additional subunits mostly remain unknown. Mitochondria are attracting increased attention nowadays since their pathological conditions are associated with neurodegenerative diseases (Dawson and Dawson, 2003; Perier *et al.*, 2012; Swerdlow, 2012) and aging (Skulachev, 2007). For this reason in the present study we will discuss enzymes which have homologs in mitochondria in some more detail.

### 1.5.3. Generation of $\Delta\tilde{\mu}_{H^+}$ by respiratory chain of mitochondria

The general scheme of catabolic reactions in mitochondria is shown in **Figure 1.5.5**. Nutrients can be oxidized yielding  $\text{CO}_2$  and water through glycolysis and the Krebs cycle (left part). During these reactions, only small amounts of ATP or GTP are synthesized. The energy is stored in the form of few reduced compounds, mainly NADH and  $\text{QH}_2$ . These compounds can be oxidized by the complexes of the electron-transfer chain (*respiratory chain*). Mitochondria have three coupling enzyme complexes which, apparently, can form supercomplexes with each other (Busch *et al.*, 2012; Mileykovskaya *et al.*, 2012). The reducing equivalents from NADH and  $\text{QH}_2$  are then transferred to a terminal acceptor. In eukaryotic mitochondria and in many aerobic prokaryotes the terminal acceptor is usually oxygen.

Standard chemical potentials for the oxidation of the carriers of reducing equivalents at physiologically relevant conditions of  $25^\circ\text{C}$  and pH 7 are shown in **Figure 1.5.6**. The separation of the electron transfer between the best potential donor of electrons (NADH) and the best potential acceptor ( $\text{O}_2$ ) into several steps allows an efficient conversion of the released energy.

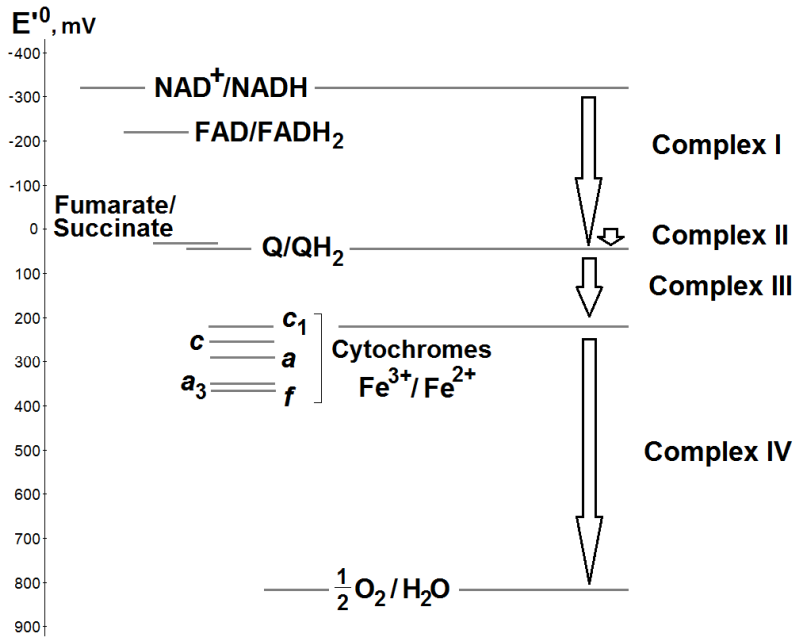


**Figure 1.5.5. General scheme of catabolic transformations in eukaryotic cell.**

Carriers of reducing equivalents are shown in magenta. ATP molecules are shown in green.

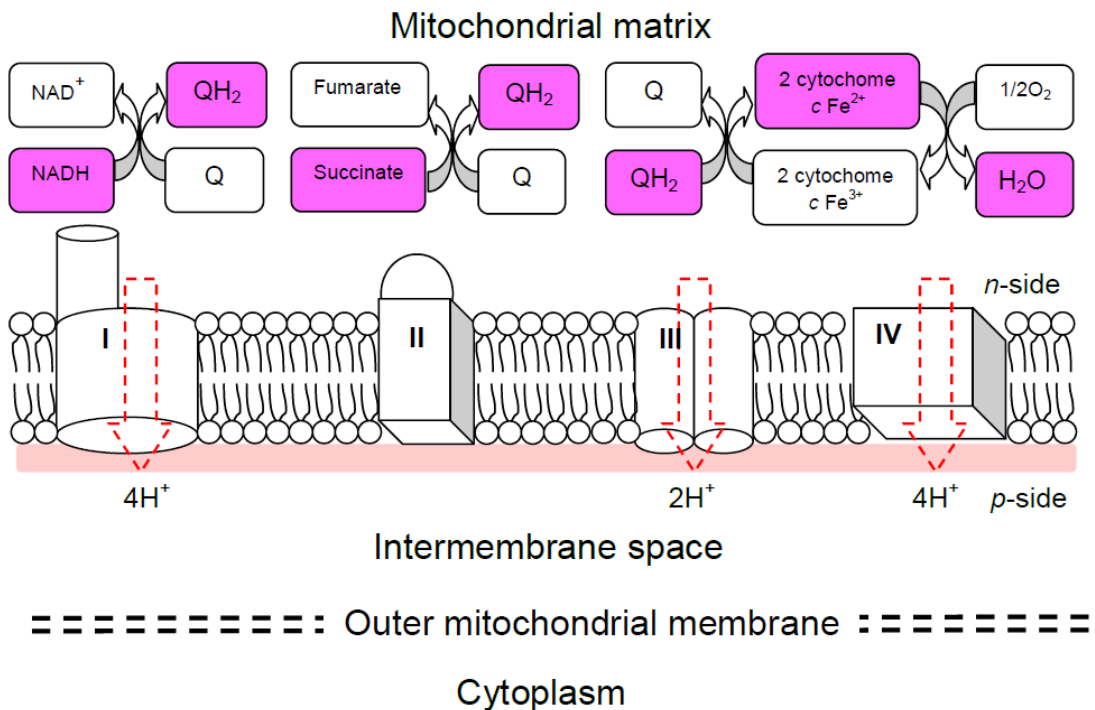
**Figure 1.5.7** shows mitochondrial energy converting enzymes which form a redox chain. These are the NADH:ubiquinone oxidoreductase (complex I), the succinate dehydrogenase (complex II), the cytochrome *bc*-complex (complex III) and the cytochrome *c* oxidase (complex IV). Each of these enzymes consists of several protein subunits. As mentioned above, only three complexes couple their redox-reactions with the transmembrane proton translocation. Succinate:ubiquinone oxidoreductase (complex II) just transduces reducing equivalents from succinate (Krebs cycle intermediate) to quinone. As shown in **Figure 1.5.6**, the difference between the standard reduction potentials of the respective redox pairs is small and does not allow the proton transport across the membrane against  $\Delta\tilde{\mu}_{H^+}$ .





**Figure 1.5.6. Standard reduction potentials of several carriers of reducing equivalents (at 25°C and pH 7).**

Names of the respiratory chain complexes performing each reaction are given on the right. Values of standard transformed reduction potentials  $E^0$  were taken from (Nelson and Cox, 2005). Value for the pair  $\text{FAD}/\text{FADH}_2$  is given for the free coenzyme; it can change upon binding to a protein.



**Figure 1.5.7. Scheme of the mitochondrial respiratory chain.**

The Roman numerals denote complexes of respiratory chain, the red arrows show the numbers of protons translocated across the membranes during each reaction shown above (i.e.  $2\bar{e}$ ). The reduced states are shown in magenta, the oxidized states are shown in white. The pink line on the  $p$ -side shows local pH decrease at the membrane surface induced by the proton translocation.

### 1.5.3.1. NADH:quinone oxidoreductase

The NADH:quinone oxidoreductase (complex I) is the largest mitochondrial membrane complex. Out of 14 core subunits corresponding to the subunits of bacterial enzymes, 7 are hydrophobic and in fact consist of  $\alpha$ -helices while the other 7 do not contain transmembrane regions but carry binding sites for the FeS-clusters, FMN (flavin mononucleotide) and NADH as a substrate. Altogether, these subunits assemble into a large L-shaped enzyme which could be divided into 3 modules:

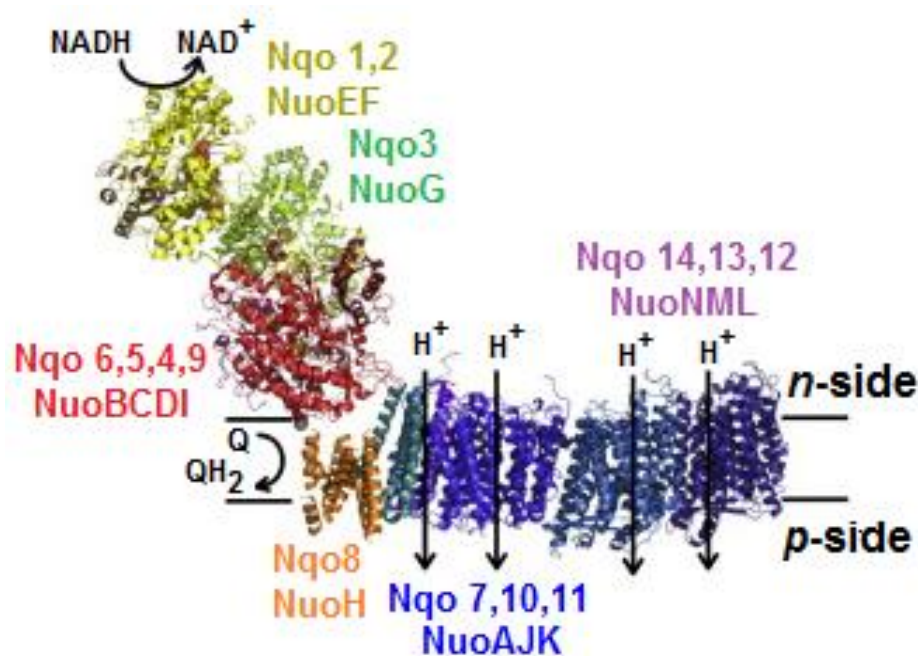
- the N-module oxidizes NADH;
- the Q-module reduces the quinone;
- the P-module performs proton translocation across the membrane.

The first two modules are located in the cytoplasmic part of the complex; the third is formed by the membrane subunits. These modules are not only used in a description of the function but they have a functional relation to several different enzymes.

The first structure of the hydrophilic domain (N- and Q-modules) was obtained for NADH:quinone oxidoreductase from *Thermus thermophilus* (Sazanov and Hinchliffe, 2006). The first structure of the P-module of NADH:quinone oxidoreductase from *E. coli* with the atomic resolution was obtained only in 2010 (Efremov *et al.*, 2010) and improved by the same team in 2011 (PDB ID 3RKO (Efremov and Sazanov, 2011)). The structure of mitochondrial complex I from *Yarrowia lipolytica* was also solved, but at a low resolution (Hunte *et al.*, 2010). The full atomic resolution structure of bacterial NADH:quinone oxidoreductase was finally solved in 2012 (Baradaran *et al.*, 2013; Efremov and Sazanov, 2012) and is shown in **Figure 1.5.8**.

NADH is a two-electron carrier, i.e. it simultaneously transfers two reducing equivalents. FMN, in turn, can accept and donate these reducing equivalents either simultaneously or in two steps. This property of FMN is of great importance as it can serve as an intermediate between NADH and FeS-clusters which accept and donate electrons one at a time. The chain of FeS-clusters between the FMN and the quinone is analogous to a "wire" and consists of 7 FeS-clusters.

All the subunits of the membrane part of the complex I, except NuoH, are depicted in the **Figure 1.5.9A**. Each of them contains a pair of transmembrane helices with the small loops in the middle formed by 5-7 amino acid residues including conserved prolines. In the **Figure 1.5.9B** these helices (7 and 12) are colored white and grey respectively. The conserved lysine residue is located in the beginning of the loops on the helix 7. It is in contact with the aspartate or glutamate residue from the helix 5. These pairs of residues together with the Glu-Tyr pair formed between the smaller subunits NuoK and NuoJ are assumed to mediate proton transfer.

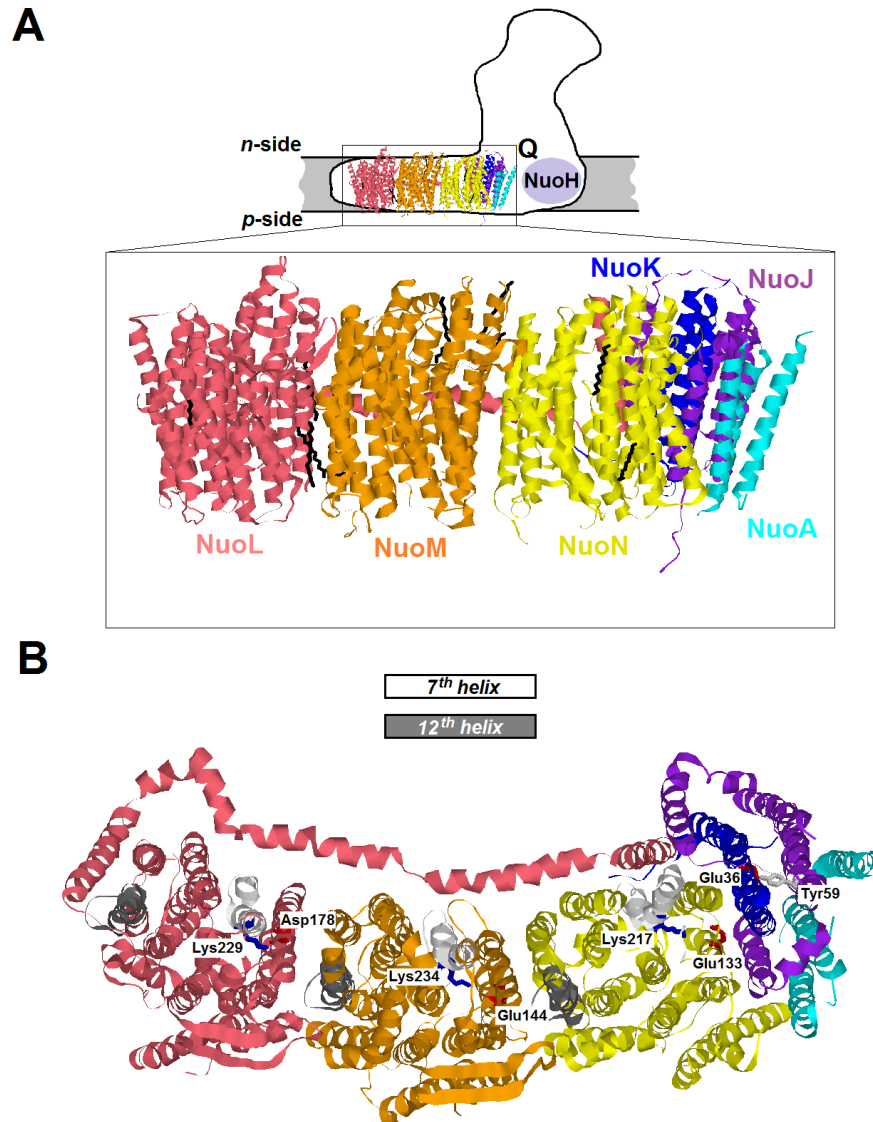


**Figure 1.5.8.** Structure of NADH:quinone oxidoreductase from *Thermus thermophilus*.

Figure taken from (Efremov and Sazanov, 2012). The transmembrane part of the complex is shown in more details in **Figure 1.5.9**.

The crystal structure supports the hypothesis that the proton transfer occurs as a result of conformational change upon quinone reduction but not during the electron transfer itself (Efremov and Sazanov, 2011). It was suggested that the lysine residues of helices 7 of the subunits NuoL, NuoM, and NuoN (see **Figure 1.5.9**) are protonated when the quinone is in the oxidized state. Reduction of the quinone leads to a conformation change which is transferred to all membrane subunits via long helix located alongside the membrane (part of

the subunit NuoL). Movement of the helices with irregular regions results in the transfer of the protons from these lysine residues to other polar residues on the *p*-side of the membrane.



**Figure 1.5.9.** Overall structure of the membrane part of NADH:quinone oxidoreductase from *E. coli*. (PDB 3RKO (Efremov and Sazanov, 2011)) (A) and the view on it from the *n*-side of the membrane (B). The helices 7 in subunits NuoL, NuoM and NuoN are colored white, the helices 12 of the same subunits are colored grey. These helices have irregular loops in the middle which make them mobile and flexible.

Complex I together with the cytochrome *bc* complex are believed to be the major sources of the reactive oxygen species in the mitochondrial respiratory chain (Balaban *et al.*, 2005). Reactive oxygen species, in turn, are among the causes of the Parkinson disease (Dawson and

Dawson, 2003) and aging (Skulachev, 2007). However, it appears that under physiological conditions, i.e. high concentrations of NADH, complex I does not produce much superoxide (Grivennikova and Vinogradov, 2006).

A decade ago, the first data on sodium translocation by NADH:quinone oxidoreductase from several bacteria including *E. coli* were published (Gemperli *et al.*, 2002). However, in the subsequent publications it was argued that the preparation could have contained an evolutionary unrelated membrane complex NQR in which the oxidation of NADH and reduction of the quinone are coupled with Na<sup>+</sup> translocation (Bertsova and Bogachev, 2004). Recently, a deactivated form of mitochondrial complex I was shown to function as a Na<sup>+</sup>/H<sup>+</sup>-antiporter (Roberts and Hirst, 2012).

**Evolution of NADH:quinone oxidoreductase.** We have already mentioned that different groups of subunits that comprise the functional core of the complex I are related to other proteins, thus the initial assembly of the complex in the course of evolution was possible from the precursors with separate original functions. This is also supported by the overall modular structure of NADH:quinone oxidoreductase (Friedrich *et al.*, 1993; Friedrich and Weiss, 1997). Below we discuss the evolutionary history of the modules (the gene names correspond to the *E. coli* enzyme).

**N-module.** Subunits NuoE and NuoF contain binding sites for the flavin, NADH and FeS-clusters. They are capable of NADH oxidation and of reduction of different possible acceptors under artificial conditions. The large subunit NuoG structurally resembles the iron-dependent hydrogenase and molybdopterin-containing proteins, formate dehydrogenase and nitrate reductase. The binding site for hydrogen or formate in NuoG is lost, but the rudimentary FeS-cluster N7 is sometimes still preserved.

**Q-module.** The pair of subunits NuoB and NuoD form the quinone-binding pocket. They resemble small and large subunits of the [NiFe]-hydrogenase, and the quinone-binding site is located at the same place as the active site of this protein (Brandt, 2006; Tocilescu *et al.*, 2010) (a detailed review on the evolution of dehydrogenases is not a subject of this thesis and is given elsewhere (Vignais and Billoud, 2007)).

**P-module.** Among seven transmembrane subunits the three largest subunits (NuoL, NuoM and NuoN) are related to several subunits of huge  $\text{Na}^+/\text{H}^+$ -antiporter complexes (Fearnley and Walker, 1992; Mathiesen and Hagerhall, 2002) (in different organisms they are called Mrp, Pha, Sha or Mnh (Swartz *et al.*, 2005)). These antiporters are widespread among prokaryotes and are usually used for pumping sodium ions out of the cells at the cost of proton influx. It turned out that the NuoK subunit also has a homologous protein in the Mrp complex denoted as MrpC (Mathiesen and Hagerhall, 2003). Probably NuoKLMN module was adopted from  $\text{Na}^+/\text{H}^+$ -antiporters.

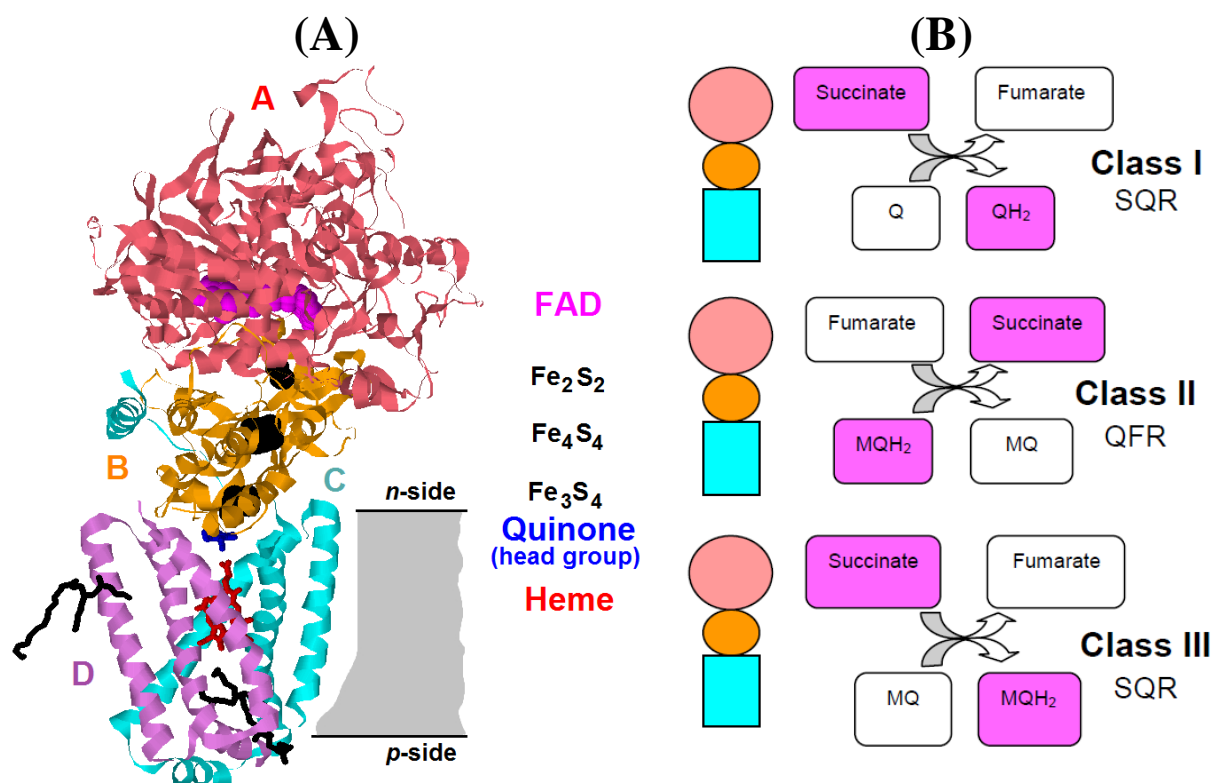
The subunit NuoH does not have close homologs with different function: all its homologs also belong to various membrane dehydrogenase complexes (Efremov and Sazanov, 2012). The cytoplasmic loops of these membrane proteins are well-alignable and contain conserved sites but the periplasmic loops can vary. NuoH is likely to provide coupling between the quinone reduction and the conformational change of the membrane part as its cytoplasmic loops in the bacterial complex take part in the formation of the quinone binding site. It has been suggested that this subunit does not perform transmembrane proton transfer itself (Efremov and Sazanov, 2012).

Alongside the typical 14-subunit complexes (the same set of subunits form the core of mitochondrial complex I) other variations are widely spread among prokaryotes. These complexes either lack the N-module or have it represented with a different protein type. 11-subunit complexes are obviously different from the hydrogenases of type 3 or type 4 and are likely to represent an ancestral form of complex I (Moparthi and Hagerhall, 2011). Phylogenetic tree based on peripheral subunits of the Q-module shows a separate clade of archaeal sequences with only 11- and 12-subunit complexes, thus presence of this type of complexes in the LUCA can be envisioned.

### 1.5.3.2. Succinate::quinone oxidoreductase

In contrast to other mitochondrial respiratory chain complexes, the succinate:quinone oxidoreductase (complex II) doesn't contribute to proton translocation across the membrane. It is also relatively small: only 4 subunits in boar mitochondria (PDB ID 1ZOY (Sun *et al.*,

2005), shown in **Figure 1.5.10A**). This structure is very similar to the *E. coli* succinate dehydrogenase crystallized before (PDB 1NEN (Yankovskaya *et al.*, 2003)). Even earlier, in 1999, C. R. D. Lancaster *et al.* have obtained a crystal structure of the fumarate reductase from *Wolinella succinogenes* with two hemes in the membrane part (PDB ID 1E7P (Lancaster *et al.*, 2001; Lancaster and Kroger, 2000; Lancaster *et al.*, 1999), **Figure 1.5.11**).

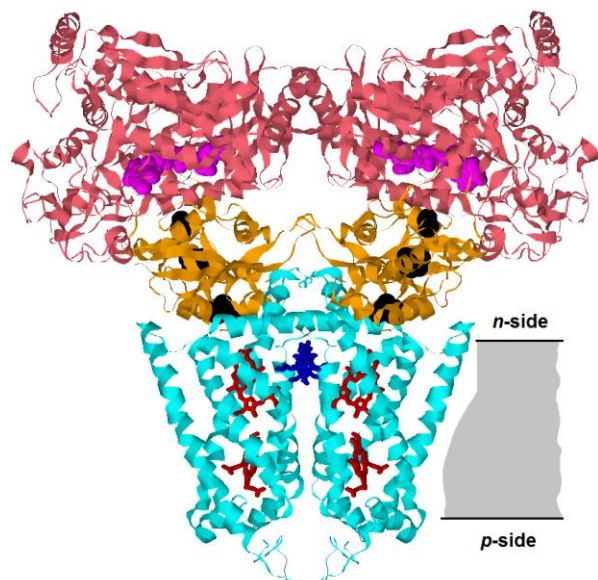


**Figure 1.5.10.** Overall scheme of succinate:quinone oxidoreductase from boar mitochondria (PDB ID 1ZOY (Sun *et al.*, 2005)) (A) and classification of succinate:quinone oxidoreductases by the type of catalyzed reactions (Hagerhall, 1997) (B).

Q denotes ubiquinone, MQ denotes menaquinone. Names of the enzymes come from the reactions they catalyze: SQR for succinate:quinone oxidoreductases or succinate dehydrogenases and QFR for quinol:fumarate oxidoreductases or fumarate reductases.

The detailed review on the succinate dehydrogenases and related enzymes was published by Hägerhäll (Hagerhall, 1997). The succinate dehydrogenase also seems to be a modular complex. The subunit A binds the substrate (succinate) and flavin adenine dinucleotide (FAD). The small subunit B binds three FeS-clusters which form a short "wire" between FAD and the membrane part of the complex. Membrane subunits C and D anchor subunits A

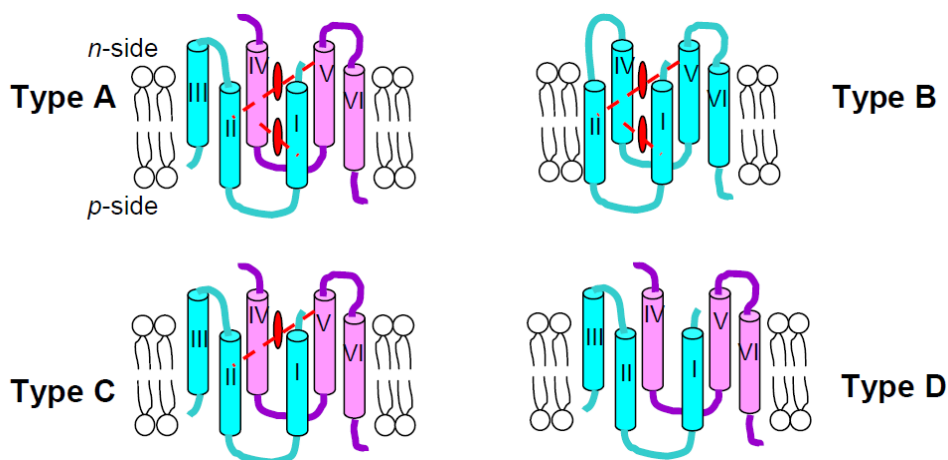
and B to the membrane. They show little sequence similarity between different species (Hagerhall, 1997).



**Figure 1.5.11. Organization of the fumarate reductase from *Wolinella succinogenes*.**

PDB 1E7P, (Lancaster *et al.*, 2001; Lancaster and Kroger, 2000; Lancaster *et al.*, 1999). Parts of the complex are colored as in **Figure 1.5.10**.

One of suggested classifications of succinate:quinone oxidoreductases is based on features of the membrane part of the complexes (Hagerhall and Hederstedt, 1996) (**Figure 1.5.12**). All of them have highly similar architecture: 4 transmembrane helices are inclined to the plane of the membrane and use histidine residues to coordinate a pair of hemes or one heme. Some complexes even do not contain hemes. The heme located at the *n*-side of the membrane is the high-potential one ( $b_H$ ) and is coordinated by histidines from helices II and V. In the complexes with the second low-potential heme ( $b_L$ ) located closer to *p*-side, its coordination is achieved by histidines from and helices I and IV.



**Figure 1.5.12. Four types of membrane part of SQR and FQR complexes (adapted from (Hagerhall, 1997)).**



Complexes without the transmembrane part are known in addition to the four types shown in **Figure 1.5.12** (Moll and Schafer, 1991). They contain two different hydrophilic subunits. Although this type of complex was first purified from crenarchaeon *Sulfolobus acidocaldarius*, it was identified afterwards in various bacteria (Lemos *et al.*, 2002). These enzymes can be also classified into three groups based on the chemical reaction they catalyze (**Figure 1.5.10B**). Class I is the most ubiquitous and includes complex II of the mitochondrial respiratory chain. Members of this class use high-potential acceptor quinone (ubiquinone), thus making succinate oxidation slightly energetically favorable. These enzymes are called succinate:quinone oxidoreductases (SQR). Other enzymes however operate with the low-potential quinones such as menaquinone and are only able to catalyze the reverse reaction, the reduction of fumarate. These enzymes belong to class II and are called quinol:fumarate oxidoreductases (QFR). Finally, enzymes from class III are able to somehow catalyze energetically unfavorable oxidation of succinate with menaquinone. They are found in several gram-positive bacteria (i.e. *Bacillus subtilis*).

**Evolution of succinate dehydrogenases.** Complexes similar to complex II have common organization and are widespread in bacteria, archaea and eukaryotes. This notion argues for the presence of their ancestor in the LUCA (Hagerhall, 1997) but cannot be considered as a proof: the same distribution could result if the complex would have been invented in one of the domains and then laterally spread to others. Phylogenetic trees of the water soluble subunits A and B do not show a specific separation of archaeal sequences and can be considered as an argument in favor of the lateral distribution of these proteins (Lemos *et al.*, 2002). The dimeric structure of the complex and overall organization of the membrane part is closely similar to those of cytochrome *bc*-complex (**Figure 1.5.11**), so that possible functional parallels were suggested (Pereira and Teixeira, 2003).

### 1.5.3.3. Quinol:cytochrome *c* oxidoreductase (cytochrome *bc* complex)

Cytochrome *bc* complexes (electrogenic quinol:cytochrome *c* oxidoreductases) play key roles in photosynthesis (cytochrome *b<sub>6</sub>f*-complex) and respiration (cytochrome *bc<sub>1</sub>* complex, mitochondrial complex III). These enzymes catalyze electron transfer from diverse

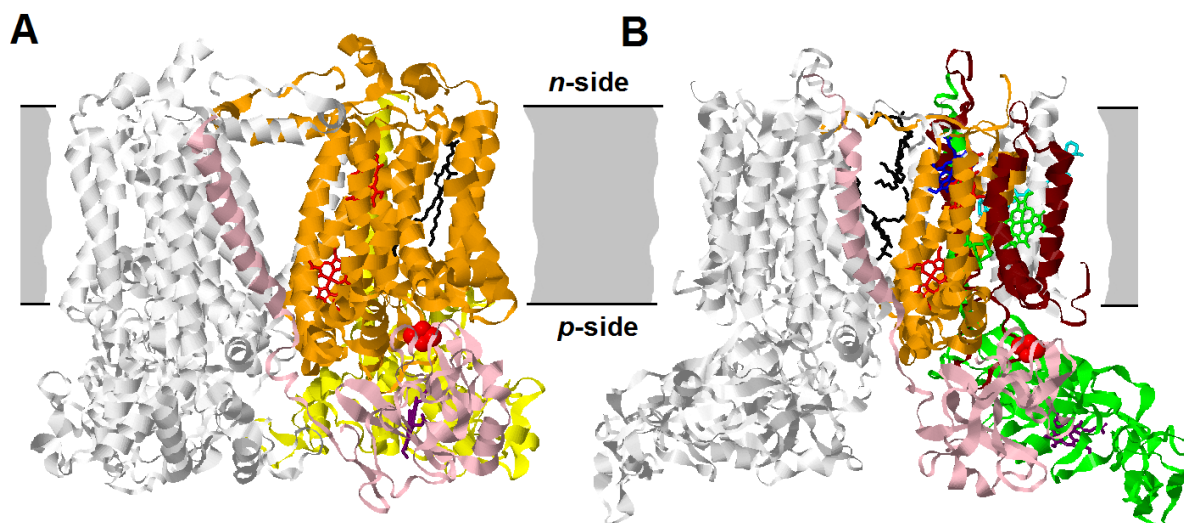
membrane quinols to high-potential redox carriers (routinely *c*-type cytochromes) and use the energy released to translocate protons across energy-converting membranes (Al-Attar and de Vries, 2012; Berry, 2002; Cramer *et al.*, 2011; Mulkidjanian, 2010).

The cytochrome *bc*<sub>1</sub> complex is an intertwined dimer (typical crystal structure is represented with proteobacterial cytochrome *bc*<sub>1</sub>-complex, PDB ID 2FYN (Esser *et al.*, 2006), **Figure 1.5.13A**). The catalytic core of each *bc*<sub>1</sub> monomer is formed by three subunits: the membrane-embedded cytochrome *b*, the [Fe<sub>2</sub>S<sub>2</sub>] cluster-carrying iron-sulfur Rieske protein, and cytochrome *c*<sub>1</sub>. The catalytic, hydrophilic domains of the two latter subunits are anchored in the membrane by single hydrophobic  $\alpha$ -helices. Each cytochrome *b* is a bundle of 8  $\alpha$ -helices that binds two (proto)hemes, one close to the p-side of the membrane (*b*<sub>p</sub>), and other close to the *n*-side of the membrane (*b*<sub>n</sub>). Because heme *b*<sub>p</sub> usually has a lower midpoint redox potential than heme *b*<sub>n</sub>, the two hemes are also denoted as the low- and high-potential hemes (*b*<sub>l</sub> and *b*<sub>h</sub>, respectively).

Proteobacteria, e.g. *Rhodobacter capsulatus*, have cytochrome *bc* complexes of only three subunits, but mitochondrial cytochrome *bc* complexes have additional, presumably, regulatory subunits (Gao *et al.*, 2003). It is noteworthy that the X-ray structure of the simplest *bc*<sub>1</sub> of *R. capsulatus*, which contains only 3 subunits, matches the structure of three catalytic subunits of mitochondrial *bc*<sub>1</sub> (Berry and Huang, 2003).

The cytochrome *b*<sub>6</sub>*f* complexes of green plants and cyanobacteria, although evolutionarily related to the cytochrome *bc*<sub>1</sub> complexes (Widger *et al.*, 1984), differ structurally from them (Cramer *et al.*, 2006; Kurisu *et al.*, 2003; Stroebel *et al.*, 2003) (**Figure 1.5.13B**). Specifically, the cytochrome *b* of the *bc*<sub>1</sub> (formed by 8 transmembrane helices) corresponds to two subunits of the *b*<sub>6</sub>*f*: the N-terminal part of the cytochrome *b* is homologous to the cytochrome *b*<sub>6</sub> (4 transmembrane helices), whereas the first 3 transmembrane helices of C-terminal part resemble the subunit IV of the cytochrome *b*<sub>6</sub>*f* complex. Thus, cytochrome *b* of cytochrome *bc* complex from *R. sphaeroides* (and also mitochondrial complex III) has 8 transmembrane helices while cytochrome *b*<sub>6</sub>*f*-complex contains only 7 helices in the corresponding part. The cytochrome *b*<sub>6</sub> subunit, besides accommodating two *b*-type hemes, as in the *bc*<sub>1</sub>, carries an additional *c*-type heme (denoted *c*<sub>n</sub> or *c*<sub>i</sub>) that does not have a counterpart in the *bc*<sub>1</sub> (Stroebel *et al.*, 2003). The iron atom of this heme is connected to the

propionate of heme  $b_n$  by a water bridge. The subunit IV binds single molecules of chlorophyll  $a$  and  $\beta$ -carotene, respectively, which are likely to be involved in photoprotection (Dashdorj *et al.*, 2005; Pierre *et al.*, 1997). In addition, the cytochrome  $f$ , which accepts electrons from the mobile FeS domain of the Rieske protein, although it carries a  $c$ -type heme, is structurally unrelated to the cytochrome  $c_1$  of the  $bc_1$  (Martinez *et al.*, 1994).



**Figure 1.5.13.** Subunit composition of cytochrome  $bc$  complex from *Rhodobacter sphaeroides* (PDB ID 2FYN (Esser *et al.*, 2006)) (A) and cytochrome  $bf$ -complex from *Mastigocladus laminosus* (PDB ID 1VF5 (Kurusu *et al.*, 2003)) (B).

Only one monomer in the complex is colored. The Rieske protein from the second monomer is shown in pink, its FeS-cluster is highlighted with red, the cytochromes  $b$  and  $b_6$  are shown in orange, and the subunit IV is colored brick-red. Quinones or inhibitors are not shown.

The cytochrome  $c_1$  is shown in yellow, the cytochrome  $f$  is shown in green.

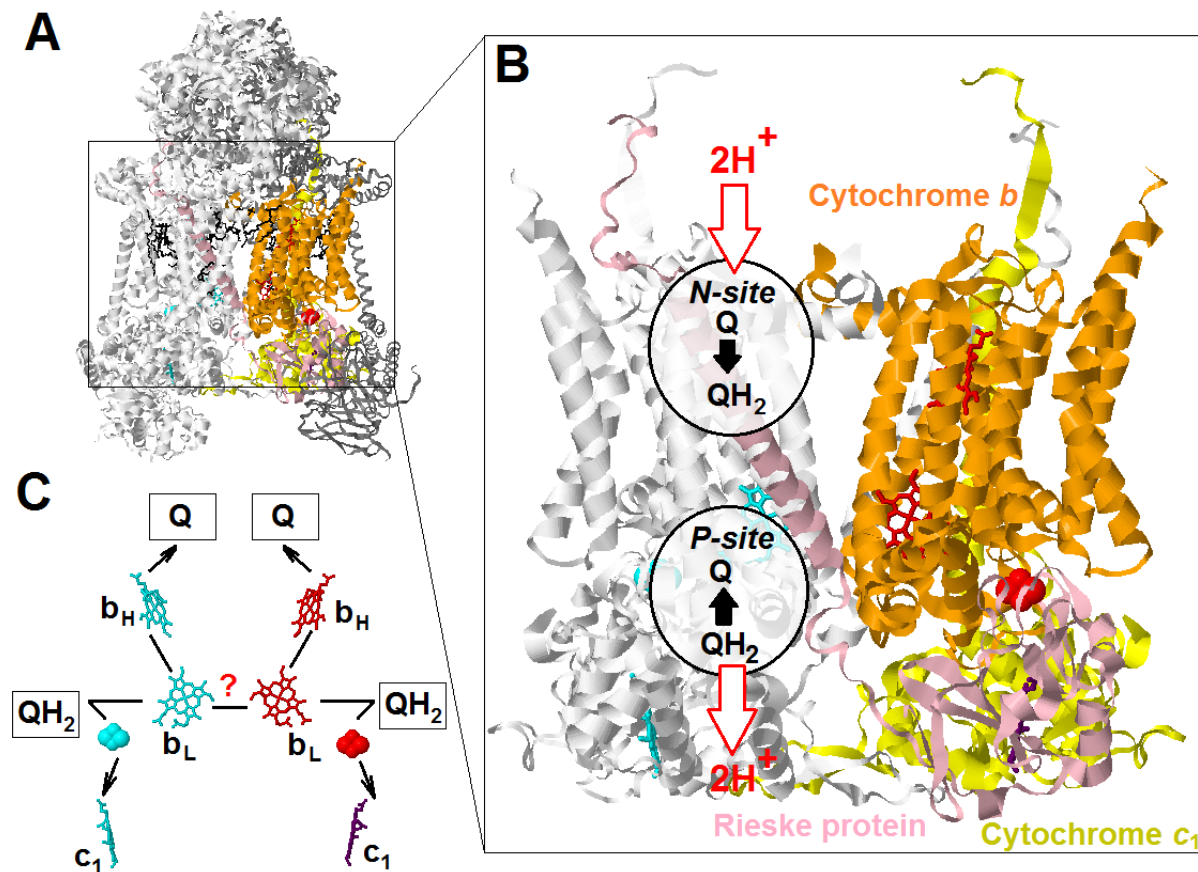
The black color depicts lipid molecules found in the crystal structures of cytochrome  $bc$  complexes, the  $\beta$ -carotene is shown in cyan.

The color code for the porphyrin-containing compounds is as follows: the hemes  $b_H$  and  $b_L$  are coloured red, chlorophyll  $a$  is green, additional heme  $c_n$  is shown in dark blue, the heme in cytochrome  $c_1$  or cytochrome  $f$  is shown in violet.

The large cavity between the monomers in the complex is assumed to be required for the exchange of the quinone/quinol substrates. In natural membranes this cavity is likely to be filled also with the lipid tails (Kurusu *et al.*, 2003). Various lipids are tightly bound to the cytochrome  $bc$  complex: for instance, in the crystal structure of the yeast complex III, 11

lipid molecules co-crystallized with the protein dimer. These lipid molecules could be important for the assembly of the cytochrome *bc* complexes (Hasan *et al.*, 2011).

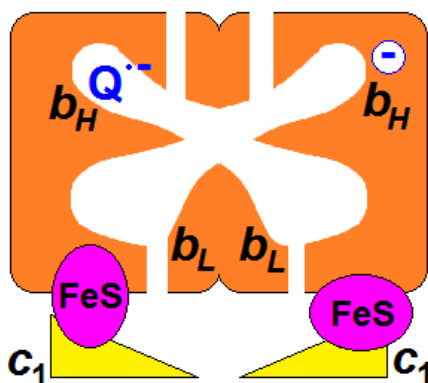
The operation of the cytochrome *bc* complexes can be described by the Mitchell's Q-cycle mechanism (Mitchell, 1975; Mitchell, 1976). According to this scheme, a substrate ubiquinol molecule gets oxidized close to the *p*-side of the membrane, in the so-called catalytic center *P*, at the interface between cytochrome *b* and the [Fe<sub>2</sub>S<sub>2</sub>] cluster-carrying domain of the Rieske protein (**Figure 1.5.14**). Upon this oxidation, one electron is accepted by the FeS domain to be transferred to the further high-potential electron acceptors, whereas the other electron, via heme *b<sub>p</sub>* and heme *b<sub>n</sub>*, crosses the bilayer and reduces an ubiquinone molecule in the catalytic center *N* close to the opposite side of the membrane.



**Figure 1.5.14.** Overall view of the yeast mitochondrial complex III (PDB 3CX5 (Solmaz and Hunte, 2008)) (A), three functionally important subunits corresponding to bacterial homologs (B) and the scheme of the electron transfer between the cofactors (C).

As a result, one ubiquinol molecule  $Q_NH_2$  is produced in center(s)  $N$  per each two molecules of substrate ubiquinol  $Q_PH_2$  that are oxidized in center(s)  $P$ . Since the nascent  $Q_NH_2$  molecules can be also oxidized in center(s)  $P$ , two protons are ultimately translocated across the membrane per each electron that passes through the  $bc_1$ .

The cytochrome  $bc_1$  complexes were shown to be functional dimers, capable of electron exchange between the monomers via closely placed heme  $b_p$  (Gopta *et al.*, 1998; Mulkidjanian, 2007; Swierczek *et al.*, 2010). In the  $bc_1$ -type complexes, under physiological conditions of a half-reduced ubiquinone pool, a total of two electrons seem to be continuously present in the dimeric cytochrome  $b$  moiety, owing to an electron equilibration with the membrane quinol pool via centers  $N$ , see (Berry *et al.*, 2000; Mulkidjanian, 2007) and references therein. Because each center  $N$  is "preloaded" by an electron (**Figure 1.5.15**), the oxidation of each quinol molecule in center  $P$  results in an immediate quinol formation in the respective center  $N$  (Mulkidjanian, 2007; Mulkidjanian, 2010).



**Figure 1.5.15.** Physiological, activated state of the cytochrome  $bc$  complex (adapted from (Mulkidjanian, 2007)) with two electrons stored in the cytochrome  $b$  moieties.

**Current views on the evolution of cytochrome  $bc$  complexes.** One of the questions on the history of the cytochrome  $bc$  complexes follows from the differences between the cytochrome  $b_6f$ -complex and the cytochrome  $bc_1$ -complex. Is the long version of cytochrome  $b$  a result of the fusion between the cytochrome  $b_6$ -type with subunit IV or, *vice versa*, the cytochrome  $b$  has split into two parts?

The cytochrome  $bc$  complexes are spread among all the three domains of life but are not ubiquitous in contrast to the rotary membrane ATPases. An apparent similarity between the

first phylogenetic tree of cytochrome *b* and the trees constructed on the base of other phylogenetic markers (e.g. 16S rRNA) led to the conclusion that lateral gene transfers of the cytochrome *bc* complexes were very rare (Lebrun *et al.*, 2006; Schutz *et al.*, 2000). As the genes of a cytochrome *bc* complex were found in the archaeon *Sulfolobus acidocaldarius* and were claimed to be functional (Schafer, 1996), it has been suggested that this complex could have been present in the LUCA (Castresana and Moreira, 1999). An alternative hypothesis suggests that archaeal proteins could be a result of lateral gene transfer from bacteria (Furbacher *et al.*, 1996) but until now this hypothesis was claimed not to be supported by phylogenetic trees (Lebrun *et al.*, 2006; Schutz *et al.*, 2000). The only acknowledged exception from the vertical inheritance rule was the cytochrome *bc* complex of *Aquifex aeolicus*, obviously transferred from proteobacteria (Schutz *et al.*, 2000). Other deviations of the cytochrome *b* trees from the taxonomical tree (for instance, actinobacterial sequences do not cluster with firmicutes despite these two bacterial groups being considered to be close relatives (Olsen *et al.*, 1994)) were explained by the mistakes in taxonomy (Lebrun *et al.*, 2006).

A recent paper described a "green clade" on the phylogenetic tree of cytochrome *bc* complexes, named after photosynthetic organisms from different phyla (*Chlorobi*, *Cyanobacteria*, *Heliobacteria*) (Nitschke *et al.*, 2010). The authors suggest that the appearance of this clade was coupled with such evolutionary events as the split of cytochrome *b* and reduction in the number of transmembrane helices. They suggested that the LUCA possessed a large cytochrome *b* of 8 transmembrane helices, which then split in some bacterial lineages (Nitschke *et al.*, 2010).

The *c*-type cytochromes in the *b<sub>6f</sub>*-type and *bc<sub>1</sub>*-type complexes seem to be unrelated and probably missing in the ancestral version of the cytochrome *bc* complex (Furbacher *et al.*, 1996). The same is suggested by the sequences of the *Chlorobium limicola* cytochrome *bc* complex: it has a number of features intermediate between cytochrome *b<sub>6f</sub>*-complexes and cytochrome *bc<sub>1</sub>*-complexes (Schütz *et al.*, 1994). Cytochrome *b* in this complex is long, but it contains only 7 helices. In the fourth helix (D), 14 residues separate two heme-binding histidines (in contrast to 13 residues found in mitochondria and proteobacteria). These two properties were previously suggested to be typical for the *b<sub>6f</sub>*-complexes (Widger *et al.*,

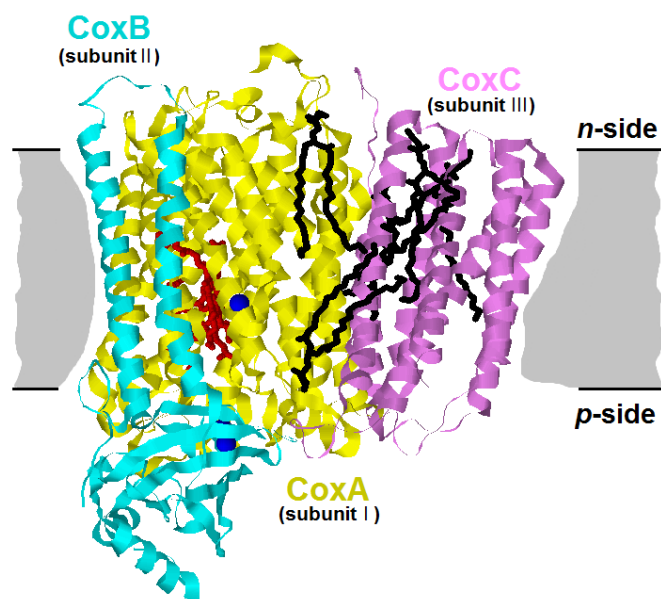
1984). Interestingly, cytochrome  $c_1$  is missing from this complex, its function seemingly is performed by a quite different cytochrome  $c$ -551 (Schütz *et al.*, 1994).

#### 1.5.3.4. Cytochrome $c$ -oxidase

Respiratory chain in mitochondria ends with the cytochrome  $c$  oxidase (complex IV) that reduces oxygen to water (Brzezinski and Gennis, 2008). The homologous bacterial cytochrome oxidase from *R. sphaeroides* contains only three subunits (**Figure 1.5.16**, PDB ID 1M56 (Svensson-Ek *et al.*, 2002)). Eukaryotic complexes usually contain additional subunits (for instance, the bovine complex IV is composed of 13 subunits). Catalytic core of the complex is formed by the subunits I and II which maintain all redox centers whereas other subunits, especially the strictly conserved subunit III, are likely to have regulatory function.

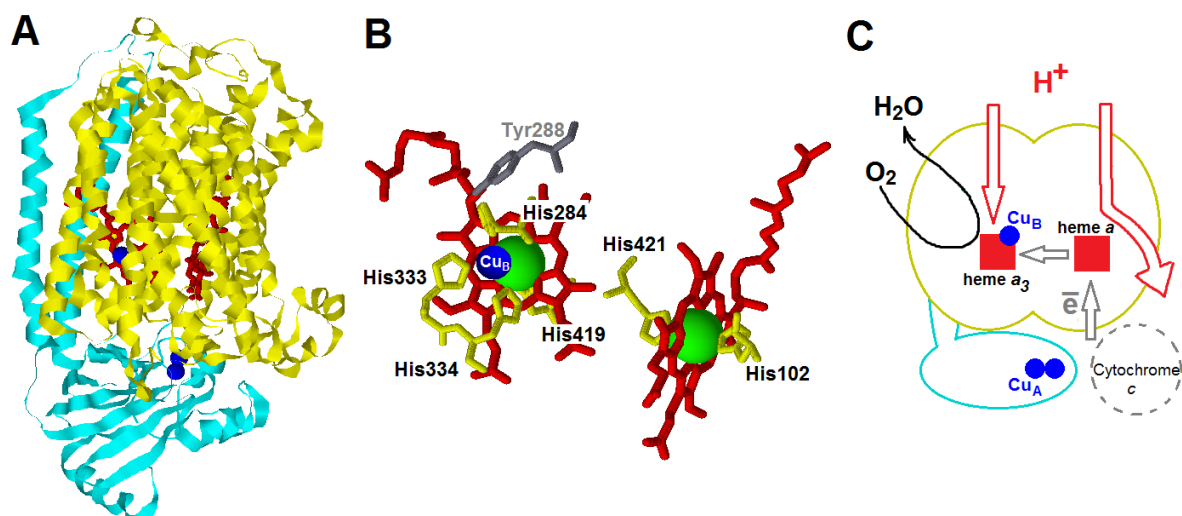
The main cofactors of the cytochrome oxidase are the pair of hemes located approximately at the same depth inside the membrane. The hemes can belong to different types, and classification of cytochrome oxidases relies on the chemical nature of the hemes. For example, mitochondrial cytochrome oxidase has both hemes of type  $a$  and thus is denoted as the  $aa_3$  type cytochrome oxidase. The numeral "3" is used because of the historical reasons to mark the heme which binds oxygen and forms a complex (a heme-copper site) with the copper ion  $\text{Cu}_B$ . **Figure 1.5.17B** shows residues which bind cofactors of the enzyme. Iron of the low-potential heme  $a$  (to the right) is coordinated by two histidine residues while the heme-copper site (to the left) is surrounded by four histidines in total (one histidine residue binds iron of the heme and the three other residues bind a copper ion).

The mechanism of this enzyme is still not fully understood. However, it is accepted that upon reduction of one oxygen molecule, four protons are collected from the  $n$ -side of the membrane to form water and additional four protons are translocated (pumped) across the membrane. The heme-copper site is responsible for the complex mechanism of oxygen reduction (this process is reviewed in details elsewhere (Brzezinski and Gennis, 2008)).



**Figure 1.5.16.** Structure of the cytochrome *c* oxidase from *Rhodobacter sphaeroides* (PDB 1M56 (Svensson-Ek *et al.*, 2002)).

Subunit I is shown in yellow, subunit II is shown in cyan, subunit III is shown in violet. Copper ions are shown as blue spheres with radii corresponding to their size, hemes are colored red. Lipid molecules are colored black.



**Figure 1.5.17.** Catalytic core of the cytochrome *c* oxidase from *Rhodobacter sphaeroides* (PDB 1M56 (Svensson-Ek *et al.*, 2002)) rotated by 90° in relation to the view shown in **Figure 1.5.16** (A), the residues that coordinate the heme *a* and the heme-copper site (B) and the overall scheme of the reactions in the cytochrome oxidase (C).

Modern classification of these enzyme complexes is based, mostly, on the type of the hemes and allows distinguishing three major groups: A (further separated into A<sub>1</sub> and A<sub>2</sub>), B and C



(Sousa *et al.*, 2011). In addition to these closely related complexes and their homolog the NO-reductase, a different family of membrane enzymes is known also to perform oxygen reduction: those are denoted as cytochrome *bd*-oxidases (Watanabe *et al.*, 1979).

**Evolution of cytochrome oxidase.** The first comparative analysis of cytochrome oxidase subunits was performed back in 1994 (Castresana *et al.*, 1994). The presence of the major subunits of this complex in archaea (sequences from *Sulfolobus acidocaldarius* and *Halobacterium halobium* were included in the study) and their separate position on the phylogenetic tree allowed the authors to conclude that this enzyme (or even a number of such enzymes) were already present in the LUCA. Gennis and Hemp, however, argued recently that phylogenetic analysis clearly shows cytochrome oxidases as a (cyano)bacterial invention, followed by their lateral transfer to archaea (Hemp and Gennis, 2008; Hemp *et al.*, 2012). A separate positioning of the archaeal sequences on phylogenetic trees these authors explained by the fast evolution of the complexes in unusual environment of the archaeal-type membrane.

The presence of cytochrome oxidase in the LUCA is, generally, unlikely. First, it is unlikely that an oxygen-reducing enzyme could emerge before the appearance of atmospheric oxygen (which happened some 2.5 Gy ago as a result of oxygenic photosynthesis in cyanobacteria). Second, the key cofactors in complex IV are copper ions, but the environments of the anoxic Earth were likely to contain only monovalent copper  $\text{Cu}^+$ , all common salts of which are insoluble in water (Ochiai, 1978; Ochiai, 1983). Concentration of the free  $\text{Cu}^+$  under anoxic conditions was estimated to be  $10^{-17}$  M which is several orders of magnitude less than for other metals (Ochiai, 1978). Only after oxygenation of the atmosphere the soluble  $\text{Cu}^{2+}$  could be recruited by enzymes (Ochiai, 1983).

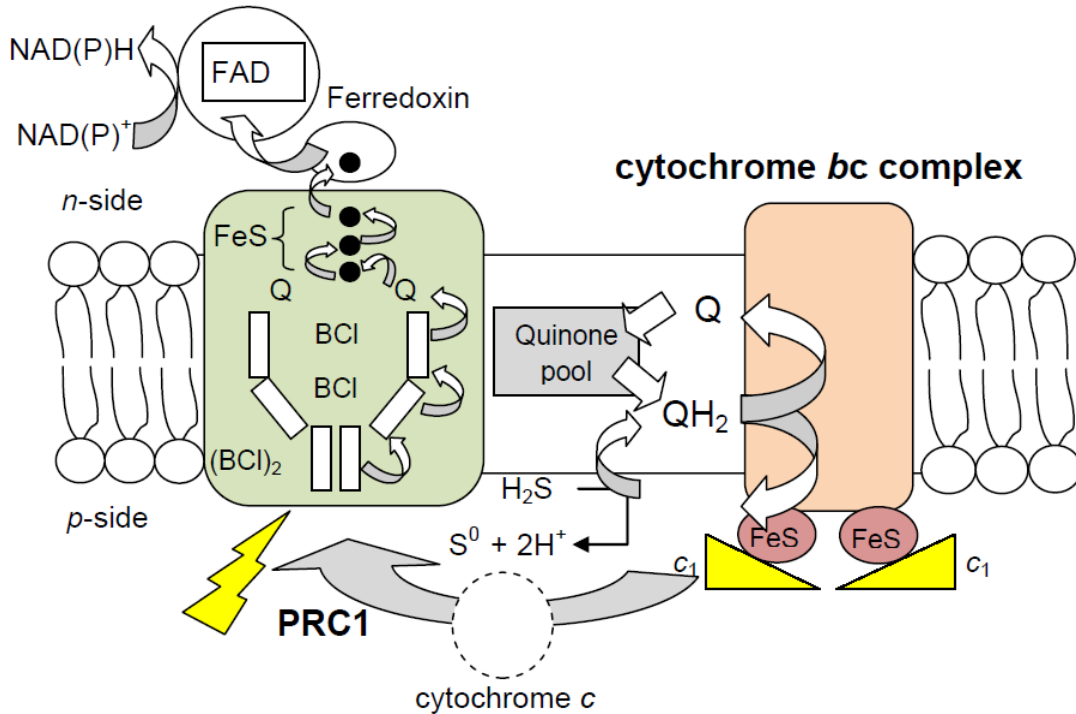
### 1.5.4. Photosynthesis

Chlorophyll-based photosynthesis is a mechanism of converting solar light energy into chemical energy (in the form of reduced cofactors, such as NAD(P)H or quinols) and into the transmembrane proton potential. Green plants and cyanobacteria use water as a reductant for CO<sub>2</sub> and perform the *oxygenic photosynthesis*, as oxygen is produced as a byproduct of water decomposition by reaction  $2\text{H}_2\text{O} \rightarrow 4\text{H}^+ + \text{O}_2 + 4\bar{e}$ . Other bacteria are capable of utilizing H<sub>2</sub>S, H<sub>2</sub>, and Fe<sup>2+</sup> as electron donors upon the *anoxygenic photosynthesis*. We will not focus closely on the structure and functioning of the photosynthetic complexes themselves (for recent reviews see (Hohmann-Marriott and Blankenship, 2011; Rutherford *et al.*, 2012)) but will rather discuss the involvement of the *bc*-complexes in this process.

While cyanobacteria and chloroplasts of green plants use complex photosynthetic machinery comprising two photosynthetic complexes, namely photosystem I and photosystem II, other phototrophic bacteria use only one light-processing complex, named Photosynthetic Reaction Center (PRC), which could be related either to photosystem I or to photosystem II.

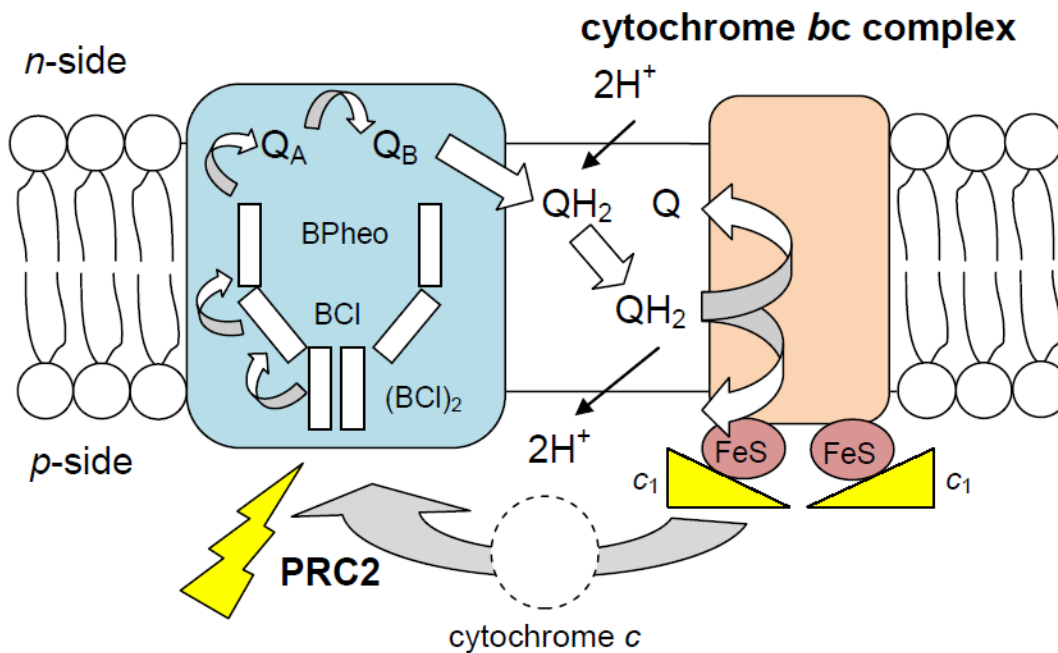
- Phototrophic representatives of *Heliobacteria* (*Firmicutes*), *Chlorobi* and acidobacteria contain PRCs similar to photosystem I (PRC1, sometimes named FeS-type PRCs). Here iron-sulfur clusters serve as ultimate acceptors of electrons (**Figure 1.5.18**). Such PRCs are capable of direct reduction of NAD(P)<sup>+</sup> to NAD(P)H.
- Phototrophic *Chloroflexi* and purple bacteria (for instance, *Rhodobacter* species) have PRCs similar to photosystem II (PRC2 or Q-type PRC) with quinones as ultimate acceptors of electrons (**Figure 1.5.19**).

The photosynthetic system of cyanobacteria and plants combines both types of PRCs and an oxygen-evolving (water-splitting) complex (**Figure 1.5.20**).



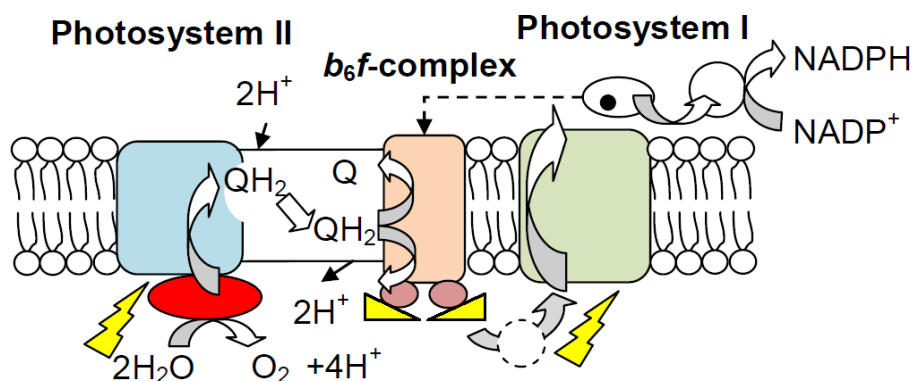
**Figure 1.5.18.** Scheme of the reactions in the PRC1 of *Chlorobium*.

(BCl)<sub>2</sub> – special pair of bacteriochlorophyll molecules, BCl – bacteriochlorophyll, Q – quinone, black dots – FeS-clusters.



**Figure 1.5.19.** Scheme of the reactions in the PRC2 of proteobacteria.

(BCl)<sub>2</sub> – special pair of bacteriochlorophyll molecules, BCl – bacteriochlorophyll, BPheo – bacteriopheophytin, Q – quinone, black dots – FeS-clusters.



**Figure 1.5.20.** Major components of the oxygenic photosynthesis chain: Photosystem II (blue), *b<sub>6</sub>f*-complex (orange) and Photosystem I (green). Water-splitting (and oxygen-evolving) complex is shown in red.

**Involvement of the cytochrome *bc* complexes in the photosynthesis.** As shown in **Figure 1.5.20**, the cytochrome *b<sub>6</sub>f*-complex is directly involved both in the linear electron transport between photosystem II and photosystem I and in the cyclic electron transfer around Photosystem I (Cramer *et al.*, 2006). In *Heliobacteria* the *b<sub>6</sub>f*-type complex is coded by the "photosynthetic" operon (together with the PRC1 and proteins responsible for chlorophyll and carotenoid biosynthesis) (Xiong *et al.*, 1998) and, apparently, is involved in the reduction of the electron vacancy in the PRC via *c*-type cytochromes (Nitschke *et al.*, 1995) in a kind of cyclic electron transfer, see **Figure 1.5.18**. The same function, most likely, is performed by the *bc<sub>1</sub>*-type complex in *Chlorobi* (Tsukatani *et al.*, 2008). Phototrophic purple bacteria use their cytochrome *bc<sub>1</sub>*-complex to cycle electrons from quinol back to the oxidized bacteriochlorophyll of the PRC2 (**Figure 1.5.19**) (Crofts and Wraight, 1983; Joliot *et al.*, 2005; Vermeglio and Joliot, 1999). In photosynthetic *Chloroflexi* the cytochrome *bc*-complex was not initially identified; instead an unrelated multisubunit complex (termed "alternative complex III") was discovered and purified from bacteria *Chloroflexus aurantiacus* (Yanyushin, 2002). Since the presence of this complex in most bacterial genomes "anticorrelates" with the presence of cytochrome *bc* complexes, this newly discovered complex was suggested to functionally replace cytochrome *bc* complex (Yanyushin *et al.*, 2005). Still a cytochrome *bc* complex was recently identified in the complete genome of *Nitrolancetus hollandicus* that belong to *Chloroflexi* (Sorokin *et al.*, 2012).

In sum, the cytochrome *bc* complexes (either of the *b<sub>6</sub>f*-type or of the *bc<sub>1</sub>*-type) are functionally coupled with photosynthetic reaction centers in all bacterial lineages.

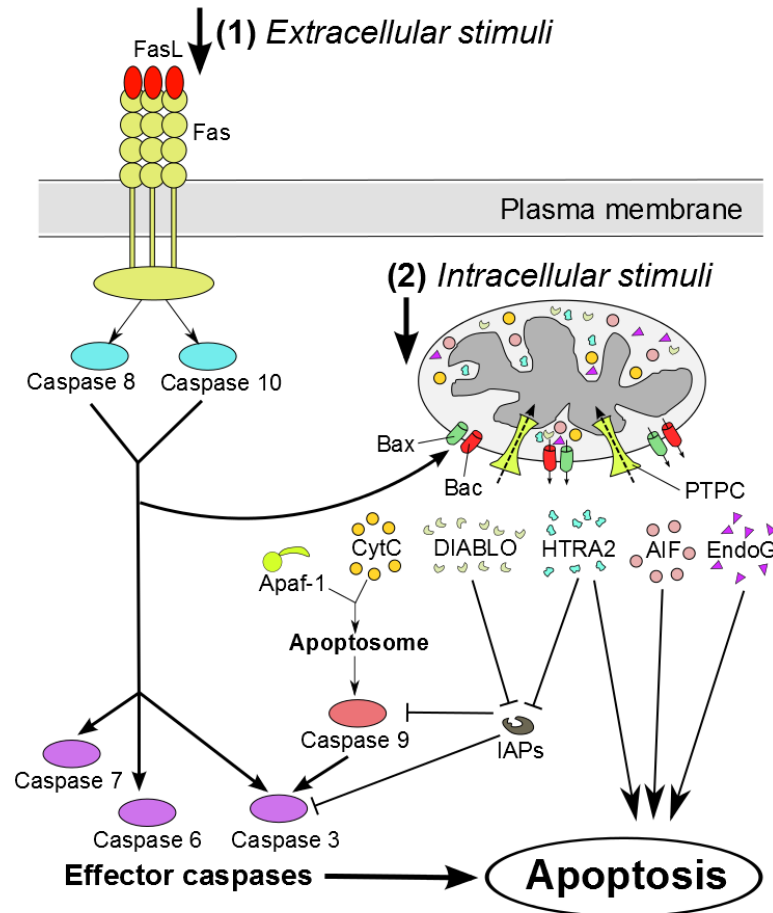
### 1.5.5. Role of reactive oxygen species and the cytochrome *bc*<sub>1</sub>-complex in apoptosis

Recently the *bc*<sub>1</sub>-complex has been suggested to serve as a putative trigger of apoptosis in animal cells (Skulachev *et al.*, 2009). Apoptosis is a mechanism of programmed cell death, which is widespread in eukaryotic organisms and well studied in *Metazoa* (Kerr, 1965; Kerr *et al.*, 1972; Lockshin and Williams, 1965; Wyllie *et al.*, 1980). For the past five decades this mechanism was extensively studied, as it turned out to be of key importance for numerous processes, including but not restricted to organism development, immune system response and aging. We are not aiming to present a thorough review of apoptosis in this section, but will briefly describe its major steps and focus on the poorly studied intrinsic apoptotic pathway, in which the enzymes of respiratory chain and in particular the cytochrome *bc* complex seem to be among the key players.

Cleavage of DNA is one of the results of apoptosis (Canman *et al.*, 1992). This step is preceded by activation of a cascade of self-cleaving cytoplasmic cysteine-dependent aspartate-directed proteases (Lazebnik *et al.*, 1994) termed caspases (Alnemri *et al.*, 1996) (for a review see (Earnshaw *et al.*, 1999)). As a result, they activate DNases which degrade DNA during apoptosis (Enari *et al.*, 1998).

The caspase cascade can be activated by extracellular stress signals that are sensed and propagated by specific transmembrane receptors; such process is sometimes called an "*extrinsic apoptosis*" (Galluzzi *et al.*, 2012; Siegel *et al.*, 2000; Wajant, 2002) (**Figure 1.5.21**). This pathway typically starts with the binding of a ligand to the receptor in the plasma membrane of the cell, which leads (through a complex cascade of caspase self-cleavage/self-activation) to the activation of the effector caspases, which directly cleave the cellular compounds, such as DNA and cytoskeleton, resulting in the cell death. In many cases, however, the activated caspase 8 triggers the Mitochondrial Outer Membrane Permeabilization (MOMP) (Kroemer *et al.*, 2007). The MOMP could also occur in the absence of the extracellular signals in a response to various types of intracellular stress conditions, including oxidative stress, and in this case one speaks of the "*intrinsic apoptosis*". The MOMP seems to be caused by the oligomerization of the pro-apoptotic members of the Bcl-2 family (Bax and Bak) in the outer membrane of mitochondria (OMM) with formation of pores, which allow the escape of small proteins from the intermembrane space of

mitochondria into the cytoplasm. Some of them have direct roles in the digestion of cell components (as endonuclease G (EndoG), apoptosis-inducing factor (AIF), high temperature requirement protein A2 (HTRA2), for a review see (Kroemer *et al.*, 2007; Wang and Youle, 2009)).



**Figure 1.5.21.** Scheme of the two pathways of apoptosis initiation (figure adapted from (Galluzzi *et al.*, 2012)).

Initiator caspases are shown in blue, effector caspases are shown in violet, caspase 9 which plays an important role in intrinsic apoptosis is shown in red.

In vertebrates, the key player is the cytochrome *c*, which otherwise translocates electrons from the cytochrome *bc*<sub>1</sub> complex to the cytochrome *c* oxidase. When cytochrome *c* finds itself in the cytoplasm, it interacts with the apoptotic protease activating factor (Apaf-1). This interaction induces the oligomerization of the Apaf-1 proteins into an apoptosome, followed by the activation of the pro-caspase-9, which triggers the caspase cascade of the apoptosis (Green and Reed, 1998; Reubold *et al.*, 2011; Wang and Youle, 2009).

The release of the mitochondrial proteins into the cytoplasm could also be caused by the assembly of the Permeability Transition Pore Complex (PTPC) within the inner mitochondrial membrane (Brenner and Grimm, 2006; Tait and Green, 2010), which can further result in the swelling of mitochondrial matrix and the rupture of the outer membrane of mitochondria.

An increase in the production of reactive oxygen species (ROS) by the electron transfer chain in mitochondria has been shown to trigger the intrinsic apoptosis pathway (reviewed in (Skulachev, 1996; Wang and Youle, 2009)). The mechanisms of ROS action are not fully understood, so that several mechanisms are under consideration:

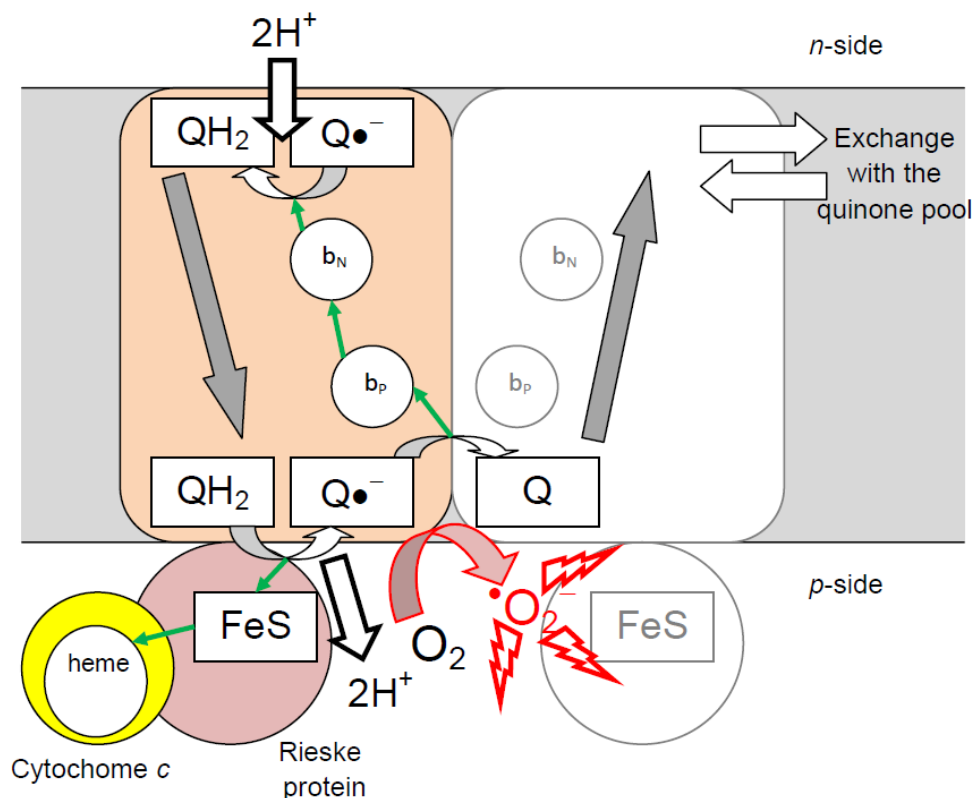
- ROS can increase the permeability of the inner mitochondrial membrane, either by directly damaging the lipids of the membrane, or by facilitating the formation of so-called mitochondrial permeability transition pores, which ultimately could lead to the rupturing of the OMM.
- ROS can trigger the formation of pores in the OMM, which are formed by Bax and Bak proteins (Kushnareva *et al.*, 2012; Shimizu *et al.*, 1999; Wei *et al.*, 2001).

The ROS are occasionally generated in the cytochrome  $bc_1$  complex because the oxidation of a quinol molecule in the centre  $P$  is accompanied by a transient formation of a low-potential unstable ubisemiquinone that quickly reduces the low-potential heme  $b_p$ , see **Figure 1.5.22** and (Mulkidjanian, 2005; Rutherford *et al.*, 2012) for reviews. The redox potential of this ubisemiquinone cannot be increased (via its stabilization by the surrounding amino acid side chains), without losses in the thermodynamic efficiency of the Q-cycle. Instead of increasing this potential, the cytochrome  $bc$  complexes are fine-tuned to minimize the electron escape to oxygen in centre  $P$  (Mulkidjanian, 2005; Rutherford *et al.*, 2012). This is achieved by keeping the lifetime of the semiquinone in centre  $P$  very short, which is reflected by the fact that this semiquinone could be measured only under very special, steady state conditions (Al-Attar and de Vries, 2012). But when the oxidation of cytochrome  $b$  via centre  $N$  is blocked (by inhibitors, or under the backpressure of membrane potential, or in response to an abrupt change in the redox balance of the electron transfer chain), the lifetime of ubisemiquinone in centre  $P$  could transiently increase, and thus electrons would be able to escape to oxygen. The first product of such reaction is superoxide, but ultimately other ROS are also formed (Andreyev *et al.*, 2005; Drose and Brandt, 2008; Rutherford *et al.*, 2012; Yin *et al.*, 2010).

Specifically, the ROS yield increases in response to the oxidation of the membrane ubiquinone pool (Drose and Brandt, 2008). Under the oxidized conditions, the  $bc_1$  can get out from the kinetically optimized activated state (see Section 1.5.3.3 and (Mulkidjanian, 2007; Mulkidjanian, 2010)), and the probability of electron escape to oxygen would increase. Transient oxidation of ubiquinol pool is a common consequence of the traumas and can occur during reperfusion (the restoration of blood flow to an organ or to tissue, e.g. after a heart attack, ischemia or a stroke). ROS can damage membrane components and, specifically, the cytochrome  $bc_1$  complex itself, which would additionally deregulate the fine tuning in this enzyme. Yin and co-workers have recently shown that gradual destruction of the cytochrome  $bc_1$  complex structure by different means (like heat inactivation or proteinase K digestion) always led to a gradual increase in superoxide production (Yin *et al.*, 2010). Hence, the cytochrome  $bc_1$  complex can get into a vicious cycle – occasional generation of ROS could eventually damage the cytochrome  $bc_1$  complex itself or, by affecting its neighbours in the membrane, change its conformation, which would lead to a further increase in the ROS production and further functional damage to the cytochrome  $bc_1$  complex. Then, apoptosis is one possible strategy to save other cells from the ROS-generating vicious cycle in damaged mitochondria by eliminating the initially affected cell (Skulachev, 1996).

The generation of ROS could occur both in complex I and complex III, but the latter is far more dangerous. ROS are generated in complex I under the conditions of reverse electron flow, which implies high membrane potential, high succinate/fumarate ratio and low NADH/NAD<sup>+</sup> ratio (Andreyev *et al.*, 2005; Korshunov *et al.*, 1997; Kushnareva *et al.*, 2002). These conditions have nothing to do with physiological conditions. Importantly, production of ROS by complex I cannot sustain the vicious cycle of self-destruction: any damage to the membrane caused by ROS would decrease the membrane potential and thereby diminish the ROS production in complex I. But the situation with the ROS production in the  $bc_1$  (mitochondrial complex III) is different from that in complex I. The ROS are produced by the forward electron flow. Generally, a drop in membrane potential should stop the production of ROS also in an intact  $bc_1$  (Korshunov *et al.*, 1997). However, if the  $bc_1$  gets damaged, e.g. by an initial burst of ROS, then such damaged  $bc_1$  would generate ROS even in the absence of membrane potential (Yin *et al.*, 2010). Thus, the ROS-producing cell with broken  $bc_1$  must be eliminated (Skulachev, 1998; Skulachev, 2006).





**Figure 1.5.22. Q-cycle mechanism for the ubiquinone-dependent cytochrome  $bc_1$  complex of aerobic organisms.**

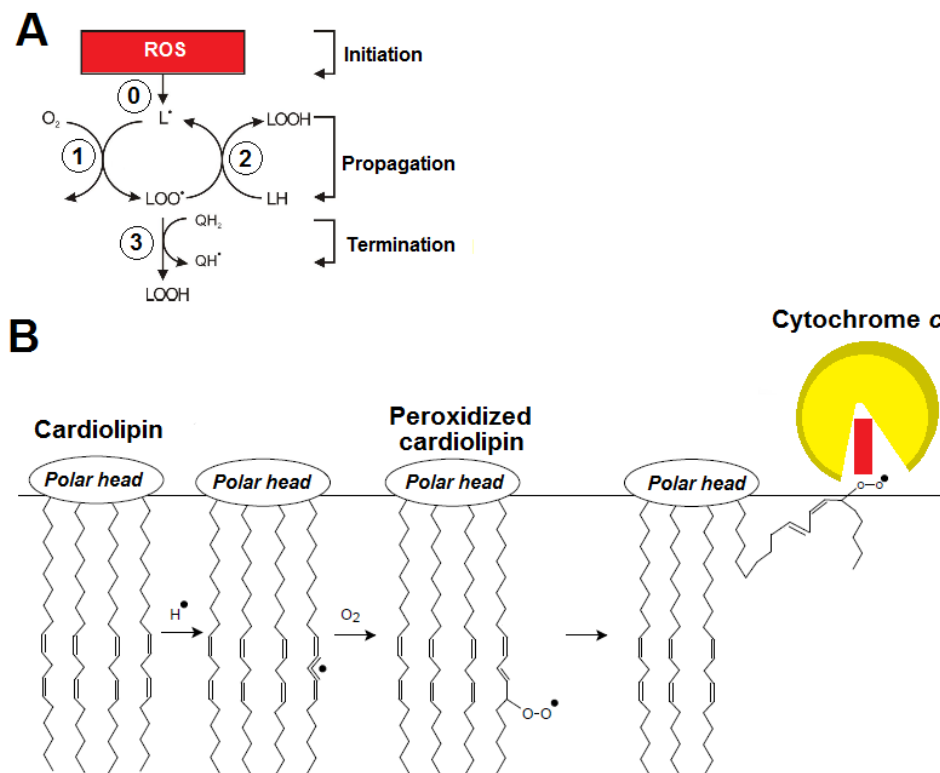
Designations: green arrows show the electron transfer steps, red arrows show uptake of protons from the  $n$ -side and their release to the  $p$ -side, grey arrows depict quinol diffusion. A possible one-electron reduction of oxygen yielding superoxide at centre  $P$  is shown with the red arrow.

Inspection of the apoptotic pathways in different multicellular organisms shows that the evolution of apoptotic cascades in vertebrates has led to development of mechanisms which diminish the ROS-induced damage to other cells by triggering the apoptosis as soon as possible, i.e. even *before* the disruption of the affected mitochondria. This is achieved by a signal amplification cascade within mitochondria, as depicted in **Figure 1.5.23A**. It has been shown that molecules of a four-tail lipid cardiolipin (CL) that is particularly susceptible to the ROS-induced peroxidation, in oxidized form can interact with the molecules of cytochrome  $c$  at the outer surface of the inner mitochondrial membrane and cause conformational change in the latter (Huttemann *et al.*, 2011; Kagan *et al.*, 2009). The affected molecules of cytochrome  $c$  attain peroxidase activity and start to produce additional ROS (including singlet oxygen). This, apparently, accelerates the formation of a pore in the

outer mitochondrial membrane and the release of cytochrome *c* into the cytoplasm (Huttemann *et al.*, 2011; Kagan *et al.*, 2009; Miyamoto *et al.*, 2012).

Cardiolipin (CL) molecules are the main targets of peroxidation in mitochondrial membranes (Ji *et al.*, 2012; Skulachev *et al.*, 2009). They are easily being peroxidized because one molecule of CL can carry four linoleate chains (this is typical, for instance, for the heart CL) and thus contain eight unsaturated bonds (a scheme of radical propagation upon peroxidation of polyunsaturated lipid containing bis-allylic hydrogen atoms is shown in **Figure 1.5.23A**). The high number of double bonds in CL is important for its structure because it enables packing of the CL molecules in the bilayer and might be crucial for interactions with the enzymes of mitochondrial inner membrane. It is known that amphiphilic molecules must have cylindrical shape with approximately similar widths of the polar head and the hydrophobic tails to form a stable membrane bilayer (Israelachvili *et al.*, 1977). This condition is still fulfilled for a CL molecule with unsaturated fatty acids even despite its bulky hydrophobic part, but unlikely for CL molecules with peroxidized, partially polar chains.

A peroxidized CL molecule would tend to stick out from the bilayer and then interact with cytochrome *c* molecules at the membrane surface, see **Figure 1.5.23B**. By inserting into the cytochrome *c*, a fatty acid chain of CL opens the heme-binding cleft, breaks the methionine-iron bond and makes the heme accessible to external ligands such as peroxy groups and oxygen (Huttemann *et al.*, 2011; Kagan *et al.*, 2009). The breakage of the methionine-iron bond should also decrease the  $E_m$  value of cytochrome *c* (Winkler *et al.*, 1997), so that the ability of superoxide generation by such modified cytochrome molecules cannot be excluded. The oxidation of CL molecules is unlikely to take place within the membrane bilayer as ubiquinol molecules of the respective pool are acting as potent antioxidants and protecting lipids from peroxidation. Ubiquinol molecules interact with peroxides (reaction 3 in **Figure 1.5.23A**) very fast, with a rate constant as high as  $3 \cdot 10^5 \text{ M}^{-1}\text{s}^{-1}$ , see (Pratt *et al.*, 2011) and references therein. And the reaction of the hydrogen atom transfer (reaction 1 in **Figure 1.5.23A**), a bottleneck in the radical propagation, is very slow. For the polyunsaturated linoleate methyl ester, its rate constant is as small as  $60 \text{ M}^{-1}\text{s}^{-1}$ .



**Figure 1.5.23. Interaction of cardiolipin and cytochrome *c* upon peroxidation of the former.**

(A) General scheme of lipid peroxidation, adapted from (Sies, 1993); L is a lipid molecule, QH<sub>2</sub> and QH• are membrane ubiquinol and its semiquinone form, respectively. (B) Scheme of transformations of a polyunsaturated cardiolipin molecule upon peroxidation according to (Kagan *et al.*, 2009; Kagan *et al.*, 2005). The black dot indicates the position of an unpaired electron. The heme in cytochrome *c* is shown with red bar.

Because the concentration of ubiquinol in the mitochondrial membrane is comparable with the concentration of polyunsaturated lipid chains, ubiquinol should fully protect the bilayer lipids from peroxidation (Lokhmatikov *et al.*, 2012). Thus, the only lipid molecules that are susceptible to peroxidation are the molecules not accessible for the pool ubiquinol molecules for some reason (Lokhmatikov *et al.*, 2012). These would be lipid molecules that are attached to membrane protein complexes and, specifically, CL. Most recent data show that the large part, if not majority, of CL molecules in mitochondrial membranes are associated with protein complexes (Althoff *et al.*, 2011; Mileykovskaya *et al.*, 2012). Cardiolipin molecules were identified in the crystal structures of many energy-converting enzymes, such as ADP/ATP carrier (Nury *et al.*, 2005; Pebay-Peyroula *et al.*, 2003), succinate dehydrogenase (Horsefield *et al.*, 2006; Yankovskaya *et al.*, 2003), cytochrome *bc*<sub>1</sub> complex (Crowley *et al.*,

2008; Hao *et al.*, 2012; Huang *et al.*, 2005; Solmaz and Hunte, 2008), cytochrome *c* oxidase (Aoyama *et al.*, 2009; Muramoto *et al.*, 2007; Muramoto *et al.*, 2010; Ohta *et al.*, 2010; Shinzawa-Itoh *et al.*, 2007; Suga *et al.*, 2011; Tsukihara *et al.*, 2003), and formate dehydrogenase (Jormakka *et al.*, 2002), see also (Althoff *et al.*, 2011; Arias-Cartin *et al.*, 2012; Mileykovskaya *et al.*, 2012; Palsdottir and Hunte, 2004) for reviews. The mitochondrial respiratory supercomplex itself appears to include hundreds of CL molecules (Althoff *et al.*, 2011). Some of them could be inaccessible to the membrane ubiquinol molecules and, hence, could serve as triggers of apoptosis.

## 1.6. Aims of the thesis

The main goal of the current work is to perform phylogenomic and comparative structural analyses of several widespread energy converting enzymes. We focus on major subfamilies of the enzymes performing reactions with nucleoside triphosphates (ATP, GTP) and on the cytochrome *bc* complex which serves as a hub in the vast majority of electron transfer chains. Specific tasks solved in this study are as follows:

- phylogenomic and comparative structural analysis of the most ancient families of proteins that perform hydrolysis of phosphoester bonds;
- phylogenomic analysis of rotary membrane ATPases/ATP synthases;
- phylogenomic analysis of the cytochrome *bc* complex
- evolutionary analysis of energy-converting enzymes that are involved in apoptosis.

Sequences of proteins are the only "fossils" from the times before 3.8 Gy of which no notable geological records are available. Thus, analysis of the sequences can be used to dramatically reduce uncertainty of evolutionary reconstructions of very early events, such as the separation of the three domains of life or even the origin of life itself. Therefore, the results of phylogenomic analysis can serve as arguments upon considering different schemes of early evolution of energy conversion and can clarify the role of energy in the early evolution of life.

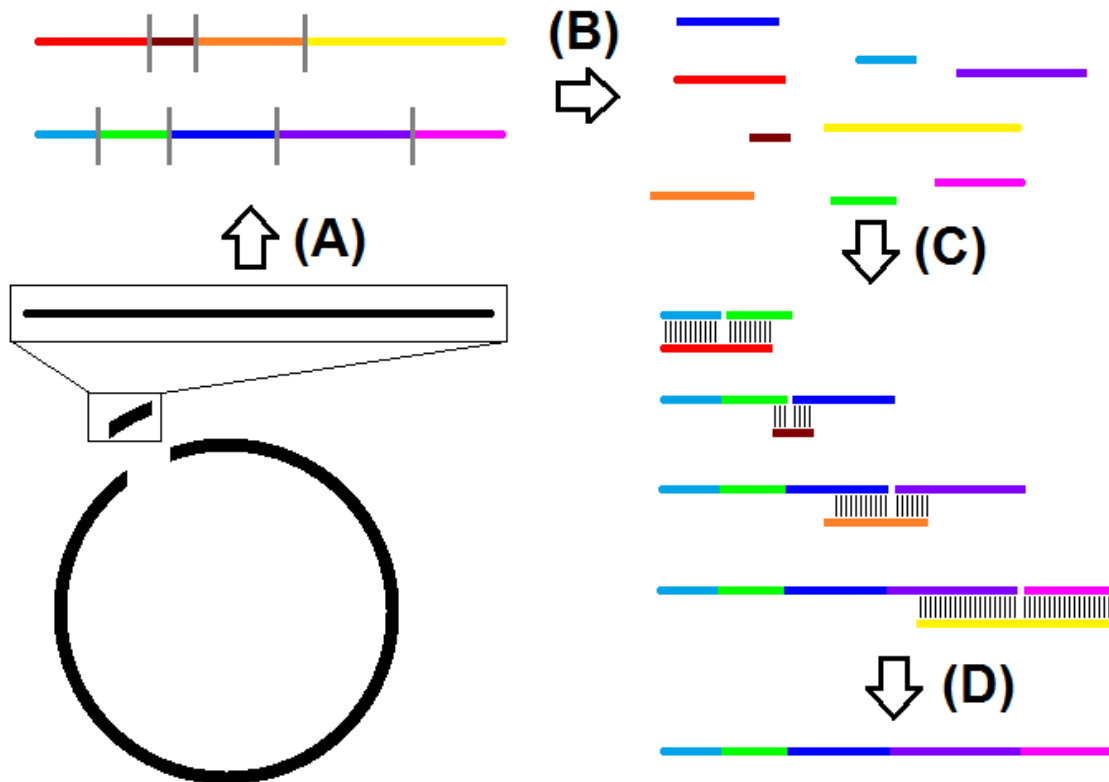
## 2. Methods

Generally, the presence of some feature (character) in a group of organisms can no longer be considered as a direct evidence for its presence in the ancestor of this group. For example, cellular membrane formation or DNA replication could be achieved by action of unrelated genes. Focusing on proteins performing particular functions seems to be more reliable. Indeed, evolution of a function that is performed by a single protein can be traced via its phylogenetic tree; regrettably, this approach does not always help to elucidate the evolution of a complex feature carried out by several different proteins. Studies of the evolution of multi-subunit complexes (as mostly involved in the cellular energy transformation) are difficult because the respective proteins could perform quite different functions before their recruitment as subunits of such enzyme complexes. It is unlikely that the complexes could have appeared at once and fully functional. Therefore, most plausible evolutionary scenarios should not only consider evolution of full-fledged complexes but also provide possible functions for their constituents (i.e. individual parts of the complex) before their merging into a complex structure.

### 2.1. Principles of phylogenomic analysis

#### 2.1.1. Sequencing of full genomes

Modern techniques allow reading nucleotide sequences of the sample yielding a set of short (around 50-300 nucleotides), overlapping *reads* (Pettersson *et al.*, 2009). These short portions can be further *assembled* by specific mathematical algorithms into longer *contigs*. Algorithms for the assembly are complex and in principle can lead to mistakes in positioning of some regions, but yet this strategy is applicable for obtaining full microbial genomes (*Figure 2.1*).



**Figure 2.1. Short description of the sequencing process.**

(A) A very large nucleic acid molecule (for example, a genome) is being split by restriction proteins into shorter molecules, which are in turn cut into very short reads (roughly around 100 nucleobases). (B) The sequences of the reads are obtained. As different short molecules can be cleaved at diverse positions, the resulting reads would cover the original short molecule, although overlapping with each other. (C)-(D) These overlaps can be used to reconstruct the molecule back from the reads. The symbols | show matches between sequences (their multitude gives the pairwise alignment).

## 2.2. Protein comparison via sequence alignment

In a nucleotide sequence of a genome (chromosome, plasmid etc.) the protein-coding regions can be predicted. Translation of the possible coding regions yields predicted protein sequences. In general, comparison of genomes relies on comparisons of individual proteins. Protein sequences can be compared by producing their alignments. Pairwise alignment means writing of the amino acid sequences one under another so that they correspond to each other in the best possible way. This definition is blurred but this reflects the idea behind the

alignment: it is constructed by the means of agreement between the sequences of letters but is expected to reflect the functional and structural similarities of proteins.

Several tools for pairwise alignment are available but the fastest and most widely used are united into the BLAST tool set (Altschul *et al.*, 1997). These are also commonly used for the search of proteins similar to query sequence in the databases. Although the exact implementation of BLAST which makes it fast and convenient to use is rather complicated, the idea behind it is plain simple and will be described below, omitting the algorithmic part.

### 2.2.1. Global and local pairwise alignment

The first widely-used computational algorithm for aligning two proteins was proposed in 1970 (Needleman and Wunsch, 1970). Its purpose was to write two sequences of letters (for example, protein sequences) one under another so that they correspond to each other the best way by inserting a special *gap* character "-" into specific places of each sequence. For two sequences of around 300 symbols each (this is a typical length of a protein) this cannot be done by exhaustive search because of huge number of possible variants and therefore unrealistic time of computation, but this algorithm allows to reduce the time enormously by checking only suitable cases. It balances between maximizing the number of matches and minimizing the number of the gaps required to get it. Algorithm of Needleman-Wunsch allows obtaining so-called *global alignment*, i.e. full length sequences are being aligned.

However, many proteins do not align globally but still can be related to each other (for instance, in compared multidomain proteins some domains can be very similar structurally and functionally, whereas other domains can be absolutely unrelated). Thus, another strategy relying on the same algorithm was proposed in 1981. Searching for regions of local correspondence, yielding *local alignment(s)* between the sequences was suggested (Smith and Waterman, 1981). This idea is now implemented in BLAST, which is called after the Basic Local Alignment Search Tool.



### 2.2.2. Position-independent amino acid substitution matrices

In the algorithms of Needleman-Wunsch and Smith-Waterman, matches and mismatches between sequences are set constant as well as the penalty for the gap insertion; all these parameters can be chosen arbitrary by the user. This approach does not treat sequences of letters as proteins: it does not account for the chemical nature of amino acid residues and ignores evolutionary pressure on amino acid replacement. Thus, such an approach does not allow distinguishing between almost inconsequential replacements (for example, change of Val to Ile, or Glu to Asp) and much more crucial mismatches (for instance, Gly to Trp, or Glu to Lys). In a famous work by Dayhoff *et al.*, the analysis of this issue allowed the authors to propose a PAM (Point Accepted Mutation) *matrix for the amino acid substitutions* (Dayhoff *et al.*, 1978). The values of the matrix were calculated based on the real replacements taking place in closely related proteins in the course of evolution. Now, instead of taking a positive value for any "match" and a negative value for any "mismatch" in the algorithms of sequence alignments, it was possible to refer to a special value for each pair of amino acids. This work was followed by construction of the BLOSUM (Blocks Substitution Matrix) matrices (Henikoff and Henikoff, 1992). These matrices are considered better than PAM for database searches as they use more distantly related proteins and thus represent the replacement scores better. BLOSUM62 matrix (where 62 refer to the 62% of minimal identity between the proteins used for its construction) is used by BLAST as a default, but can be optionally changed to other BLOSUM or PAM variants. The matrices described are *position-independent* since the score of substitution is considered the same for all proteins and for all positions in the protein.

### 2.2.3. Global multiple alignment. Similarity searches.

Algorithms for pairwise sequence alignments allow comparing two proteins, but larger numbers of similar proteins can be compared by building a *multiple alignment* rather than through a series of pairwise comparisons. A multiple alignment, just as a pairwise alignment, is a record of sequences one under another where "overall similarity is maximized". One would surely prefer to have amore accurate definition, but it is what it is: the alignment is constructed in the hope that it will reflect the structural and functional similarity between the

proteins, but all available algorithms are heuristic and can only maximize different sophisticated scores without providing the exact solution. To summarize, multiple alignment, as compared to a pairwise alignment, allows obtaining more information, but its construction is a more complex task, and the respective algorithms are generally more sophisticated. CLUSTAL, the first algorithm suitable for a fast alignment of numerous sequences was proposed in 1988 (Higgins and Sharp, 1988) and further improved (CLUSTAL W) in 1994 (Thompson *et al.*, 1994). This later version and its updates are extensively used until now, but a number of new algorithms have been proposed (they mostly differ in computational details, as described elsewhere (Wallace *et al.*, 2005)). In the present work, the MUSCLE algorithm was preferred (Edgar, 2004) since, based on previous experience, it aligns the less conserved parts rather accurately and thus is well suitable for the alignment of membrane proteins.

We do not focus here on computational details of algorithms used for multiple alignment. Instead, we analyse the alignments from the biological viewpoint, which is discussed in Section 2.3.

The BLAST tool is fast enough to make quick searches through the vast protein databases in a time range from seconds to minutes. But one should keep in mind that plain BLAST relies on BLOSUM or PAM matrices, and the flaws of these matrices for the alignments of membrane proteins is apparent: they were mostly built based on the known families of water soluble enzymes which are remarkably different from membrane proteins in both amino acid composition and amino acid replacement frequency. Thus a search for membrane proteins with BLAST can be biased, and distantly homologous sequences can be missed. However, this obstacle can be overcome with the use of approaches based on PSSM (Position-Specific Scoring Matrix, in contrast to the position-independent BLOSUM and PAM) (Gribskov *et al.*, 1987; Henikoff and Henikoff, 1996; Tatusov *et al.*, 1994). This is of particular interest for this thesis where many membrane protein sequences are analyzed.

A PSSM is obtained from a multiple alignment of protein sequences. A multiple alignment can be treated as a set of columns (also called positions, or characters), each containing one symbol from each of aligned sequences. This symbol could be either an amino acid residue or a gap. Each column can be characterized by a vector of, to put it simple, an expected probability to detect each amino acid inside it; a set of columns thus yields a matrix (*Figure*

**2.2B**). Different possible methods could be used to estimate the expected probability, with simple occurrence frequency being the most obvious and imperfect (Tatusov *et al.*, 1994).

In principle, all columns of a multiple alignment can be used for construction of PSSM, but positions without any conservation do not provide any specific information as frequencies in them could reflect at best only a global occurrence frequencies of amino acids. In practice, only a subset of columns of a multiple alignment is used for construction of PSSM. For a particular alignment, a subset of positions which contain at least partially conserved residues in all sequences is termed a *block* (Henikoff *et al.*, 1990; Posfai *et al.*, 1989). Blocks are usually chosen so that they contain only ungapped regions (Castresana, 2000; Henikoff and Henikoff, 1991; Henikoff *et al.*, 1999; Talavera and Castresana, 2007; Tatusov *et al.*, 1994), however this is done mostly for algorithmical simplicity and is not an absolute requirement.

Amino acids (~ 20 positions)							
	A	C	D	E	F	.	W
A	X <sub>AA</sub>	X <sub>CA</sub>	X <sub>DA</sub>	X <sub>EA</sub>	X <sub>FA</sub>	.	X <sub>WA</sub>
C	X <sub>AC</sub>	X <sub>CC</sub>	X <sub>DC</sub>	X <sub>EC</sub>	X <sub>FC</sub>	.	X <sub>WC</sub>
D	X <sub>AD</sub>	X <sub>CD</sub>	X <sub>DD</sub>	X <sub>ED</sub>	X <sub>FD</sub>	.	X <sub>WD</sub>
E	X <sub>AE</sub>	X <sub>CE</sub>	X <sub>DE</sub>	X <sub>EE</sub>	X <sub>FE</sub>	.	X <sub>WE</sub>
F	X <sub>AF</sub>	X <sub>CF</sub>	X <sub>DF</sub>	X <sub>EF</sub>	X <sub>FF</sub>	.	X <sub>WF</sub>
.	.	.	.	.	.	.	.
W	X <sub>AW</sub>	X <sub>CW</sub>	X <sub>DW</sub>	X <sub>EW</sub>	X <sub>FW</sub>	.	X <sub>WW</sub>

**(A)**

Positions (~ length of the domain)							
	1	2	3	4	5	.	N
A	X <sub>1A</sub>	X <sub>2A</sub>	X <sub>3A</sub>	X <sub>4A</sub>	X <sub>5A</sub>	.	X <sub>nA</sub>
C	X <sub>1C</sub>	X <sub>2C</sub>	X <sub>3C</sub>	X <sub>4C</sub>	X <sub>5C</sub>	.	X <sub>nC</sub>
D	X <sub>1D</sub>	X <sub>2D</sub>	X <sub>3D</sub>	X <sub>4D</sub>	X <sub>5D</sub>	.	X <sub>nD</sub>
E	X <sub>1E</sub>	X <sub>2E</sub>	X <sub>3E</sub>	X <sub>4E</sub>	X <sub>5E</sub>	.	X <sub>nE</sub>
F	X <sub>1F</sub>	X <sub>2F</sub>	X <sub>3F</sub>	X <sub>4F</sub>	X <sub>5F</sub>	.	X <sub>nF</sub>
.	.	.	.	.	.	.	.
W	X <sub>1W</sub>	X <sub>2W</sub>	X <sub>3W</sub>	X <sub>4W</sub>	X <sub>5W</sub>	.	X <sub>nW</sub>

**(B)**

**Figure 2.2. Difference between position-independent matrices of amino acid substitutions as BLOSUM or PAM (A) and position-specific matrices used by HMMer or PSI-BLAST (B).**

**(A)** Position-independent matrices could be symmetric alongside the main diagonal (shown in grey), thus  $X_{AC} = X_{CA}$  and so on. They show the average relative score of the replacement of one amino acid by another (which is usually negative for the replacements remarkably changing the properties of the residue, shifting its size, acidity etc, and positive for more conservative replacements), and thus are applicable for aligning of different proteins.

**(B)** Position-specific matrix is describing a particular protein domain (or, more widely, any multiple alignment). It is not square and its actual size depends on the alignment length. The matrix is being constructed based on the occurrence of amino acids in each particular position of multiple alignment.

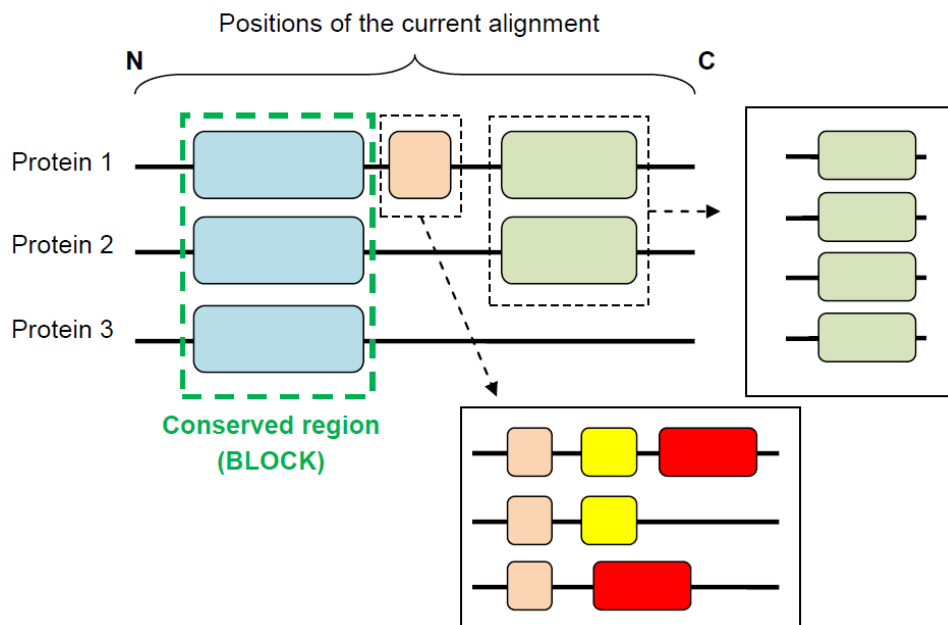
The *sequence logos* were introduced as visual representations of simple PSSMs (Schneider and Stephens, 1990). Each logo consists of stacks of letters (of single-letter amino acid code or of single-letter nucleotide code), one stack for each position in the multiple alignment. The height of the stack reflects the sequence conservation at that position, while the height of each letter in the stack indicates the relative frequency of corresponding amino or nucleic acid at that position.

Commonly PSSM is used to characterize a *protein domain* or a particular protein family which contains several domains, see **Figure 2.3**. The term "protein domain" is not restricted to the aforementioned definition: it is more naturally described as a structurally separated part of the protein with a specific function. Under this definition such widely known databases as SCOP (Murzin *et al.*, 1995) and CATH (Orengo *et al.*, 1997) provide structural comparisons and classification of proteins. A comprehensive database of protein domains Pfam in addition to classification of the domains based on the sequences has cross-links with the PDB database of protein structures and the manual annotation for many of its entries (Finn *et al.*, 2010).

PSSMs can serve as a test whether a subject sequence belongs to the family described by it or not (**Figure 2.4**). In short, if amino acid residues of the sequence are highly expected by PSSM, this sequence is denoted as a *match* for this PSSM. Thus, as the query sequence in normal BLAST, such a match can be used upon similarity searches.

PSI-BLAST, named after Position-Specific Iterated BLAST (Altschul *et al.*, 1997), is actually a combination of simple BLAST with the PSSM-based approach. It progresses through iterations. On the iteration #1 a simple BLAST is performed for a query sequence. As any other BLAST search, it results in a set of local alignments. Each alignment can be characterized by a *raw score*: a sum of all amino acid substitution scores (taken from the substitution matrix, for instance BLOSUM62) and the penalties for gaps insertion. The program also estimates the expect value or *e-value* for each alignment. This value is similar to probability of obtaining such alignment by chance, but it, in principle, can exceed 1 as it represents the number of different alignments with scores equivalent to or better than the current score that is expected to occur in a database search by chance. E-value is used as a measure of alignment reliability: e-value of  $10^{-5}$  is frequently taken as a threshold for considering an alignment non-random. Low e-values (depending on the particular case, but

usually lower than  $10^{-20}$ ) are typical for orthologs (proteins with the common origin and the same function), e-values between  $10^{-20}$  and  $10^{-5}$  could be attributed to both distant orthologs and paralogs, while e-values higher than  $10^{-5}$  in general indicate that the relation between two proteins should be clarified by other methods than pairwise sequence comparison.



**Figure 2.3. Schematic representation of the multiple alignment which reflects complex domain structure of the proteins.**

Region in the sequences colored blue is conserved in all proteins of current alignment and thus can be termed as a block of it. Other parts of the proteins (hardly detectable on current alignment and shown in dotted boxes) can be attributed to distinct domains (possible multiple alignments for them are boxed).

Search hits are listed and sorted according to their e-values, and the user is allowed to manually choose the hits which are assumed to be the correct ones (or just take all hits with e-value lower than given value). The chosen set is then a subject to multiple alignment construction and calculation of the PSSM, which would be used as a query upon the iteration #2. The new hits of the further iterations could be added to the alignment and thus PSSM is improved after every iteration until no new hits are identified and the process stops.

Iteration #1 can be skipped if the multiple alignment of some proteins of interest was previously obtained from different source. Then it can be supplied to PSI-BLAST or used in the separate HMMer program (Finn *et al.*, 2011).

### Multiple alignment

Protein	Positions									
Prot 1	M	M	A	A	L	L	V	L	E	E
Prot 2	M	Q	A	A	L	L	L	V	E	E
Prot 3	M	N	A	A	L	L	V	V	D	E
Prot 4	M	F	K	A	L	L	V	V	D	E
Prot 5	M	G	K	A	L	L	V	V	D	E

### PSSM

Amino Acid	Positions									
	1	2	3	4	5	6	7	8	9	10
A			0.6	1						
C										
D									0.6	
E									0.4	1
F		0.2								
G		0.2								
H										
I										
G		0.2								
K			0.4							
L					1	1	0.2	0.2		
M	1	0.2								
N										
P										
Q		0.2								
R										
S										
T										
V							0.8	0.8		
W										

### Examples of query sequences

Query	Positions									
Query1	M	G	A	A	L	A	L	V	D	E
Query2	G	F	H	A	A	V	A	A	K	R
Query3	M	R	K	A	L	H	V	G	E	E

Perfect hit  
 Very weak hit  
 Moderate hit

**Figure 2.4.** Example of a PSSM (middle table) based on the multiple alignment (top table) with simple frequencies of amino acids used as a scores.

Empty cells are considered to be filled with zeros. Query sequences can be checked for compliance with this matrix in each of their positions and then in total (bottom table): the color reflects the correspondence with the PSSM (red – poor, yellow – moderate, green – good).

### 2.3. Critical analysis of multiple alignment from the biological point of view

As the algorithms used for multiple alignment construction are sophisticated, one should care to analyze the output from the biological point of view. An alignment must not be treated as a black box or as an "output of the program" but as a piece of biologically valuable data which anyone can examine with eyes. The steps of such a biological analysis, as routinely performed throughout this work, are listed below.

- **The alignment should be checked for unexpectedly long or short sequences.**

If such deviating sequences appear as single sequences (not inside a specific group in alignment) they can frequently be a result of the mistakes in the sequence database, especially when RefSeq or other large database is used. In some cases, *short proteins* really represent the truncated version of the protein, but particularly if the "normal" protein is already present in the same organism, these proteins can be deleted from the sequence sample as database errors (but this should be mentioned when describing the results). If the protein of interest is a multidomain protein (this can be checked with the Pfam database of protein domains), one should check whether this short version is just a single domain of a multidomain protein and whether anything is known about the function of this domain. *Long proteins*, especially if the long non-aligned part occurs from the N- or C-terminus, can be either (1) fusions with other proteins (which could reflect the real case or an artificial merge because of the erroneous missing of stop codon in a database) or (2) a database error in the identification of the gene start codon. One can discriminate between these two cases by doing a BLAST search for the non-aligned part of the protein (or better a gene itself) alone: if no hits are obtained, then it is very likely to be the latter case.

In the cases when no other sequences from the same organism are identified except for the suspicious one, the reason of its abnormal length should be still clarified. After deleting all suspicious sequences from the sample, it is recommended to re-align all the other sequences, since the presence of abnormal proteins could essentially affect the initial accuracy of the alignment.

- **All the possible additional information should be mapped to the alignment.**

If the 3D structure of any protein from the studied family is available, its sequence should be used to locate known structural features: (1) elements of secondary structure, (2) ligand binding sites, (3) binding interfaces with other subunits (if the studied protein forms a complex). If no 3D structure was solved for any member of the respective family (which is common, for instance, for membrane proteins), it is worthwhile to try to predict the secondary structure of the given protein by using one of the available algorithms, for example Jpred, and applying it to the multiple alignment (Cole *et al.*, 2008). If the studied protein is a membrane protein, its transmembrane regions should be either extracted from the 3D structure or predicted from sequence of one of the proteins in the alignment (for example, by TMHMM algorithm (Krogh *et al.*, 2001)). Finally, if the protein is a multidomain protein, the known domains should be indentified (for instance, from the Pfam database). Information about functionally important residues can be also derived from literature.

- **The overall quality of the alignment should be estimated based on additional data.**

For example, while poor conservation of the N- or C-termini as well as the loop regions is common, the poor conservation of the substrate binding sites/interfaces should be considered as an indication of a poor quality of the alignment. The reasons behind it can be different, but until the issue is fixed, one should treat the alignment data with care and keep in mind that any reconstructions based on it (i.e. phylogenies) could be unreliable.

- **Manual improvement should first focus on the well-aligned parts of the alignment.**

Careful inspection of the overall well-aligned part allows identifying singular sequences which are not aligned properly or even are not aligned at all. On one hand, this might be due to some algorithmic issue, when the sequence is erroneously shifted by several residues, and should be fixed manually, if possible. On the other hand, such a poor alignment could reflect specific feature of this sequence (for example, a loss of the active site), and, as such, should be taken into account.



Generally all these operations are covered by the term "manual improvement" as used throughout this thesis.

## 2.4. The problem of "the same" genes and the minimal gene set concept

The start of comparative genomics, a field where full genomes are analyzed, can be dated back to the 1996, when the first two sequenced genomes were compared (Mushegian and Koonin, 1996). These were the genomes of parasitic species *Haemophilus influenzae* (gram-negative bacteria) and *Mycoplasma genitalium* (gram-positive bacteria) belonging to very diverged taxonomical groups. One of the results of this paper was the *minimal gene set* concept: if the *same gene* is observed in distinct species, then it should be important for the cell functioning, so that identification of such genes should allow constructing the minimal set of genes possibly required for the cell survival.

An important question concerns "the same gene" term: which genes in different genomes should one call "the same" in this context? Two procedures were worked out to address this problem:

1. Identification of related genes that evolved from one ancestral sequence;
2. Discrimination of homologous genes that diverged as a result of speciation (orthologs) from those homologous genes that diverged after a gene duplication (paralogs) – these terms are defined more rigorously below in Section 2.5.2.

The first procedure can be almost completely automated (see Section 2.2.3), although manual checking of the results is usually required (Section 2.3). The discrimination of orthologs and paralogs is more difficult because of diversity in evolutionary events happening with genes; currently used approaches are discussed in Section 2.5.

In addition, two issues demand attention and analysis even if we succeeded in identification of the "same" genes:

- Sometimes one gene can be functionally replaced by an absolutely unrelated gene (this is called a *non-orthologous gene displacement* (Koonin, 2000; Koonin *et al.*, 1996; Leipe *et al.*, 1999)). Then, even the most perfect search of homologs based on sequence similarity will fail to identify the gene with "the same" function. Such cases could be solved by analyzing patterns of occurrence of different genes: the ones which underwent non-orthologous displacement are likely to have mostly mutually

exclusive patterns (i.e. only one of the two genes is present in the genome while the second is absent). In principle, the non-orthologous gene displacement can be viewed as an emanation of the "analogy" phenomenon (like the relationship between wings of a fly and wings of a bird).

- Genes very often undergo Lateral Gene Transfer, or LGT (reviewed in (Gogarten and Townsend, 2005)). This is a phenomenon of obtaining particular gene from a different contemporary species rather than from the ancestral cell (this "normal" way is called vertical inheritance). The lateral gene transfer is frequently performed through plasmid exchange or by bacteriophage infection.

The second question, and rather important one, concerns the conditions under which one can expect any elucidated set of genes to be minimal. Growth conditions can alter the set of strictly required genes, so that the investigator should distinguish between genes necessary only under certain conditions and genes obligatory needed under all conditions. This problem is usually addressed by providing environment with all the possible chemical compounds required for growth and without any stresses including competition with other species (i.e., laboratory conditions with rich medium). This approach was used in the first experimental attempts to find dispensable and strictly required genes in species of *Mycoplasma* genus which had the smallest known genome among organisms able to grown in pure cultures (Glass *et al.*, 2006; Hutchison *et al.*, 1999). The authors have performed global mutagenesis of *Mycoplasma* species forced by insertion of a specific transposon and observed some insertions which led to viable phenotypes whereas many other insertions disrupted necessary genes leading to non-viable phenotypes. As a result, among 482 protein-coding genes of *M. genitalium* 382 were found to be essential (Glass *et al.*, 2006). Recently, creation of fully synthetic genome, as developed *in silico*, that was able to successfully control life and reproduction of a mycoplasma organism was reported (Gibson *et al.*, 2010).

However the artificial conditions considered in such experimental works are far from natural. No wild bacteria live in such comfort, and thus the term "minimal gene set" is not related to natural circumstances. Generally speaking, because of the huge variety of specific metabolic pathways and requirements for nutrients, as well as diversity of conditions in the "wild" microbial world, an identification of a "real" minimal gene set is likely to be impossible.

The minimal gene set is not obligatory coupled with the possible genes inherited from the common ancestor of all cellular forms (LUCA). However, it appears that the ubiquitous genes, all present in the "minimal gene set", are likely to correspond to the genes of this ancestor. Several cases of non-orthologous displacement within the "minimal gene set" could be used as milestones for analysis of life evolution (Koonin, 2000).

## 2.5. Phylogenomic analysis and its difference from phylogenetic analysis

### 2.5.1. Phylogenetic analysis. Phylogenetic tree construction methods

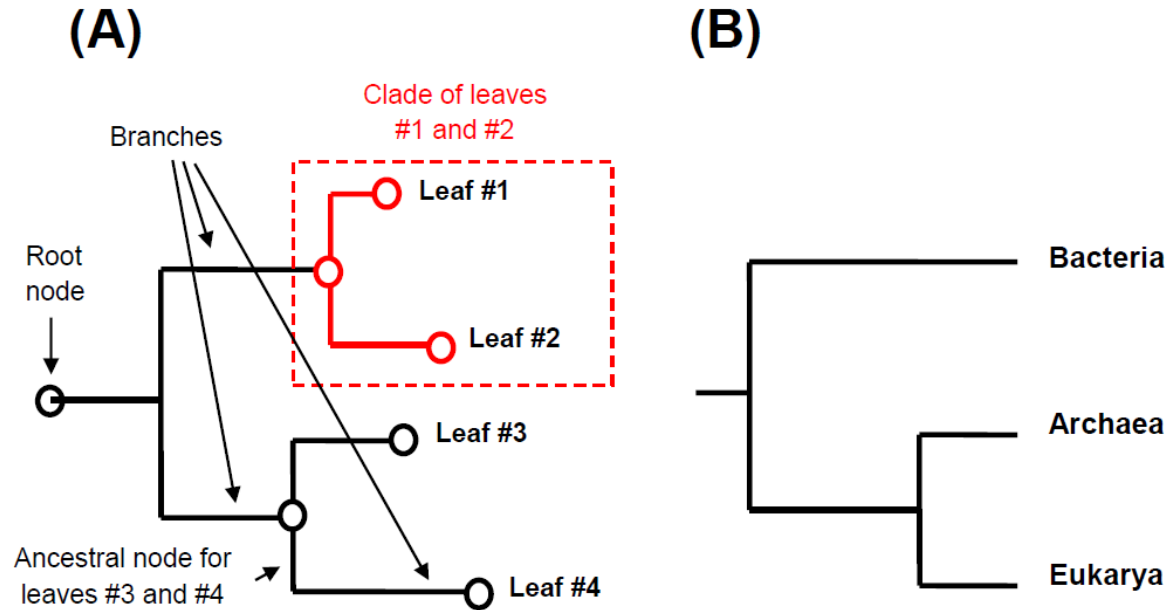
Phylogenetics, in a narrow sense, is a part of biological science focused on revealing the evolutionary relationships between different species from sequences of DNA, RNA or proteins. Charles Darwin was the first to suggest that the evolution of species can be represented with the tree-like structure, "the tree of life". Before the burst in knowledge of biological polymer sequences in the last decades of the XX century, the classification of the species was mainly restricted to relatively big and complex eukaryotic organisms belonging to *Metazoa* or *Metaphyta*. The classification was mostly based on listing the traits of each species and on paleontological data. However, prokaryotes could not be reliably classified neither by morphological features (as they could be too small to discover and could vary) nor by paleontology records, which are essentially lacking in this case.

A *phylogenetic tree*, or *phylogeny* is a reconstruction of evolutionary events, see **Figure 2.5A** for the example of phylogeny with the terms specifying its parts and **Figure 2.5B** for the summarized version of the tree of life by Woese and co-workers (see Section 1.1.3), as an example.

A multiple alignment is believed to contain information on a set of events which produced the observed sequences. Generally, a number of tentative reconstructions of these events could be proposed, and thus a formal algorithm should be applied to distinguish between different reconstructions and choose the best. The literature on different methods of phylogenetic analysis is vast; here we will cover only three most popular algorithms used, which all work under the assumption of independent point mutations (single replacements of amino acids/nucleotides, or single insertions and deletions) as a common mechanism of

proteins evolution. Insertion, deletion or replacement of a single character are considered to be the simplest *evolutionary events*.

- The first method relies on the idea that the evolution should be parsimonious (economical). Thus, for instance, the scenario which implies 10 evolutionary events to explain the origin of a given sequence is less parsimonious and thus worse than a scenario that includes only 5 events. This method is called Maximum Parsimony (MP) method and was proposed for nucleotide sequences in 1971 (Fitch, 1971).
- The second approach is called Neighbor-Joining method, or NJ, and was suggested nearly two decades later (Saitou and Nei, 1987). This method requires a definition of the *distance* between every pair of sequences. The essence of the method is to group sequences with the smallest distance ("neighbors") together by introducing a new interior node between them. After this the distances are recalculated, as distances to the new node should be calculated, while distances for the sequences already grouped together are no longer needed. Algorithm on the next step again searches for the smallest distance, and if one of the new nodes is involved, it is further grouped with the neighbor. NJ works faster compared to other methods (Saitou and Imanishi, 1989), but the major flaw of this method is its dependence on the pre-calculated matrix of distances between all proteins in multiple alignment. Distances between proteins are, in general, assumptions under a specific model, and, moreover, they can differ only slightly if the proteins of interest are diverged or remarkably biased as compared to typical proteins that were used for construction of substitution matrices (which essentially contribute to the distance calculation). As discussed in Section 2.2.3, membrane proteins are characterized by specific substitution matrices, and thus NJ method can result in biased topology when applied to membrane proteins.
- The Maximum Likelihood (ML) method is a common statistical method which was applied to sequence alignments (Felsenstein, 1973). It is the slowest among the considered methods, but recent improvements in the algorithm as well as overall acceleration of computing allows to use it extensively even on big samples (Guindon *et al.*, 2010; Guindon and Gascuel, 2003; Tamura *et al.*, 2011). The idea behind the method is to estimate parameters of the model (e.g. the branching order of a tree), under which the probability of obtaining the data (given sequences) is maximized.



**Figure 2.5.** Basic terms denoting the elements of phylogenetic trees (A) and a scheme of the phylogenetic tree proposed by Woese *et al.* (B).

Leaves of the tree are biological sequences from different species. Root node is mostly placed arbitrary because the algorithms for tree reconstruction cannot define this position automatically. Bifurcations on the tree show events of formation of two genes from one ancestral version during either gene duplication or species divergence. Branch lengths can be used to estimate the evolutionary distance (commonly calculated as a number of substitutions per one position in an alignment, or a *site*) between the leaf and its reconstructed parent node.

An important issue in phylogenetic studies is the reliability of the branches. *Bootstrap test* was introduced into the field (Felsenstein, 1985). Briefly, this test allows to create a required number (usually 100 or 1000) of artificial alignments of the same size as the original one which are constructed by random sampling of the characters (the whole columns of the alignment) with replacements, thus each of these alignments (also called *bootstrap replicas*) will contain some characters duplicated and some missing. Then the corresponding tree can be estimated for each replica, and thus each branch in original tree can be tested for the presence in these replicas. A bootstrap support value for the branch is then a percentage of replicas containing this branch.

In this thesis, the strategy of maximum likelihood was mostly used as the one showing the best results according to (Philippe *et al.*, 2005). In some cases, the approximate Likelihood Ratio Test (aLRT) was used instead of or in addition to the common bootstrap test (Anisimova and Gascuel, 2006). aLRT is implemented with the SH-like corrections in the

last version of the PhyML software (Guindon *et al.*, 2010). Further investigations suggest that this type of test is as powerful as a bootstrap test but is much faster (Anisimova *et al.*, 2011). We have used the LG model of amino acid substitutions (Le and Gascuel, 2008) which could be still not perfectly-suited for membrane proteins, but should be, at least, less biased towards globular water-soluble proteins, than the WAG matrix created on the basis of alignments containing at least one protein with known 3D structure by that time (Whelan and Goldman, 2001). The authors of LG matrix have created a large sample based on more than one and a half thousand Pfam families.

Altogether, these and other methods not described here are tools for *phylogenetic analysis* which is, in general, a reconstruction of sequences of evolutionary events based on a multiple alignment for any set of sequences.

### **2.5.2. Phylogenomic analysis**

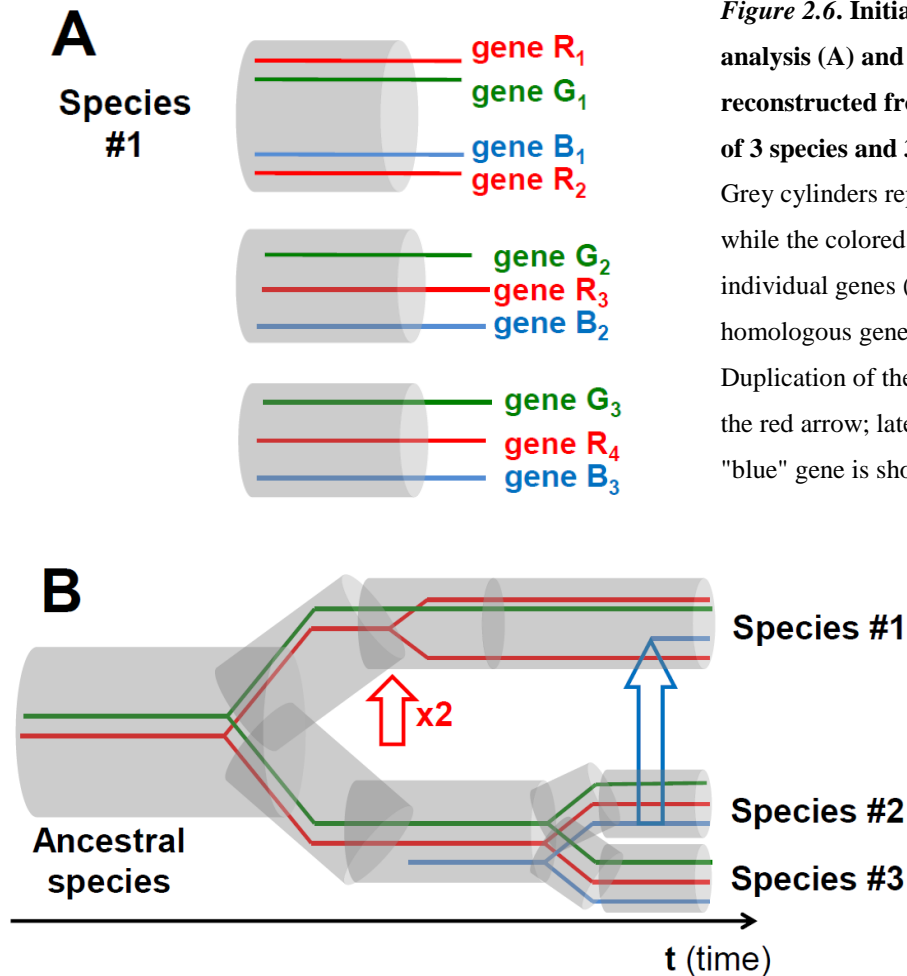
The approach used for analysis of each enzyme in this thesis is called *phylogenomic analysis*. It utilizes the whole methodology of phylogenetics as discussed in the sections above, but requires the use of sequences from fully sequenced genomes. At the first glance, this seems not to be a very important condition, but, in fact, only on the base of information on full genomes one can ask certain questions; answering these specific questions is the subject of phylogenomic analysis.

Simple analysis of similar sequences can be very useful for predicting functionally important residues in a particular protein. Accordingly, an analysis of a (small) set of specific genes, easily distinguishable from all other genes and occurring always in one copy per genome (as 16S rRNA), can be used for reconstruction of species taxonomy. However, evolutionary history of a gene itself (and thus history of a particular function associated with it) cannot be elucidated from the analysis of a random set of homologous sequences since such a set might contain sequences with different functions. Therefore, first all the possible homologs of this gene, as present in considered genomes, should be checked for their function. Also if incomplete genome information is being used, one cannot be sure on the absence/presence of a particular gene in a particular species and thus cannot trace the events of gene loss or gain. Reliable information on the absence of a gene and thus of its function is not only important

for evolutionary considerations. It could stimulate searching for alternative executors of the respective function (for instance, successful search for archaeal pathway of heme biosynthesis was encouraged by the absence of enzymes for several well-studied reactions of bacterial pathway (Storbeck *et al.*, 2010)).

The majority of the genes do not evolve through simple vertical inheritance (which is achieved through cell division that leads to species diversification and results in the emergence of genes called *orthologs*) and limited point mutations. Genes also undergo gene duplications (yielding rapidly diversifying *paralogs*), gene losses (sometimes followed by non-orthologous displacement) and lateral gene transfers. Thus, similarity between two genes, even if similarity represents their *homology*, common origin, does not indicate that these genes (or their products) have the same function. No algorithm is known that allows an automatic identification of homologous genes with the same function (orthologs), but still this task is simplified by following the clustering approach, as implemented in the COG (Clusters of Orthologous Groups) database (Tatusov *et al.*, 2003; Tatusov *et al.*, 1997). The procedure is based on comparing sequences from a completely deciphered genome with sequences from other full genomes and, concomitantly, with sequences from the same genome. This simple but powerful procedure of additional use of genomic information allows distinguishing between paralogs (proteins separated by inter-genomic duplication with typically different functions) and orthologs, which, in contrast, tend to preserve a common function.

Each COG is aimed to represent a set of orthologous proteins from different genomes, but due to overall complexity of the problem it sometimes also includes paralogs; even such high-quality and manually annotated system as COGs can only be used as a supporting tool upon actual analysis. **Figure 2.6** illustrates this complexity with a very simple example. Our initial data are the full genomes of different species each containing numerous separate genes (**Figure 2.6A** depicts only 3 genomes and 3 genes while the real number of only prokaryotic genomes exceeds 1500 and each of them usually contains more than 3000 genes).



*Figure 2.6. Initial data for phylogenomic analysis (A) and evolutionary scheme to be reconstructed from it (B) (in the example of 3 species and 3 genes).*

Grey cylinders represent species/genomes while the colored lines inside them are individual genes (each color shows homologous genes sharing the same origin). Duplication of the "red" gene is shown with the red arrow; lateral gene transfer of the "blue" gene is shown with the blue arrow.

Phylogenomic analysis could uncover the evolutionary history of the genes in the example shown in *Figure 2.6*.

- In a very restricted number of cases ("green" gene in *Figure 2.6*) each concerned genome contains one copy of a gene and individual phylogenetic tree of this gene generally follows taxonomic tree of organisms. The best way to explain such data is to suggest (1) vertical inheritance of this gene from the common ancestor of all species and (2) similar function for individual genes  $G_1$ ,  $G_2$  and  $G_3$ .
- However, the presence of only one gene copy in each genome is not sufficient to claim their vertical inheritance, as could be seen on example of the "blue" gene in *Figure 2.6*. Phylogenomic analysis of this gene will clearly show that genes  $B_2$  and  $B_1$  are remarkably more similar than  $B_2$  and  $B_3$ , while taxonomy of organisms in the case of vertical inheritance should suggest the opposite. Thus, instead of vertical inheritance, the investigator would rather propose lateral transfer. Laterally



- transferred genes in the new genome can undergo a functional shift: for example, a member of the long biochemical pathway (or a subunit of protein complex) after being transferred to a different genome, obviously, can no longer perform its original function alone. As we still observe this gene, we should propose that it was not dismissed in evolution and thus should be of some importance to the host genome.
- Finally, the presence of two similar "red" genes in the same genome (**Figure 2.6**) can only be explained by phylogenomic analysis, which also can be used to predict whether the function of genes  $R_1$  and  $R_2$  is the same or not. If, for instance, gene  $R_1$  is similar to  $R_3$  and  $R_4$  to much more extent than to  $R_2$ , this can be interpreted as a clue for ancestral gene duplication leading to escape of  $R_2$  from the evolutionary pressure (as its function could have been carried out by  $R_1$ ), its divergence and gaining of a new function. However, if  $R_1$  will be more similar to  $R_2$  than to other homologs, a rather recent duplication with preservation of the functions of both genes can be proposed. In both cases  $R_2$  would be paralogous to all other genes (which are orthologous), but the latter case would be hardly detectable by automatic procedure of identification of orthologs.

This example shows that the knowledge about orthologs and paralogs is important not only in evolutionary issues but also in solving one of the central problems of applied bioinformatics: the prediction of protein function.

## **2.6. Summary of classical bioinformatical algorithms, software and databases used in this study**

Initial search of the sequences in the database of fully sequenced genomes RefSeq (Pruitt *et al.*, 2007) was routinely performed with the PSI-BLAST algorithm, implemented in the BLAST package (Altschul *et al.*, 1997). Initial multiple alignments were reconstructed with the Muscle software (Edgar, 2004) and further improved using various sources of additional information (see Section 2.3 for details). Phylogenetic trees were reconstructed by positions aligned in most sequences from the sample, as this procedure was suggested to improve accuracy of the phylogenies (Talavera and Castresana, 2007). Regions of the alignment

formed by these positions (blocks) were mostly chosen manually. Phylogenetic trees were reconstructed with the PhyML 3.0 (Guindon *et al.*, 2010) or MEGA5 (Tamura *et al.*, 2011). MEGA5 was used as a visualizing program for all the trees. Protein domains were identified by similarity to known domains described in the Pfam database (Finn *et al.*, 2010). Detailed information on metabolic pathways was extracted from the KEGG database (Kanehisa *et al.*, 2010). Some data on the experimentally studied enzymes were obtained from the manually curated BRENDA database and checked in respective articles (Chang *et al.*, 2009). Logo diagrams were visualized with the WebLogo service (Crooks *et al.*, 2004). Protein 3D structures were obtained from the Protein Data Bank (PDB) (Berman *et al.*, 2000). Finally, for visualization of 3D structures of proteins we used the RasMol (Sayle and Milner-White, 1995) and Jmol (an open-source Java viewer for chemical structures in 3D; available at <http://www.jmol.org/>) software.

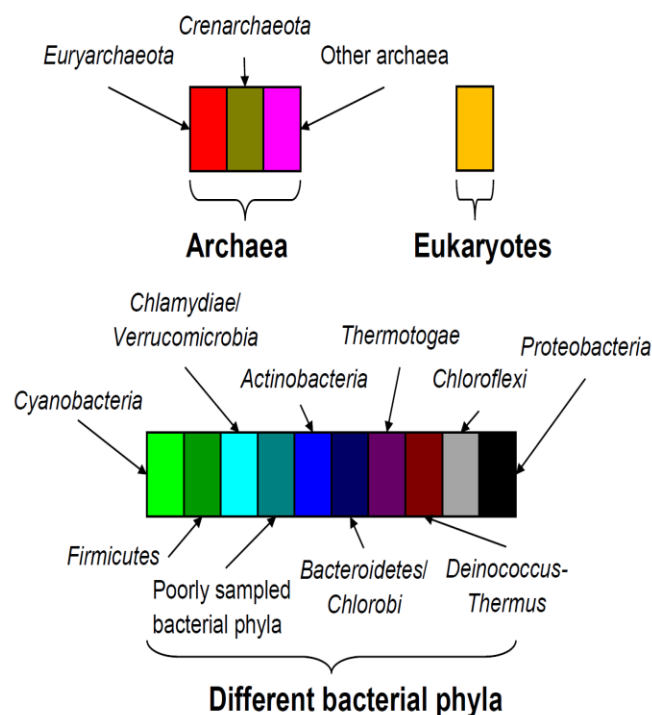
**A typical protocol that was applied to each protein that was analyzed in this study.**

Phylogenomic analysis of each protein in this thesis has routinely included the following steps: **(1)** identification of presumably homologous sequences in the database of full genomes, if possible – using COGs, **(2)** multiple alignment of these sequences with the manual correction and phylogenetic tree construction, **(3)** classification of the sequences on the basis of the tree and (if possible) other sequence features, such as the domain or operon structure, active site composition etc, **(4)** elucidation of the sequences which significantly differ from those described in literature, analysis of differences and, if possible, prediction of the protein function, **(5)** analysis of the occurrence of the studied proteins in different taxonomical groups, **(6)** reconstruction of possible ancestral function of the respective protein (or its components) and estimation of the stage of its appearance (before or after the divergence of bacteria and archaea).

In fact, only the first step of this "pipeline" could be automated to some extent (the available algorithms are described in the sections above). All other steps were performed manually for each protein.

## 2.7. Application of the described bioinformatical methods to the particular proteins

We relied upon the COGs for the phylogenomic analysis of some enzymes, as described for each case below. The assignment of proteins from the RefSeq release 45 (Jan 07, 2011) to COGs was obtained from the NCBI FTP site (<ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/Prok1202/>). The list of prokaryotic organisms was manually compacted by removing closely related species. This resulted in a list of 179 species (**Table S2**) (bacteria and archaea from all major phyla) used for a global analysis. Eukaryotic sequences from 35 sampled genomes (**Table S3**), belonging to all major taxonomic units of eukaryotes, were obtained, where necessary, by a search with HMMer (Finn *et al.*, 2011) using the PSSM constructed from alignment of prokaryotic members of the same COG. Phylogenetic trees were colored to show the taxonomic identity of each sequence. We have used the color code shown in **Figure 2.7**.



**Figure 2.7. Color code for the names of sequences sampled from different phyla of bacteria and archaea.**

The branches that led to eukaryotic sequences were coloured orange, to discriminate them from the archaeal and bacterial branches that were coloured black.

In our work we usually have used the RefSeq database, a comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and

protein sequences (Pruitt *et al.*, 2007). Specific smaller databases were routinely used for obtaining detailed annotation of proteins. For the needs of this study, well-annotated records from the UniProt protein database (UniProtConsortium, 2012) were also used.

### 2.7.1. Phylogenomic analysis of GTPases and ATPases

#### 2.7.1.1. EF-Tu and other translation factors

The representative sequences of translation factors EF-Tu/EF1, EF-G/EF2, IF1 and SelB/eIF2g from archaea, bacteria and eukaryotes were manually chosen using BLAST searches with the respective *E. coli* proteins as queries. These 12 proteins were initially aligned with the Muscle software and the alignment was manually improved using previously described multiple alignment of sequences from TRAFAC family (Leipe *et al.*, 2002). Sequences of several experimentally studied K<sup>+</sup>-dependent GTPases (their IDs are as follows: MNME\_ECOLI, YQEH\_BACSU, DER\_BACSU, NP\_228080 from *Thermotoga maritima* and YP\_139129 from *Streptococcus thermophilus*) were manually aligned with translation factors.

The Asp residue in position -3, related to the Lys of the P-loop, was conserved in all aligned translation factors. In order to check if the conservation is preserved also in other EF-Tu/EF1 proteins we have extracted members of COG0050 and COG5256 which describe EF-Tu and EF1 (respectively, 148 and 29 proteins). The multiple alignment for all these proteins was constructed with the Muscle software under default parameters. Manual curation of the alignment resulted in the removal of 3 proteins (*Table 2.1*).

**Table 2.1. Proteins removed from the sample upon manual curation of the multiple alignment of EF-Tu/EF1 homologs.**

#	GI	Organism	Cause of removal
1	15895803	<i>Clostridium acetobutylicum</i> ATCC 824	Unexpectedly short sequence
2	51473345	<i>Rickettsia typhi</i> str. Wilmington	Alignment is ambiguous
3	110637001	<i>Cytophaga hutchinsonii</i> ATCC 33406	Alignment is ambiguous

The remaining 174 proteins were re-aligned. The region of interest (P-loop) was unambiguously aligned and did not require a manual correction. The logo diagram was constructed for the positions from the -12 to +9 from the conserved lysine residue.

### 2.7.1.2. Recombination proteins RecA/RadA

We have extracted 227 members of COG0468 which describes RecA/RadA proteins. Then we added eukaryotic proteins selected as described above (totally 53 proteins) to the sample and performed multiple alignment with the Muscle software under default parameters. The archaeal experimentally studied protein with known 3D structure (PDB 1XU4 (Wu *et al.*, 2005)) belongs to *Methanococcus voltae*, which is absent in the list of 179 species, thus the corresponding sequence was added manually. Manual curation of the alignment resulted in the removal of 23 proteins (**Table 2.2**). The resulting 257 sequences were realigned with the Muscle software, and alignment was manually improved. Seven conserved blocks (total 141 positions) of the alignment were manually chosen for the tree construction. Phylogenetic tree was constructed with the PhyML by using SPR algorithm for the tree construction and under other parameters set to default values. The tree was visualized with the MEGA5 software; the names of prokaryotic sequences were colored by the corresponding taxonomical phyla (color code is shown on **Figure 2.7**).

**Table 2.2.** Proteins removed from the sample upon manual curation of the multiple alignment of RecA/RadA homologs.

#	GI	Organism	Cause of removal
1	285019368	<i>Xanthomonas albilineans</i> GPE PC73	Unexpectedly long sequence, alignment is ambiguous
2	301629465	<i>Xenopus (Silurana) tropicalis</i>	Unexpectedly long sequence, contains domain fusions
3	15827469	<i>Mycobacterium leprae</i> TN	Unexpectedly long sequence with loop-like insertions
4	15842276	<i>Mycobacterium tuberculosis</i> CDC1551	
5	171186310	<i>Pyrobaculum neutrophilum</i> V24Sta	Divergent, possibly paralogous group of proteins
6	159899943	<i>Herpetosiphon aurantiacus</i> DSM 785	

7	86158309	<i>Anaeromyxobacter dehalogenans</i> 2CP-C		
8	225873140	<i>Acidobacterium capsulatum</i> ATCC 51196		
9	225874745	<i>Acidobacterium capsulatum</i> ATCC 51196		
10	162450010	<i>Sorangium cellulosum</i> So ce56		
11	86158035	<i>Anaeromyxobacter dehalogenans</i> 2CP-C		
12	42522000	<i>Bdellovibrio bacteriovorus</i> HD100		
13	119719549	<i>Thermophilum pendens</i> Hrk 5		
14	15897679	<i>Sulfolobus solfataricus</i> P2		
15	288947751	<i>Allochromatium vinosum</i> DSM 180		
16	159897928	<i>Herpetosiphon aurantiacus</i> DSM 785		
17	258405739	<i>Desulfohalobium retbaense</i> DSM 5692		
18	226228137	<i>Gemmatimonas aurantiaca</i> T-27		Unexpectedly short sequence, alignment is ambiguous
19	119719654	<i>Thermophilum pendens</i> Hrk 5		
20	182415863	<i>Opitutus terrae</i> PB90-1		
21	124028477	<i>Hyperthermus butylicus</i> DSM 5456		
22	221635933	<i>Thermomicrobium roseum</i> DSM 5159		
23	238909168	<i>Eubacterium eligens</i> ATCC 27750		

### 2.7.1.3. Molecular chaperone GroEL

We have extracted 258 members of COG0459, which describes GroEL. Then we added eukaryotic proteins selected as described above (totally 295 proteins) to the sample and performed a multiple alignment with the Muscle software under default parameters. Manual curation of the alignment resulted in the removal of 11 proteins (**Table 2.3**). The resulting 542 sequences were realigned with the Muscle software, and alignment was manually improved. Six conserved blocks (total 317 positions) of the alignment was manually chosen for the tree construction. The phylogenetic tree was constructed with the PhyML software by

using the SPR algorithm for the tree construction and under other parameters set to default values. The tree was visualized with the MEGA5 software; the names of prokaryotic sequences were colored by the corresponding taxonomical phyla (color code is shown on *Figure 2.7*).

**Table 2.3. Proteins removed from the sample upon manual curation of the multiple alignment of GroEL homologs.**

#	GI	Organism	Cause of removal
1	30913055	<i>Schizosaccharomyces pombe</i>	Unexpectedly long sequence, alignment is partial
2	341941090	Mouse	
3	300669693	Human	
4	347595800	Yeast	
5	47115589	Fruit fly	
6	341940954	Mouse	Alignment is ambiguous
7	11133565	Human	
8	66773863	<i>Pongo abelii</i>	
9	3024693	Pig	Sequence is truncated
10	2501139	Pig	
11	1729866	Axolotl	

#### 2.7.1.4. GHKL superfamily

We have extracted a seed of the Pfam domain PF02518 that describes a GHKL superfamily of ATPases named after Gyrase, Hsp90, bacterial histidine and mitochondrial serine protein Kinases, DNA mismatch repair protein MutL. A seed of a Pfam domain is a representative subset of sequences from this domain which is constructed semi-manually by the maintainers of the database; it usually includes proteins that have known 3D structures and/or are experimentally characterized. The PF02518 domain is observed in nearly 130000 sequences while the seed contained only 700 proteins. We have improved the original Pfam alignment for this superfamily for the further use.

### 2.7.1.5. Membrane pyrophosphatases

We have extracted 70 members of COG3808, which describes the membrane pyrophosphatases. We added to them 14 proteins with experimentally studied ion specificity and dependence on  $K^+$  which belonged to organisms not included into the initial list. Another 2 experimentally characterized proteins (#9 and #16 in **Table 2.4**) were already included into the alignment. All the experimentally characterized proteins are listed in **Table 2.4**.

**Table 2.4.** List of proteins with known experimental properties (data is taken from (Luoto *et al.*, 2011)).

All proteins except #9 and #16 were manually added to the sample. The cells corresponding to  $H^+$ -translocating pyrophosphatases are filled with pink, the cells corresponding to  $Na^+$ -translocating pyrophosphatases are filled with blue. The  $K^+$ -dependence is shown by the grey filling of the corresponding column.

#	GI	Organism	Coupling ion	$K^+$ -dependence
1	3834302	<i>Arabidopsis thaliana</i>	$H^+$	No
2	6007754	<i>Pyrobaculum aerophilum</i>	$H^+$	No
3	7212770	<i>Rhodospirillum rubrum</i> ATCC 11170	$H^+$	No
4	21221966	<i>Streptomyces coelicolor</i> A3(2)	$H^+$	No
5	21226803	<i>Methanosarcina mazei</i> Go1	$H^+$	No
6	125973939	<i>Clostridium thermocellum</i> ATCC 27405	$H^+$	No
7	78043894	<i>Carboxydotherrmus hydrogenoformans</i>	$H^+$	Yes
8	183222788	<i>Leptospira biflexa</i> serovar Patoc	$H^+$	Yes
9	146299239	<i>Flavobacterium johnsoniae</i> UW101	$H^+$	Yes
10	21226802	<i>Methanosarcina mazei</i> Go1	$Na^+$	Yes
11	28210139	<i>Clostridium tetani</i> E88	$Na^+$	Yes
12	83590196	<i>Moorella thermoacetica</i> ATCC 39073	$Na^+$	Yes
13	95930454	<i>Desulfuromonas acetoxidans</i> DSM 684	$Na^+$	Yes
14	167745362	<i>Anaerostipes caccae</i> DSM 14662	$Na^+$	Yes
15	189346691	<i>Chlorobium limicola</i> DSM 245	$Na^+$	Yes
16	15642948	<i>Thermotoga maritima</i> MSB8	$Na^+$	Yes



The resulting 86 sequences were realigned with the Muscle software, no proteins were removed during manual curation, and the alignment was manually improved. Eleven conserved blocks (total 496 positions) of the alignment were manually chosen for the tree construction. The phylogenetic tree was constructed with the PhyML software by using the SPR algorithm for the tree construction and under other parameters set to default values. Tree was visualized with the MEGA5 software; names of prokaryotic sequences were colored by the corresponding taxonomical phyla (color code is shown on **Figure 2.7**).

## **2.7.2. Phylogenomic analysis of the N-ATPases**

### **2.7.2.1. Search for cyanobacterial ATP synthases subunits**

We have performed a PSI-BLAST search with the  $\alpha$ -subunits of *E. coli* ATP synthase (ATPA\_ECOLI) against a RefSeq release 39 (Jan 23, 2010) records belonging to *Cyanobacteria*. Sequences retrieved after third iteration were aligned with  $\alpha$ - and  $\beta$ -subunits of *E. coli* ATP synthase in order to sort them into these two groups. In five cyanobacteria, namely *Acaryochloris marina* MBIC11017 *Cyanothece sp.* ATCC 51142, *Synechococcus sp.* PCC 7002, *Cyanothece sp.* CCY0110 and *Nodularia spumigena* CCY9414 we have observed two sets of  $\alpha$ - and  $\beta$ -subunits. We have manually analyzed genome neighborhoods of these proteins in aforementioned organisms and documented different operon structure for the two groups. We then retrieved the *c*-subunits from operons of these organisms (totally 10 proteins) and aligned them with the sodium-specific *c*-subunits of ATP synthases of *Ilyobacter tartaricus* (ATPL\_ILYTA) and *Propionigenium modestum* (ATPL\_PROMO). The groups appeared to be also different in the predicted coupling ion specificity: sequences from one group contained a Na<sup>+</sup>-binding site, while sequences from the other group did not contain the respective ligands.

### **2.7.2.2. Search for the subunits of prokaryotic ATP synthases**

We have performed a global PSI-BLAST search with the  $\alpha$ -subunits of the *E. coli* ATP synthase (ATPA\_ECOLI) against a RefSeq release 39 (Jan 23, 2010) with a strict e-value threshold ( $10^{-30}$ ) in order to retrieve only  $\alpha$ -subunits. For 36 cases when more than two  $\alpha$ -

subunits were found in a single genome we have observed a division into two groups based on both operon structure and predicted coupling ion specificity. In this way the proteins of the N-ATPases were separated for the further analysis.

### **2.7.2.3. Concatenated phylogenetic tree construction**

In order to prove the separation of the retrieved N-ATPases from the common F<sub>1</sub>F<sub>0</sub>-type ATP synthases we have constructed a phylogenetic tree. The  $\alpha$ -,  $\beta$ -,  $\epsilon$ -,  $c$ - and  $b$ -subunits were obtained from the operons of N-ATPases and from the F-type ATPase operons of *Bacillus subtilis*, *Escherichia coli*, yeast and human. Subunits of each type were aligned separately, poorly aligned segments were removed, and the resulting alignments were concatenated producing a single sequence for each organism. The phylogenetic tree was constructed using the NJ algorithm implemented in the MEGA program. Bootstrap values were calculated from 100 samples.

### **2.7.3. Phylogenomic analysis of the cytochrome $b$**

The sequences of cytochromes  $b$  were identified in the full genomes sampled from RefSeq using PSI-BLAST. To reduce the number of genomes in the study, sequences from closely related species or different strains of one species have been excluded. Sequences of cytochromes  $b$  that are "split" into the N-terminal cytochrome  $b_6$ -like and C-terminal subunit IV-like part were artificially concatenated. The initial multiple sequence alignment was built with the Muscle software. The resulting alignment was then manually adjusted based on the two sequences of proteins with known 3D structures, the "split" cytochrome  $b$  of the cyanobacterial  $b_6f$  complex from *Mastigocladus laminosus*, PDB entry 1VF5 (Kurisu *et al.*, 2003), and the "long" bovine cytochrome  $b$ , PDB entry 1NTM (Gao *et al.*, 2003).

The blocks aligned in all sequences (3 blocks in the N-terminal part [170 positions] and 2 blocks in the C-terminal part [54 positions]) were used for the construction of the phylogenetic tree by using the PhyML software (with the SPR method for the search of the topology and 10 random trees to start with).

## 2.7.4. Multiple alignments of other components of the cytochrome $bc_1$ complex

### 2.7.4.1. Construction of multiple alignment for the cytochrome $c_1$

All sequences belonging to the COG2657, which describes cytochrome  $c_1$ , were extracted. As expected, they belong to the *Proteobacteria* and *Aquificae*. We have obtained the homologous proteins from 35 eukaryotic genomes listed in **Table S3**. Sequences of proteins from complexes with known 3D structures, namely from yeast (PDB 3CMX (Solmaz and Hunte, 2008)), bovine (PDB 1PP9 (Huang *et al.*, 2005)) *Paracoccus denitrificans* (PDB 2YIU (Kleinschroth *et al.*, 2011)) and *Rhodobacter sphaeroides* (PDB 2FYN (Esser *et al.*, 2006)) cytochrome  $bc_1$  complexes were added to the alignment. We have collapsed the list of proteobacterial species and further used only sequences from the genomes listed in **Table S4**. Manual curation of the multiple alignment of total 71 sequences resulted in the removal of 10 proteins (**Table 2.5**).

**Table 2.5. Proteins removed from the sample upon manual curation of the multiple alignment of cytochrome  $c_1$  homologs.**

#	GI	Organism	Cause of removal
1	27377597	<i>Bradyrhizobium japonicum</i> USDA 110	Unexpectedly long sequence, contains domain fusions
2	83313188	<i>Magnetospirillum magneticum</i> AMB-1	
3	6226530	Yeast	Unexpectedly long sequence, alignment is ambiguous
4	6226532	Yeast	
5	6226531	Yeast	
6	171184597	<i>Thermoproteus neutrophilus</i> V24Sta	
7	30249219	<i>Nitrosomonas europaea</i> ATCC 19718	
8	374287882	<i>Bacteriovorax marinus</i> SJ	Sequence is truncated
9	156366200	<i>Nematostella vectensis</i>	
10	156397929	<i>Nematostella vectensis</i>	

#### **2.7.4.2. Construction of multiple alignment for the subunit 8 (9.5 kDa subunit) of the eukaryotic *bc* complex.**

The 9.5 kDa subunit of the yeast cytochrome *bc*<sub>1</sub> complex was used as a query for the PSI-BLAST search against the RefSeq database. After the third iteration, no new hits were obtained, and totally 74 proteins were extracted. Only proteins which belong to the list of 35 species listed in Supplementary Table 3 were used for further analysis. The sequences of proteins from complexes with known 3D structures, namely from yeast (PDB 3CMX (Solmaz and Hunte, 2008)) and bovine (PDB 1PP9 (Huang *et al.*, 2005)) cytochrome *bc*<sub>1</sub> complexes were added to the alignment. The multiple alignment was built with the Muscle program and did not require a manual correction.

#### **2.7.4.3. Conservation of positively charged residues in the sequences of cytochromes *c*<sub>2</sub>.**

We have used a cytochrome *c* from the horse (CYC\_HORSE) to perform a psi-BLAST search against RefSeq database. After the third iteration 168 proteobacterial, 56 fungal and 209 metazoan sequences of cytochromes *c* were obtained. Multiple alignment, as constructed with the Muscle software, was used for production of logo diagrams.

### **3. Evolution of the mechanisms of phosphodiester bonds hydrolysis: from K<sup>+</sup> ions to the lysine or arginine "fingers"**

#### **3.1. Inorganic ion requirements of ubiquitous cellular systems**

As discussed in Section 1.1.3, the set of ubiquitous proteins which are coded in all genomes is likely to comprise the core of the LUCA proteins (Koonin, 2003). This set is presented in *Table 3.1*; apparently it is enriched with ribosomal proteins and with various ATPases. In *Table 3.1*, the list of ubiquitous proteins is supplemented with data on the inorganic ion requirements and affinities of the experimentally studied representatives of each enzyme family (for a full table with references see *Table S1*). In addition to the preference for Zn and Mn, which has been detected and discussed previously (Mulkidjanian and Galperin, 2009; Mulkidjanian and Galperin, 2010), we have found that several proteins and functional systems require K<sup>+</sup>, whereas none of the surveyed ancestral proteins specifically requires Na<sup>+</sup>. In this chapter we analyze the preference for K<sup>+</sup> in some more detail (the results were partly published in (Mulkidjanian *et al.*, 2012).

The preference of ubiquitous enzymes for K<sup>+</sup> provides support for the view that the first cells may have emerged in K<sup>+</sup>-rich environments, as first suggested by Archibald Macallum in 1926. He noted that, although similarities between seawater and organismal fluids, such as blood and lymph, indicate that the first animals emerged in the sea, the inorganic composition of the cell cytosol, which Macallum was first to determine, dramatically differs from that of modern sea water (Macallum, 1926). Specifically Macallum has pointed out the prevalence of potassium over sodium in cytosol; he speculated that the first cells could thrive in K<sup>+</sup>-rich environments. Macallum wrote "*...the very earliest organisms must have been of the micellar or ultramicroscopic kind... These had as yet no nuclei and an enclosing membrane could have been only of the most elementary character*" (Macallum, 1926). Indeed, the ion-tight membranes of modern cells are extremely complex energy conversion and transport systems that obviously are products of long evolution and were unlikely to exist in the first protocells. According to the available reconstructions, the first lipids were simple and single-tailed

(Deamer, 2008; Deamer and Dworkin, 2005; Gotoh *et al.*, 2007; Mulkidjanian and Galperin, 2010). The experiments with such lipid compounds have shown that vesicles made of fatty acids (Deamer and Dworkin, 2005; Mansy *et al.*, 2008) or of phosphorylated isoprenoids (Nomura *et al.*, 2001) can reliably entrap polynucleotides and proteins. Such membranes, however, are leaky to small molecules (Deamer, 2008; Nomura *et al.*, 2001). Hence, the membranes of first cells probably could occlude biological polymers and even facilitate their transmembrane translocation but could not prevent (almost) free exchange of small molecules and ions with the environment. Furthermore, before the emergence of diverse membrane translocators, the exchange of small molecules via leaky membranes should have been of vital importance for the first cells, which also implies that their interior was equilibrated with the surroundings, at least with respect to small molecules and ions (Deamer, 1997; Deamer, 2008; Mulkidjanian *et al.*, 2009; Nomura *et al.*, 2001; Szathmary, 2007; Szostak *et al.*, 2001; Szostak and Ricardo, 2009). According to Macallum *"there must have been an adjustment in the composition of very simple organisms to that of their medium,.... which diffusing into each minute multi-micellar mass brought into it the inorganic elements in the proportions in which they obtained in the external medium"* (Macallum, 1926). This idea was later generalized as a "chemistry conservation principle": the chemical features in the organisms (i.e. biochemical pathways or suitable ion concentrations) are more conserved than the changing environment and thus could retain information about ancient environmental conditions (Mulkidjanian and Galperin, 2007).

A potential alternative to this explanation is that the chemical differences between the intracellular milieu and the environment are unrelated to the conditions under which the first cells evolved (Dupont *et al.*, 2010). Then, the dramatic enrichment of modern cells for  $K^+$  could be viewed as a relatively late shift that came after the emergence of powerful ion-translocating membrane pumps and was driven by the growing demand of the newly evolving enzymes for particular inorganic ions as catalysts or substrates. Our analysis of the ubiquitous proteins helps to distinguish between these two possibilities. These proteins have developed their preferences for inorganic ion already at the stage of LUCA or even before and therefore are expected to provide information about the habitats of the first cells. Particularly informative are  $K^+$ -dependent translation factors EF-Tu and EF-G that result from duplication event that preceded LUCA (Atkinson *et al.*, 2008; Inagaki and Ford

Doolittle, 2000; Nakamura and Ito, 1998) (see Section 1.1.3 above) and, therefore, carry information about the pre-LUCA times. Based on data from **Table 3.1**, a search for geochemical conditions with  $K^+/Na^+ > 1$  and high levels of phosphate and transition metals has been performed. The only environments with such properties were found to be geothermal fields overlaying the vapor dominated zones of inland geothermal fields (Mulkidjanian *et al.*, 2012). This finding has prompted a scenario on the origin of first cells at anoxic geothermal fields, see Section 7.3 and (Mulkidjanian *et al.*, 2012) for details.

**Table 3.1. Products of ubiquitous genes and their association with essential inorganic cations and anions.**

The lists of ubiquitous genes were extracted from refs. (Charlebois and Doolittle, 2004; Koonin, 2000). The data on the dependence of functional activity on particular metals were taken from the BRENDA database (Chang *et al.*, 2009). According to the BRENDA database, the enzymatic activity of most  $Mg^{2+}$ -dependent enzymes could be routinely restored by  $Mn^{2+}$ . As concentration of  $Mg^{2+}$  ions in the cell is ca.  $10^{-2}$  M, whereas that of  $Mn^{2+}$  ions is ca.  $10^{-6}$  M, the data on the functional importance of  $Mn^{2+}$  were not included in the table for many enzymes. The presence of metals in protein structures was as listed in the PDB entries. The table includes all enzymes represented by orthologs in all cellular life forms as well as several cases when a function is ubiquitous (e.g., DNA polymerase, DNA primase) whereas the enzymes responsible for that function are represented by two or more nonorthologous forms (Koonin, 2003). Upward arrows indicate the activation by the particular ion and downward arrows indicate the inhibition by this ion. If low ion concentrations activate the enzyme while high amounts of the same ion cause its inhibition then the sign  $\uparrow\downarrow$  is used.

Protein function	EC number (if available)	Functionally relevant inorganic anions	Monovalent cations		Divalent cations	
			Functional dependence	Presence in at least some structures	Functional dependence	Presence in at least some structures
<b>Translation and ribosomal biogenesis</b>						
Ribosomal proteins	-	-	-	* $MC^+$	-	$Mg^{2+}$ , $Cd^{2+}$ , $Zn^{2+}$
Conserved translation factors (EF-G, EF-Tu, IF-1, IF-2, eIF5- a)	3.6.5.3	$PO_4^{3-}$	$\uparrow K^+$ , $NH_4^+$ , $\downarrow Na^+$	-	$Mg^{2+}$	$Mg^{2+}$ , $Zn^{2+}$
Most tRNA synthetases	6.1.1.-	$PP_i$	$\uparrow K^+$ , $\downarrow Na^+$	$K^+$	$Mg^{2+}$ , $Zn^{2+}$	$Mg^{2+}$ , $Zn^{2+}$
Pseudouridylate synthase	5.4.99.1 2	$PO_4^{3-}$	-	$K^+$	$Mg^{2+}$ , $Zn^{2+}$	$Mg^{2+}$ , $Zn^{2+}$
Methionine aminopeptidase	3.4.11.1 8	-	-	$MC^+$	$Fe^{2+}$	$Mn^{2+}$ , $Zn^{2+}$

<b>Transcription</b>						
DNA-directed RNA polymerase [ $\alpha$ , $\beta$ , $\beta'$ ]	2.7.7.6	$\text{PO}_4^{3-}$	-	$\text{Na}^+$	$\text{Mg}^{2+}$ , $\text{Zn}^{2+}$	$\text{Mg}^{2+}$ , $\text{Mn}^{2+}$ , $\text{Zn}^{2+}$
<b>Replication</b>						
Clamp loader ATPase (pol III, subunit $\gamma$ and $\tau$ )	2.7.7.7	$\text{PO}_4^{3-}$	-	-	$\text{Mg}^{2+}$	$\text{Mg}^{2+}$ , $\text{Zn}^{2+}$
Topoisomerase IA	5.99.1.2	-	$\text{MC}^+$	-	$\text{Mg}^{2+}$	$\text{Zn}^{2+}$ , $\text{Hg}^{2+}$
<b>Repair and Recombination</b>						
5'-3' exonuclease (including N- terminal domain of PoII)	3.1.11.-	$\text{PO}_4^{3-}$	-	-	$\text{Mg}^{2+}$	$\text{Mn}$ , $\text{Zn}^{2+}$
RecA/RadA (Rad51) recombinase	-	$\text{PO}_4^{3-}$	$\text{K}^+$	$\text{K}^+$	$\text{Mg}^{2+}$	$\text{Mg}^{2+}$
<b>Chaperone function</b>						
Chaperonin GroEL	3.6.4.9	$\text{PO}_4^{3-}$	$\text{K}^+$	$\text{K}^+$	$\text{Mg}^{2+}$	$\text{Mg}^{2+}$
<b>Nucleotide and amino acid metabolism</b>						
Thymidylate kinase	2.7.4.9	$\text{PO}_4^{3-}$	-	$\text{Na}$	$\text{Mg}^{2+}$	$\text{Mg}^{2+}$
Thioredoxin reductase	1.8.1.9	-	-	-	-	$\text{Mg}^{2+}$
Thioredoxin	-	-	-	-	-	$\text{Cd}^{2+}$ , $\text{Zn}^{2+}$
CDP-diglyceride-synthase	2.7.7.41	$\text{PO}_4^{3-}$	$\uparrow\text{K}^+$ , $\downarrow\text{Na}^+$	No entries	$\text{Mg}^{2+}$	No entries
<b>Energy conversion</b>						
Phosphomannomutase	5.4.2.8	-	-	-	$\text{Mg}^{2+}$	$\text{Mg}^{2+}$ , $\text{Zn}^{2+}$
Catalytic subunit of the membrane ATP synthase	3.6.3.14	$\text{PO}_4^{3-}$	-	-	$\text{Mg}^{2+}$	$\text{Mg}^{2+}$
Proteolipid subunits of the membrane ATP synthase	3.6.3.14	-	-	-	-	-
<b>Coenzymes</b>						
Glycine hydroxymethyltransferase	2.1.2.1	-	$\downarrow\text{MC}^+$	-	$\downarrow\text{Mg}^{2+}$ , $\text{Mn}^{2+}$ , $\text{Ca}^{2+}$	-
<b>Secretion</b>						
Preprotein translocase subunit SecY	-	-	-	-	-	$\text{Zn}^{2+}$
Signal recognition particle GTPase FtsY	3.6.5.4	$\text{PO}_4^{3-}$	-	$\text{K}^+$	$\text{Mg}^{2+}$	$\text{Mg}^{2+}$



Miscellaneous						
Predicted GTPase (YchF, PF06071, 1JAL, 2OHF, 2DBY, 2DWQ, 1NI3)	-	PO <sub>4</sub> <sup>3-</sup>	K <sup>+</sup>	-	Mg <sup>2+</sup>	-
DNA primase (dnaG)	2.7.7.-	PP <sub>i</sub>	-	-	Zn <sup>2+</sup>	-
S-adenosylmethionine dimethyltransferase (KsgA)	2.1.1.48	-	-	-	Mg <sup>2+</sup>	-

### 3.2. Potassium dependence of the enzymes catalyzing reactions of phosphate group transfer

All the ubiquitous proteins that show dependence on potassium ions are either GTPases or ATPases. It is noteworthy that as early as in 1960, Lowenstein, while studying, in the presence of Mg<sup>2+</sup> or Mn<sup>2+</sup> as divalent cations, the non-enzymatic transphosphorylation reaction  $ATP + P_i \rightarrow ADP + PP_i$  has shown its three-fold stimulation by large monovalent cations (K<sup>+</sup>, Rb<sup>+</sup> or NH<sub>4</sub><sup>+</sup>), while the smaller cations (Na<sup>+</sup>, Li<sup>+</sup>) did not affect the reaction rate (Lowenstein, 1960). In case of the archaeal RadA protein, it was demonstrated that K<sup>+</sup> ions are occupying the same position as the catalytically important Lys residue in the homologous bacterial enzyme RecA (Wu *et al.*, 2005). The suggested emergence of the first cells in potassium-rich environments could imply the initial involvement of K<sup>+</sup> ions as catalysts upon the cleavage of phosphoester bonds in the first organisms, followed by replacement of K<sup>+</sup> ions by amino groups of amino acid sidechains, e.g. arginine or lysine "fingers" (Scheffzek *et al.*, 1997), in the course of evolution.

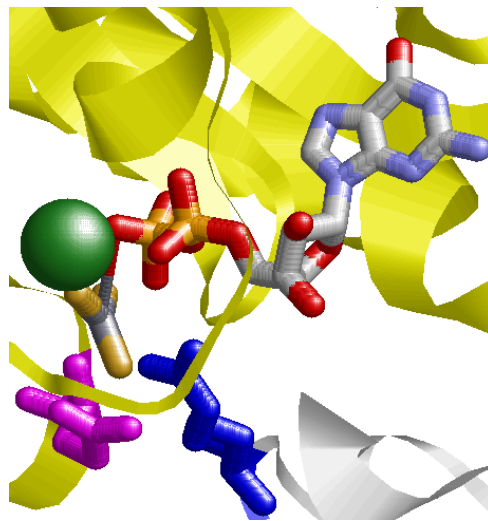
The enzymes that participate in transformations of phosphoester bonds belong to a number of unrelated groups of proteins, see (Gorbalenya and Koonin, 1990; Milner-White *et al.*, 1991; Saraste *et al.*, 1990; Schweins and Wittinghofer, 1994; Vetter and Wittinghofer, 1999). These protein (super)families contain the most widespread protein folds, such as:

- the mononucleotide-binding fold (P-loop NTPases);
- the protein kinase fold;
- the histidine kinase, topoisomerase II and chaperone Hsp90 fold;
- the fold of RNase H and chaperone Hsp70.

In subsequent sections we analyze a number of unrelated or distantly related  $K^+$ -dependent hydrolases that belong to these families. We show that, within protein families,  $K^+$  ions could be functionally and structurally replaced by positively charged residues (one/two either Arg or Lys) in the course of evolution.

### 3.2.1. Example #1: P-loop GTPases

The P-loop NTPase fold is the most popular protein fold in a majority of cellular organisms and comprises up to 18% of all gene products (Koonin *et al.*, 2000). The  $K^+$ -dependent ribosomal GTPases of the ubiquitous protein set (see Table 3.1) belong to this protein family together with the well-known oncogenic Ras protein (Fernandez-Medarde and Santos, 2011). In the Ras GTPase, the best studied of the P-loop GTPases, the hydrolysis of GTP is controlled by the interaction between the GTPase and the GTPase Activating Protein (GAP) (*Figure 3.1*).



*Figure 3.1.* "Arginine finger" residue provided by the GAP protein in the active site of the Ras GTPase (PDB 1WQ1 (Scheffzek *et al.*, 1997)).

The Ras protein is colored yellow, the activating GAP protein is colored white. The arginine finger is colored blue, ATP is colored by atoms and shown as a wireless model, the  $Mg^{2+}$  ion is colored green and is shown as a sphere of Van-der-Waals radius. The glutamate residue which is mediating hydrolysis is colored magenta.

The transition state is stabilized by the  $Gln^{cat}$  residue of Ras while insertion of the  $Arg^{GAP}$  residue of GAP (so-called "arginine finger") stimulates the hydrolysis. But this mechanism is not applicable to a large number of different P-loop GTPases which lack the catalytic  $Gln^{cat}$

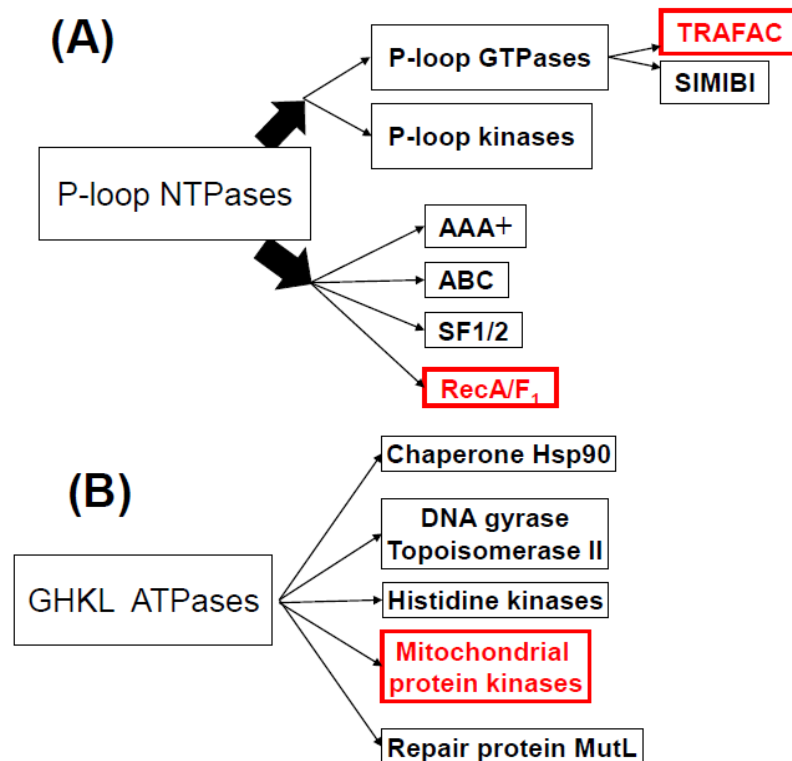
residue and do not require an additional protein (i.e. GAP) for the effective hydrolysis of GTP (Mishra *et al.*, 2005). These proteins were named HAS-GTPases after the Hydrophobic Amino Acid Substitution which takes place in the Gln<sup>cat</sup> position.

For a number of proteins referred to as HAS-GTPases, the hydrolysis of GTP was found to be potassium-dependent but not sodium-dependent. These include the ribosome assembly GTPase YqeH (Anand *et al.*, 2010), the essential protein EngA (Foucher *et al.*, 2012) from *Bacillus subtilis*, the tRNA modification GTPase MnmE (Scrima and Wittinghofer, 2006) from *E. coli*, the TrmE GTPase (Yamanaka *et al.*, 2000) from *Thermotoga maritima*, and the G-protein-coupled ferrous iron transporter FeoB (Ash *et al.*, 2010) from *Streptococcus thermophilus*. Binding sites for K<sup>+</sup> and its potential ligands were described together with the possible mechanism of activation where K<sup>+</sup> ion plays the same role as "arginine finger" in Ras (Anand *et al.*, 2010; Ash *et al.*, 2010; Foucher *et al.*, 2012; Scrima and Wittinghofer, 2006; Yamanaka *et al.*, 2000).

As early as 1964 it has been shown that the rate of GTP hydrolysis and synthesis of polyphenylalanine in the ribosome-containing extract of the *E. coli* cells depend on the concentration of K<sup>+</sup> and NH<sub>4</sub><sup>+</sup>, but not Na<sup>+</sup> (Conway and Lipmann, 1964). Translation factor EF-Tu from the ubiquitous gene set was identified as a K<sup>+</sup>-dependent GTPase: the amount of hydrolysed GTP was shown to be considerably more effectively increased by K<sup>+</sup> and NH<sub>4</sub><sup>+</sup> than by Na<sup>+</sup> (Fasano *et al.*, 1982). However, the mechanism of the activation of EF-Tu by monovalent cations, to the best of our knowledge, has remained unknown.

The K<sup>+</sup>-dependent translation factors belong to the same superfamily of the P-loop GTPases (TRAFAC) as the HAS-GTPases (see **Figure 3.2A**). We have aligned the sequences of translation factors from archaea, bacteria and eukaryotes; the alignment of representative sequences of initiation factors (IF2 and eIF2g) and elongation factors (EF-Tu and EF-G) in the most conserved motifs is shown on the bottom of **Figure 3.3**. Aforementioned K<sup>+</sup>-dependent GTPases are shown on the top of **Figure 3.3**. The asparagine residue, a crucial K<sup>+</sup> ligand, shown by the yellow arrow in **Figure 3.3**, is replaced with aspartic acid in all translational factors. However, mutation of this asparagine to the aspartic acid were shown to affect GTPase activity only slightly, thus both could be present in functionally active proteins (Anand *et al.*, 2010). Structure of the possible K<sup>+</sup> binding site, as inferred from visualization of corresponding residues in the structure of EF-Tu factor, is also similar to the structure of

the  $K^+$ -binding site in the structure of MnmE GTPase (*Figure 3.4*). Thus, the translation factors, most likely, could be activated by potassium ions via the same mechanism as is utilized by several characterized  $K^+$ -dependent P-loop GTPases.



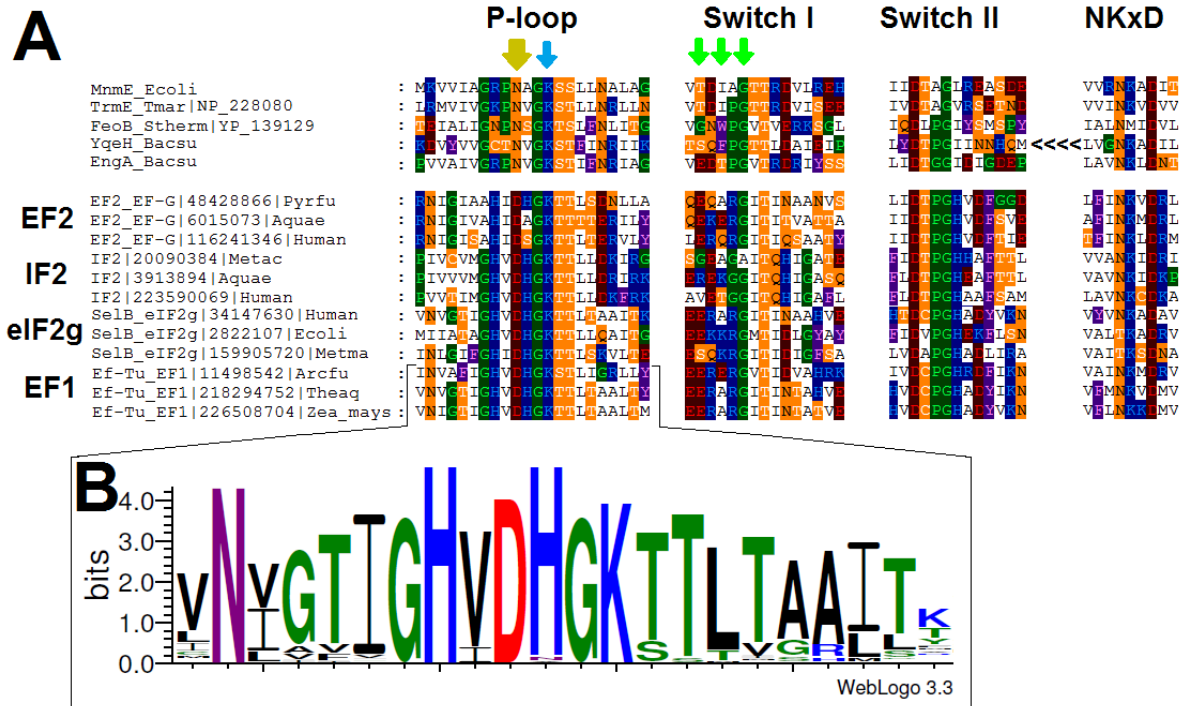
**Figure 3.2. Major subfamilies of the two widespread NTPase domains.**

(A) P-loop NTPases, kinase-GTPase group (top) and ASCE group (bottom), the red frames show positions of RecA/F<sub>1</sub> family (catalytic subunits of ATP synthase and recombinase RecA belong to this family) and TRAFAC superfamily (named after translational factors).

(B) GHKL ATPases, the red frame highlights mitochondrial protein kinases.

During previous phylogenetic analysis of the P-loop GTPases family (Leipe *et al.*, 2002), several protein subfamilies were predicted to be evolutionarily old. Among those are the translation factors themselves. In other evolutionarily old proteins (obg, YyaF/Ygr210, HflX, YyaW) the Asn residue that corresponds to the potassium-binding residue (marked with yellow in the *Figure 3.3*) is also mostly conserved (see Figure 2 in (Leipe *et al.*, 2002)).

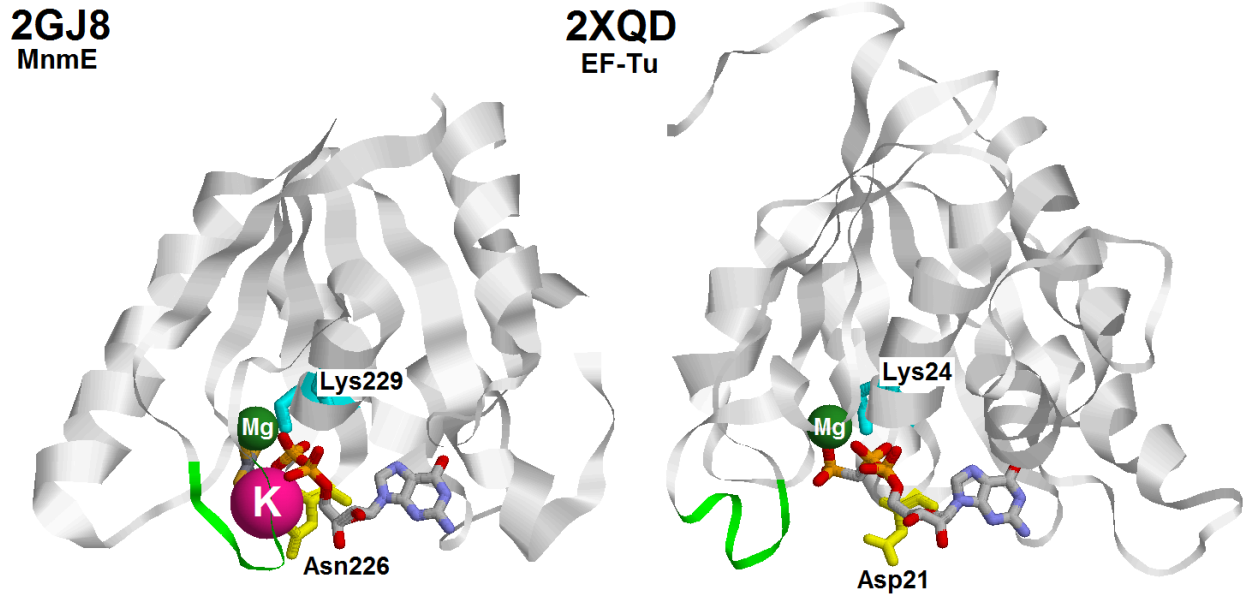
Hence, our analysis indicates that in the family of the P-loop GTPases the activation of catalysis by potassium ions looks as an evolutionary ancient trait that was independently replaced by activation via side chain amino groups in several lineages.



**Figure 3.3. Multiple alignment of the potassium-dependent GTPases and different translation factors from archaea, bacteria and eukaryotes.**

(A) Amino acid residues are colored according to the chemical nature of their side chains (basic residues are in blue, acidic residues are in red, polar residues are in orange, non polar aliphatic residues are in white, aromatic residues are in violet, proline and glycine residues are in green). This color code for the amino acid residues will be further used throughout the thesis. The bold yellow arrow shows the key residue that is important for  $K^+$  binding (Asn→Asp in this position in YqeH didn't affect much the GTPase activity while Asn→Leu or Asn→Gln abolished both the activity and its potassium-dependence (Anand *et al.*, 2010)). One of the green arrows shows the glycine residue that is not directly involved in the binding, but seems to provide required flexibility of the chain in this region. The two residues marked with further green arrows present their backbone oxygen residues for binding, thus their nature is not of much importance. The YqeH protein is circularly permuted and its NKxD motif is located at the N-terminus of the protein (this fact is shown by the <<<< sign).

(B) The Logo diagram for 174 EF-Tu and EF1 proteins sampled as described in Section 2.7.1.1. Logo was constructed for the positions from the -12 to +9 as counted relative to the position of the conserved lysine residue in the P-loop. The aspartic acid residue in the position -3 is absolutely conserved within this family.



**Figure 3.4.** Structures of the P-loop GTPase domains of MnmE protein (left) and translation elongation factor Tu (right).

Proteins are shown in ribbon representation. GDP and AlF<sub>3</sub> (together forming a GTP analogue) in the PDB entry 2GJ8 (Scrima and Wittinghofer, 2006) and GCP (guanosine 5'-( $\alpha,\beta$ -methylene) triphosphate, a GTP analog) in the PDB entry 2XQD (Voorhees *et al.*, 2010) are colored by atoms while the metal ions are shown as spheres (Mg<sup>2+</sup> in green, K<sup>+</sup> in pink). The green loops correspond to the region marked with green arrows in the **Figure 3.3** (Switch I region). The lysines of the P-loops are shown in cyan. The Asn residue considered to be responsible for the binding of the K<sup>+</sup> ion is shown in yellow in the MnmE structure. The corresponding Asp residue in the EF-Tu structure is also colored yellow.

### 3.2.2. Example #2: RadA and RecA

RadA and RecA are ATPases that belong to a different subfamily of P-loop NTPases than the previously discussed P-loop GTPases (**Figure 3.2A**). These ATPases are called ASCE (Additional Strand, Catalytic E) ATPases as they bear an additional strand between the P-loop (Walker A) and the Walker B motifs and have a conserved, proton abstracting Glu residue acting upon hydrolysis (Leipe *et al.*, 2003).

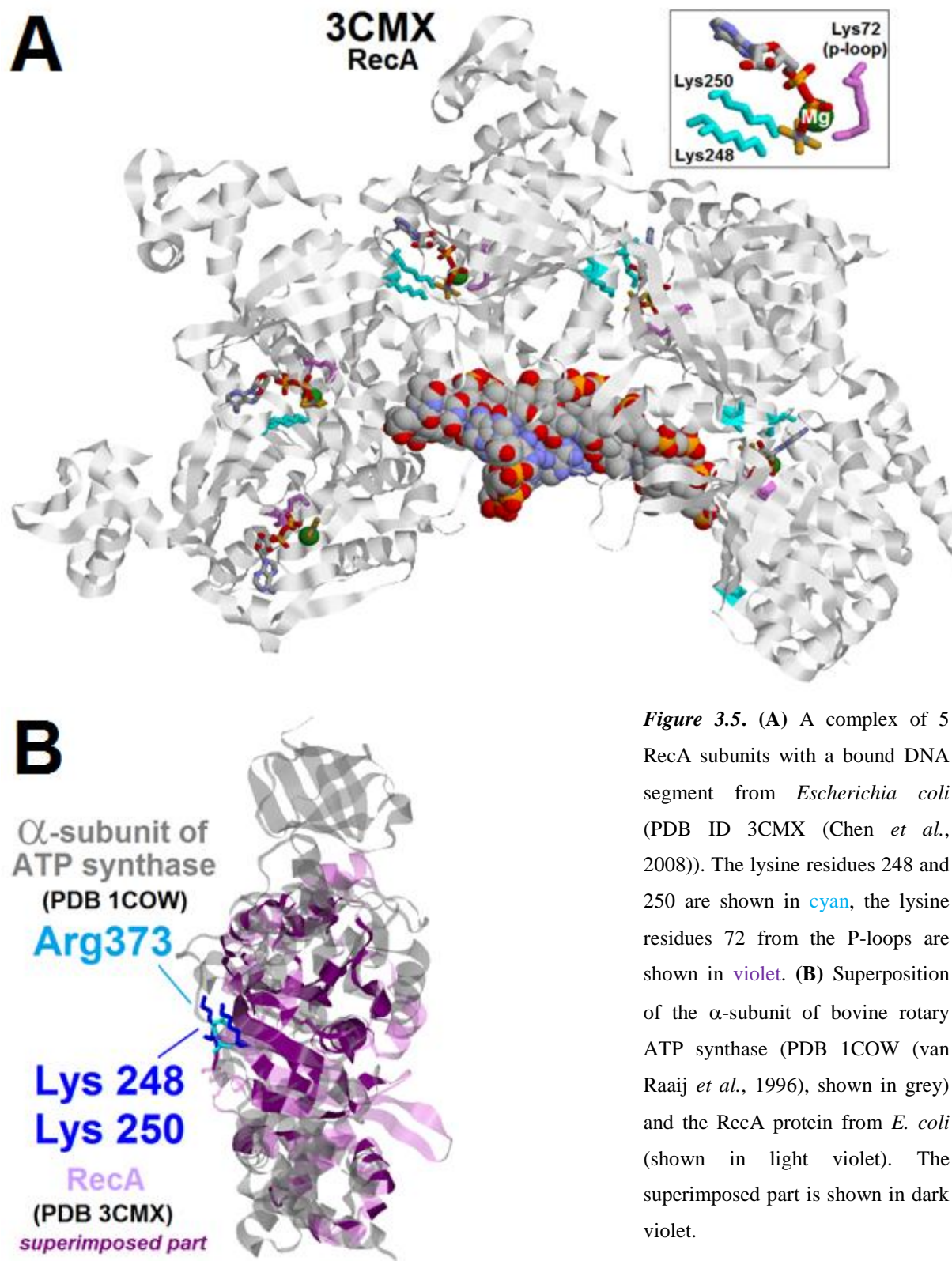
The activity of the RadA protein from an archaeon *Methanococcus voltae* showed strong potassium dependence: almost no P<sub>i</sub> (inorganic phosphate) release was observed without K<sup>+</sup> added to the media (Wu *et al.*, 2005). The crystal structure of RadA (PDB ID 1XU4) shows two binding sites for potassium ions (see Figure 2 from (Wu *et al.*, 2005)). One ion is bound

in the same orientation relative to ATP and  $Mg^{2+}$  as in the P-loop GTPases (the same orientation was theoretically proposed by Lowenstein in 1960 (Lowenstein, 1960)). Interestingly, if no  $K^+$  is added to the medium, the binding sites are occupied with water molecules, as was already observed for the FeoB GTPase (Ash *et al.*, 2010).

The bacterial homolog of Rad is a recombinase RecA. The crystal structure of RadA from *Methanococcus maripaludis* (PDB ID 3EW9) was compared with the crystal structure of *Escherichia coli* RecA (Li *et al.*, 2009). After superposition of P-loops (see Figure 4 from (Li *et al.*, 2009)), the  $K^+$  ions in RadA lay over the lysine residues K248 and K250 of RecA which are important for the catalysis (Chen *et al.*, 2008). Although these residues are located on the surface of RecA monomer far away from the ATP binding site, they are inserted into the ATP binding site of the next monomer in the 3D structure of the complex with DNA (PDB 3CMX (Chen *et al.*, 2008), see **Figure 3.5A**).

Our phylogenetic tree for the RecA/RadA superfamily shows that bacteria mostly have two lysine residues, whereas archaea and eukaryotes (except for those few which received this protein as a result of lateral gene transfer from bacteria) have conserved aspartic acid (Asp302 in *M. voltae*) in the same place in the alignment (see **Figure 3.6** for the schematic representation of the tree and **Figure 3.7** for the sample from multiple alignment). This residue was shown to directly bind one potassium ion with its side chain (Wu *et al.*, 2005). Another residue which binds the second  $K^+$  ion with its side chain is almost absolutely conserved (Glu151 in *M. voltae*) and does not determine the potassium dependence (Qian *et al.*, 2006).

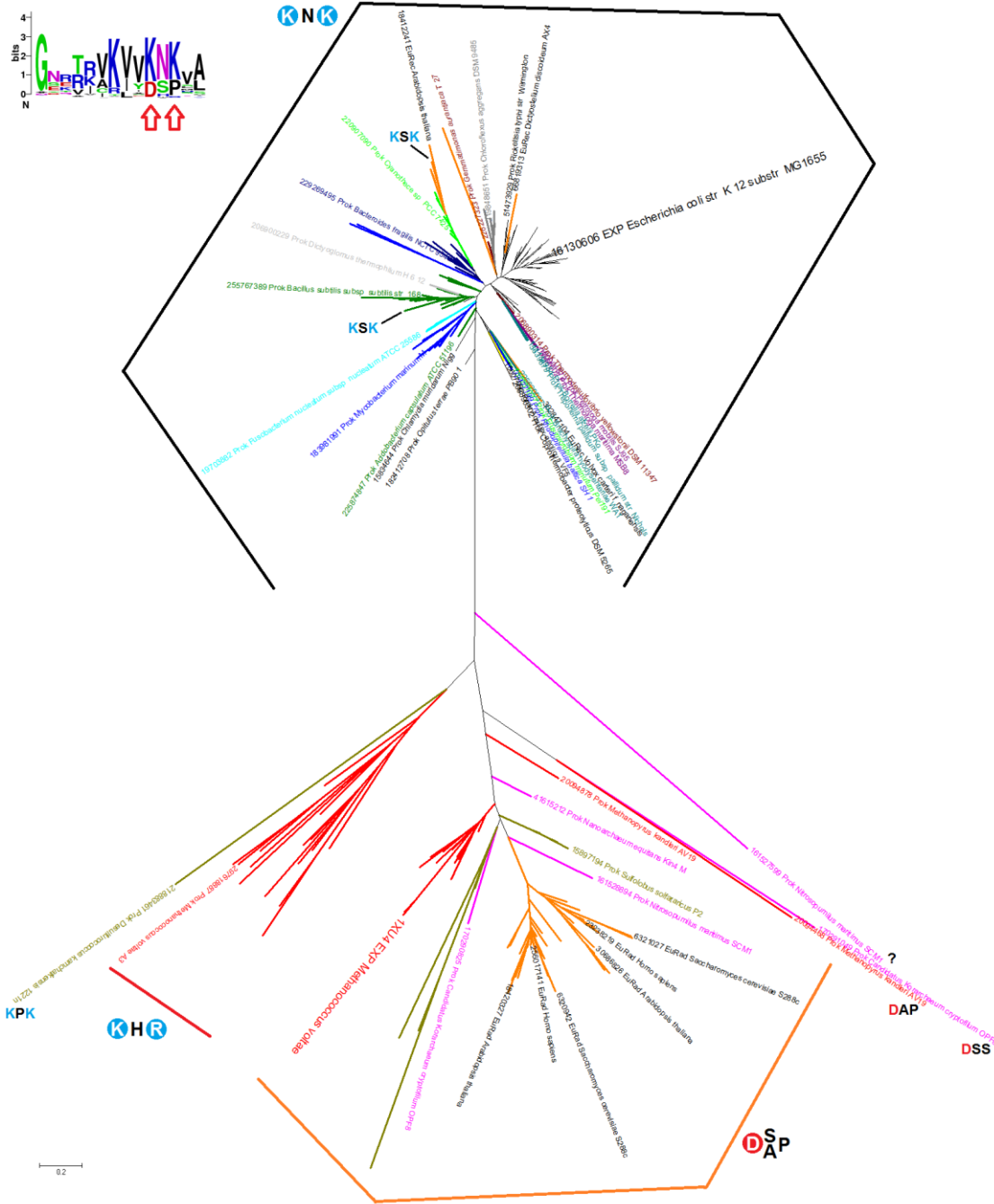
The strict distinction between the bacterial and archaea/eukaryotic clades does not allow making a clear statement whether the primordial form of this enzyme contained a potassium-binding aspartic acid or a lysine pair. However, there is at least one indication in favor of the primacy of the potassium-binding site.



**Figure 3.5.** (A) A complex of 5 RecA subunits with a bound DNA segment from *Escherichia coli* (PDB ID 3CMX (Chen *et al.*, 2008)). The lysine residues 248 and 250 are shown in cyan, the lysine residues 72 from the P-loops are shown in violet. (B) Superposition of the  $\alpha$ -subunit of bovine rotary ATP synthase (PDB 1COW (van Raaij *et al.*, 1996), shown in grey) and the RecA protein from *E. coli* (shown in light violet). The superimposed part is shown in dark violet.

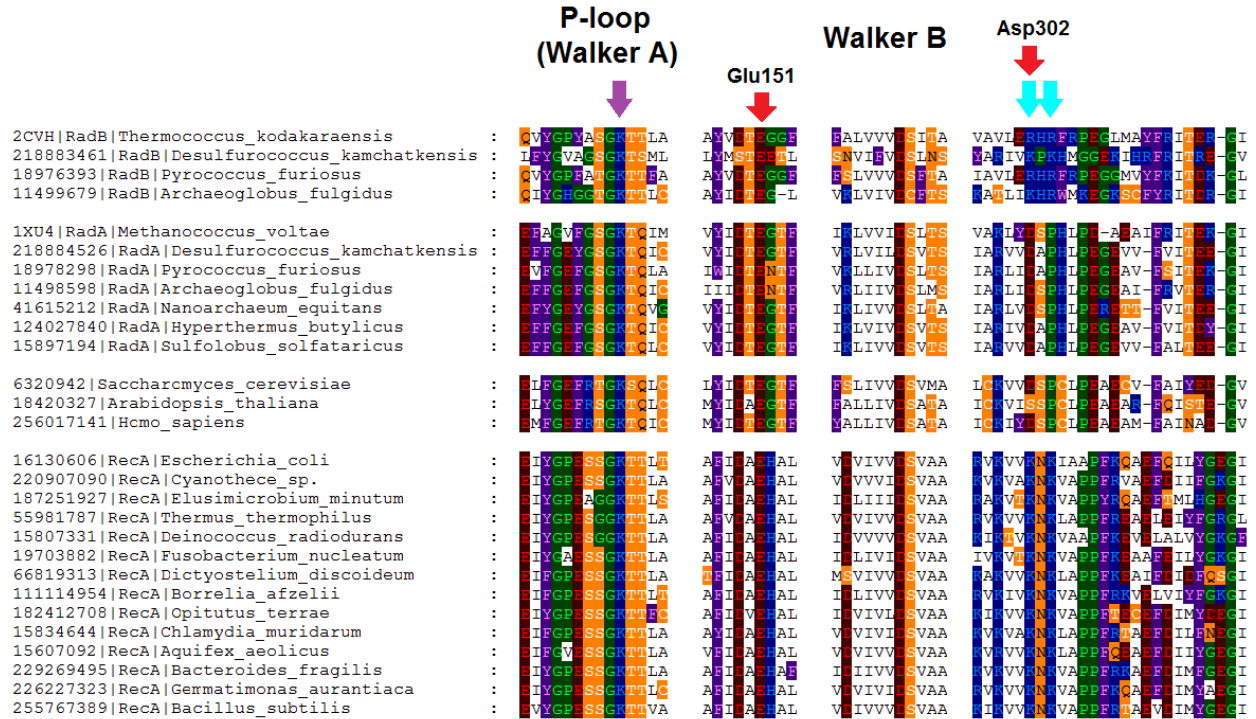


Euryarchaeal genomes all contain two sequences, one of which groups in the clade with eukaryotic and other archaeal sequences (RecA), whereas the other forms a distinct branch on the tree (RadB (Haldenby *et al.*, 2009; Sandler *et al.*, 1999)). This seems to be a result of ancient gene duplication (Lin *et al.*, 2006). In RadB proteins, the aspartic acid is replaced with two positively charged residues; they, however, could be either lysine and arginine or two arginine residues in contrast to the conserved pair of lysine residues in all bacterial RecA. The sequence from the creanarchaeon *Desulfurococcus kamchatkensis* groups with the same clade. The ATPase activity of the RadB from *Pyrococcus furiosus* does not increase with DNA binding and remains more than 5 times lower than the ATPase activity of the RadA even in the absence of DNA (Komori *et al.*, 2000). The [K,R]HR motif was previously proposed to play a role in the DNA binding as the alanine mutations in this motif led to a greatly reduced DNA binding (Guy *et al.*, 2006). However, later the structure of *E. coli* RecA in complex with DNA has shown that the two lysine residues alignable with this motif are directly in contact with the ATP molecule (Chen *et al.*, 2008). A structural superposition of bovine ATP synthase  $\alpha$ -subunit and RecA from *E. coli*, as performed with the protein structure comparison service Fold at European Bioinformatics Institute (<http://www.ebi.ac.uk/msd-srv/ssm>) (Krissinel and Henrick, 2004), shows that their core elements including all  $\beta$ -strands of the large  $\beta$ -sheet could be well aligned (**Figure 3.5B**). A single Arg residue ("arginine finger") is absolutely conserved in all  $\alpha$ -subunits; in RecA proteins, exactly the same space is occupied by two lysine residues. Thus, in homologous proteins from subfamilies of RecA/RadA/Rad51 and  $\alpha/\beta$ -subunits of the ATP synthase, different possible arrangements of positive residues are observed (KNK motif in bacterial RecA, KHR motif in RadB and a single arginine residue in  $\alpha$ - and  $\beta$ -subunits of ATP synthase), while the  $K^+$ -binding motif with a key Asp residue remains conserved in archaeal RadA and eukaryotic Rad51. Thus, within the family that unites RecA/RadA/Rad51 proteins and  $\alpha/\beta$ -subunits of rotary ATPases, a conserved  $K^+$ -binding site appears to precede the different arrangements of the positively charged residues in evolution.



**Figure 3.6. Scheme of phylogenetic tree of the RadA/RecA subfamily.**

The sample for the alignment construction and phylogenetic tree reconstruction consisted of members of COG0468 from a set of 179 genomes from all bacterial and archaeal phyla and RadA/RecA orthologs (Rad51, Dmc1) from 35 eukaryotic species (*Table S2* and *Table S3*).



**Figure 3.7. Multiple alignment of RecA/RadA/RadB and their Rad51-like eukaryotic homologs.**

Amino acid residues are colored in a same way as in **Figure 3.3**.

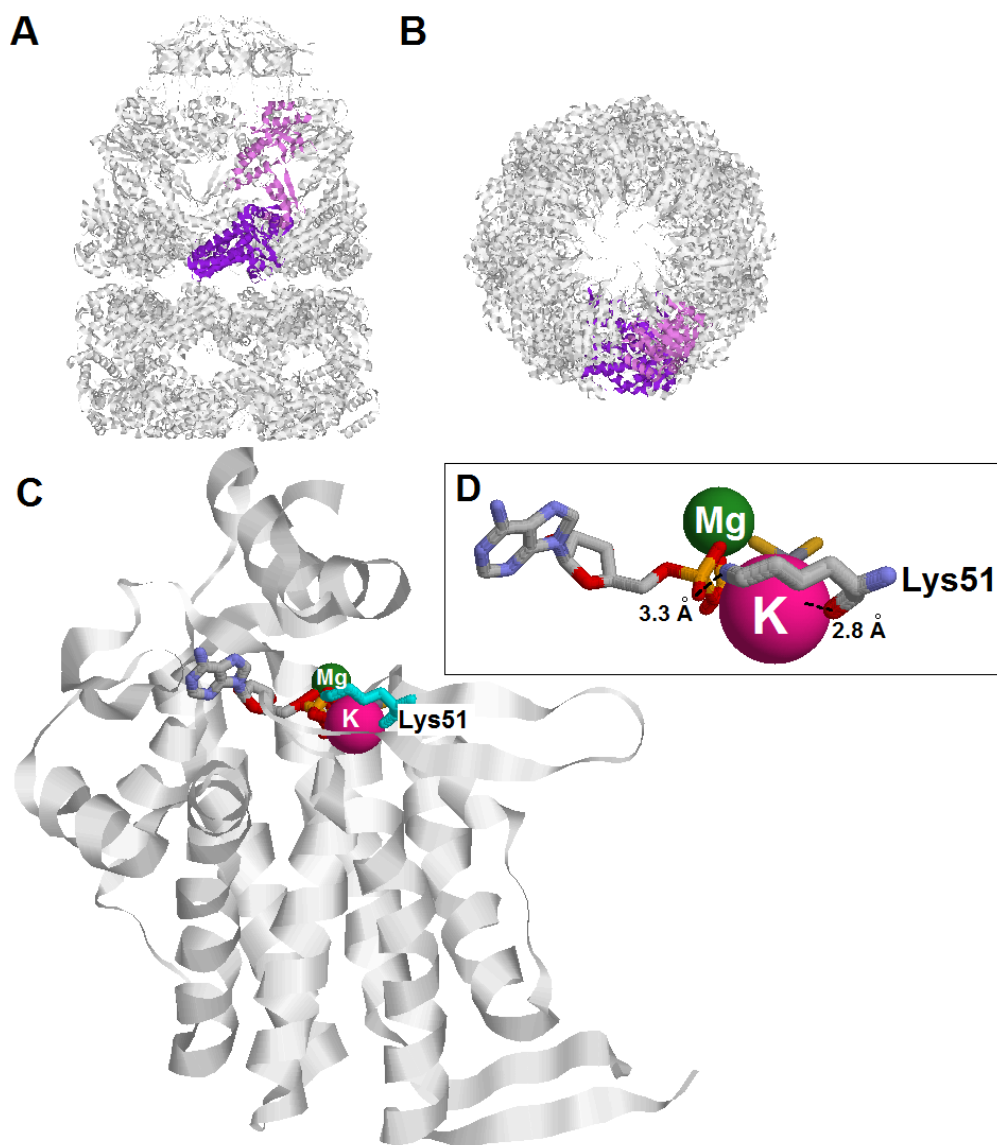
The violet arrow shows the P-loop lysine residue whereas the two cyan arrows show the two lysine residues in RecA sequences which are inserted into the ATPase binding site of the next monomer in complex. Red arrows mark two acidic residues involved in potassium binding in *M. voltae* (the numbers correspond to the *M. voltae* sequence).

### 3.2.3. Example #3: chaperonine GroEL

Chaperonins are large, oligomeric proteins that act as containers upon the folding of other proteins (Ranson *et al.*, 1998). Together with its co-protein GroES, GroEL mediates protein folding in an ATP-dependent manner (Horovitz *et al.*, 2001). Its orthologs are found also in eukaryota (Hsp60 (Ostermann *et al.*, 1989)) and in archaea (Phipps *et al.*, 1993; Trent *et al.*, 1991). The nucleotide influences the affinity of a chaperonine to its protein substrate: GroEL shows low affinity with ATP bound and high affinity in the absence of ATP (Wang and Boisvert, 2003). The ATPase activity of the *E. coli* GroEL was claimed to show an absolute requirement for  $K^+$  (Viitanen *et al.*, 1990), but later the GroEL complexes from the chloroplasts of plants were shown to hydrolyze ATP in a  $K^+$ -independent manner (Viitanen *et al.*, 1995).

A potassium ion is observed in the nucleotide-binding site of the GroEL crystal structure together with  $Mg^{2+}$  (Chaudhry *et al.*, 2003; Page and Di Cera, 2006; Wang and Boisvert, 2003) (**Figure 3.8**). It is noteworthy that the relative positions of ATP,  $Mg^{2+}$  and  $K^+$  are same as in the structures of Rad51 (see Section 3.2.2 above), potassium-dependent HAS-GTPases (see Section 3.2.1 above) and branched-chain  $\alpha$ -ketoacid dehydrogenase kinase (see Section 3.2.4 below). The potassium ion is coordinated by the oxygen atoms of the backbone carboxyl groups of Lys51, Asp52 and Thr30, as well as by the side chain hydroxyl groups of Thr30 and Thr90. Interestingly, the same Lys51 residue is rotated with its side chain amino group towards the ATP molecule (**Figure 3.8D**).

This Lys51 residue is not conserved in the GroEL sequences. It is frequently changed to Asn residue. We have mapped the residues in corresponding position to the phylogenetic tree for the GroEL/Hsp60 sample (**Figure 3.9**). The Asn residue is conserved among almost all archaea, in the eukaryotic sequences which do not cluster with bacterial sequences (nature of duplicated chaperonins in eukaryotes is discussed elsewhere (Archibald *et al.*, 2000)), except for one branch, and in a number of distinct bacterial taxons. Hence, the phylogenomic analysis is compatible with the evolutionary primacy of an Asn-containing  $K^+$ -dependent version of GroEL.



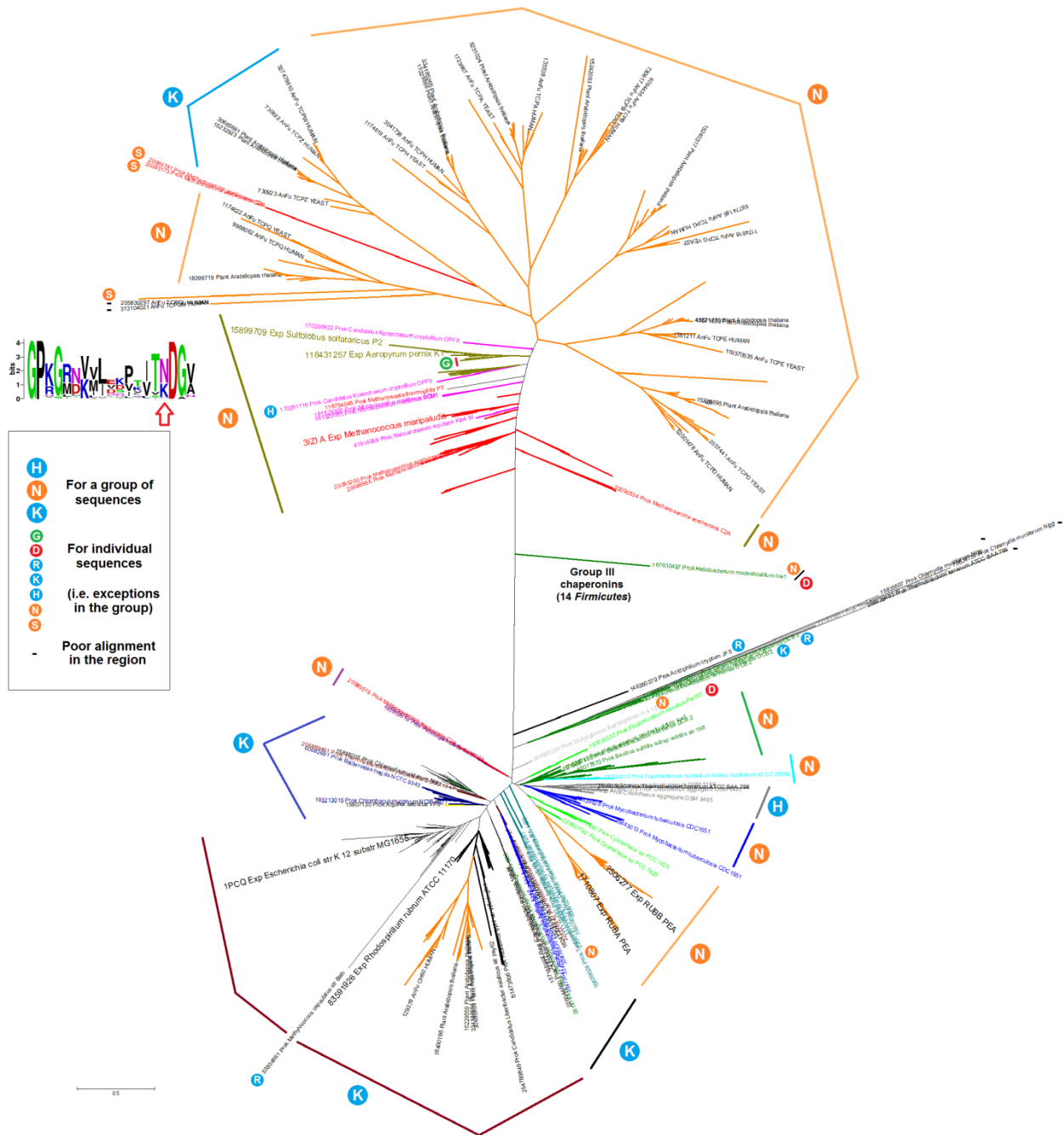
**Figure 3.8. GroEL and GroES complex (PDB 1PCQ (Chaudhry *et al.*, 2003)) from *E. coli*.**

(A) A side view of the oligomeric complex of GroEL with a "cap" of GroES at the top. One subunit (chain A) is colored violet, a part of it with the ATP binding site is colored purple.

(B) Top view of the same complex.

(C) The part of one subunit of GroEL is shown by ribbons, ADP and AlF<sub>3</sub> (that together occupy the space of a bound ATP molecule) are colored by atoms, and metal ions are shown as spheres of Van-der-Waals radius (Mg<sup>2+</sup> in green, K<sup>+</sup> in pink). The proximal Lys51 residue is colored cyan.

(D) The Lys51 residue is forming bonds with both the potassium ion (via backbone carboxyl group) and the  $\alpha$ -phosphate group of ADP.

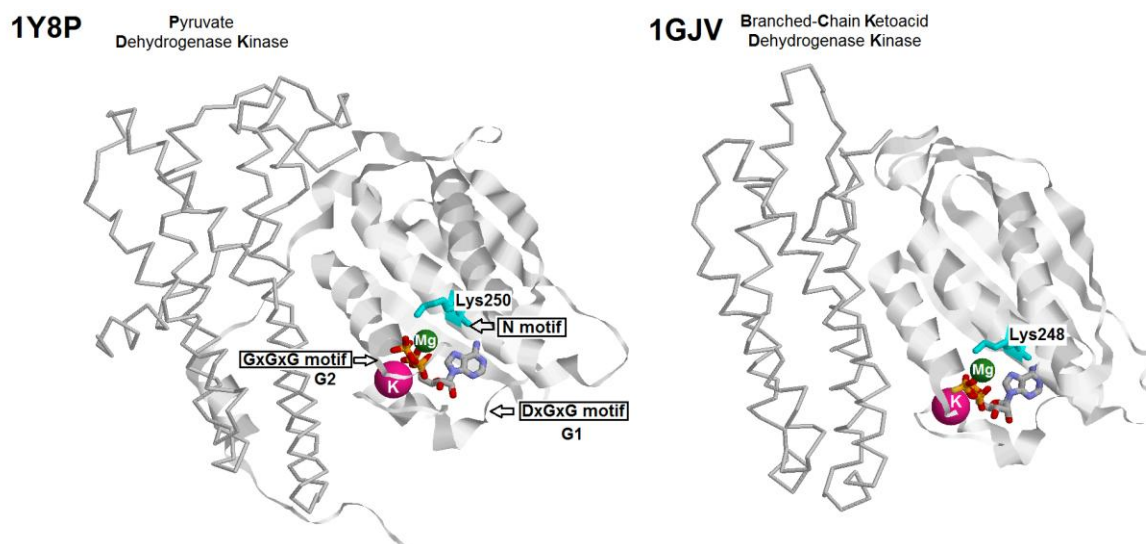


**Figure 3.9. Phylogenetic tree of the GroEL.**

Sample for the alignment construction and phylogenetic tree reconstruction consisted of the members of COG0459 from a set of 179 genomes comprising all bacterial and archaeal phyla as well as proteins from 35 eukaryotic genomes from different kingdoms (*Table S2* and *Table S3*).

### 3.2.4. Example #4: BCK (branched-chain $\alpha$ -ketoacid dehydrogenase kinase)

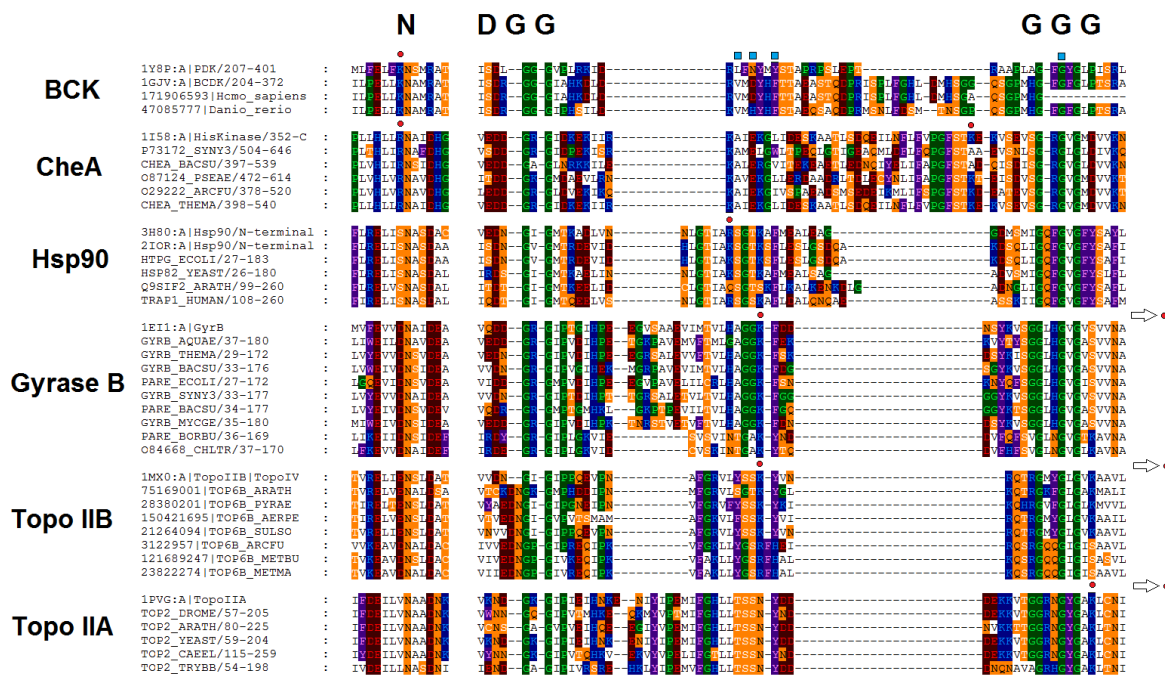
The mitochondrial branched-chain  $\alpha$ -ketoacid dehydrogenase (BCKD) complex catalyzes the oxidative decarboxylation of branched chain  $\alpha$ -ketoacids such as derived from leucine, isoleucine and valine. Similar to the well-known pyruvate dehydrogenase complex, it is composed of a 24-subunit cubic core of dihydrolipoyl transacylase (E2), several copies of branched-chain  $\alpha$ -ketoacid decarboxylase dehydrogenase (E1), homodimeric dihydrolipoamide dehydrogenase (E3) and a number of copies of regulatory proteins: BCKD kinase and BCKD phosphatase (see (Machius *et al.*, 2001) for references). BCKD kinase together with pyruvate dehydrogenase kinases form a distinct family of protein kinases (**Figure 3.2B**) (Harris *et al.*, 1995) inside a superfamily of GHKL (Gyrase, Hsp90, bacterial histidine and mitochondrial serine protein Kinases, DNA mismatch repair protein MutL) ATPases (discovered in (Bergerat *et al.*, 1997), see (Dutta and Inouye, 2000) for a review). In the crystal structures of the BCKD kinase and pyruvate dehydrogenase kinase the potassium ion is bound in the nucleotide-binding site and is located relative to  $Mg^{2+}$  and ATP (or its analogue) as in all the cases considered above (**Figure 3.10**).



**Figure 3.10. Pyruvate dehydrogenase kinase (Kato *et al.*, 2005) (PDB ID 1Y8P) and branched chain  $\alpha$ -ketoacid dehydrogenase kinase (Machius *et al.*, 2001) (PDB ID 1GJV).**

The ATPase domains are shown as ribbons, the N-terminal domains are shown as a backbone. ATP $\gamma$ S (adenosine-5'-[ $\gamma$ -thio]-triphosphate, ATP analog which is hydrolyzed slowly) or ATP are colored by atoms and metal ions are shown as spheres of Van-der-Waals radius ( $Mg^{2+}$  in green,  $K^+$  in pink). Lysine residues proximal to ATP are shown and colored cyan. The positions of the three conservative motifs are shown in the left figure.

The ATPase domain of mitochondrial protein kinases is common for numerous proteins. In the database Pfam (Finn *et al.*, 2010), which contains domains inferred from sequence comparisons, it is represented by a record PF02518 (HATPase\_C) with around 130000 sequences containing this domain. The same ATPase domain is described as a separate fold called "ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase" (number 55873) in the SCOP database of domains inferred from known 3D structures (Murzin *et al.*, 1995). The positions of the three most conserved motifs (N, G1 and G2) in the 3D structure of the pyruvate dehydrogenase kinase are shown in **Figure 3.10**. The residues from these motifs are directly involved in the binding of ATP and metal ions. In particular, Lys250/Lys248, which is located closer than 3.5Å to the  $\gamma$ -phosphate of ATP, is a residue preceding strongly conserved asparagine residue in the N motif, whereas the middle glycine in the G2 motif provides its backbone oxygen atom for the potassium binding.

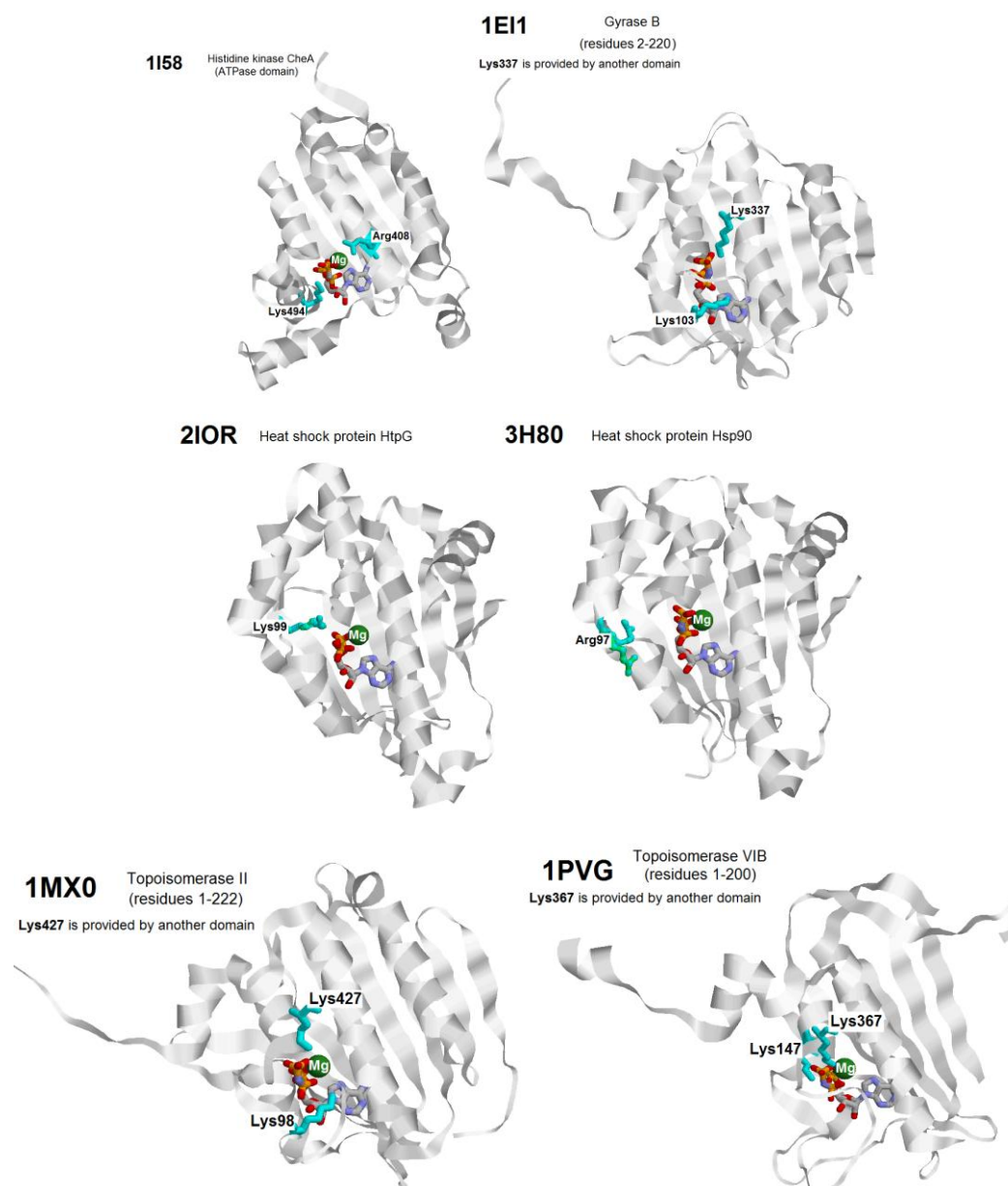


**Figure 3.11. Multiple alignment of ATPase domains of proteins from GHKL superfamily.**

The most conserved regions are shown: the N motif, DxGxG motif and GxGxG motif with the linker between them. Only representative sequences from each subfamily are shown. Alignment is based on the Pfam seed sample and manually corrected. Lysine and arginine residues located in the close proximity to the phosphate groups of ATP (also shown in **Figure 3.10** and **Figure 3.12**) are marked with red dots. The dot is given with an arrow to indicate the cases when a second lysine/arginine residue is located within another domain. Potassium binding site in BCK subfamily is shown with blue rectangles.



The 3D structures of proteins from different subfamilies of the GHKL superfamily show localization of up to two positively charged residues near the phosphates of ATP (**Figure 3.12**). They are conserved within a subfamily but vary between different subfamilies (for multiple alignment see **Figure 3.11**). Thus, there is no evidence that these residues are an ancestral trait of the family. It is likely that they have independently appeared during the evolution of each subfamily and that the ancestral form of the enzyme was  $K^+$ -dependent.



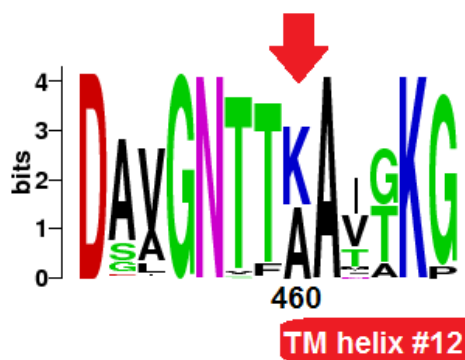
**Figure 3.12.** Lysine and arginine residues in different proteins from GHKL ATPase/kinase superfamily near the phosphate groups of the ATP. Colors used repeat the colors of **Figure 3.10**.

### 3.2.5. Example #5: membrane pyrophosphatases

The enzymes from this family couple hydrolysis of pyrophosphate ( $\text{PP}_i$ ) with an active transport of cations ( $\text{H}^+$  or  $\text{Na}^+$ ) across the membrane (Maeshima, 2000) and show no significant sequence similarity to other known protein families. The Pfam record PF03030 covers this family and shows that its representatives are widely distributed among all domains of life. The proteins are highly hydrophobic: the *Thermatoga maritima* enzyme contains 16 transmembrane helices connected by loops (Kellosalo *et al.*, 2012). Until recently no 3D structure was known for these proteins, but in 2012 two crystal structures of membrane pyrophosphatases were finally solved. These are proteins from *Thermatoga maritima* (PDB ID 4AV6 (Kellosalo *et al.*, 2012)) and *Vigna radiata* (PDB ID 4A01 (Lin *et al.*, 2012)).

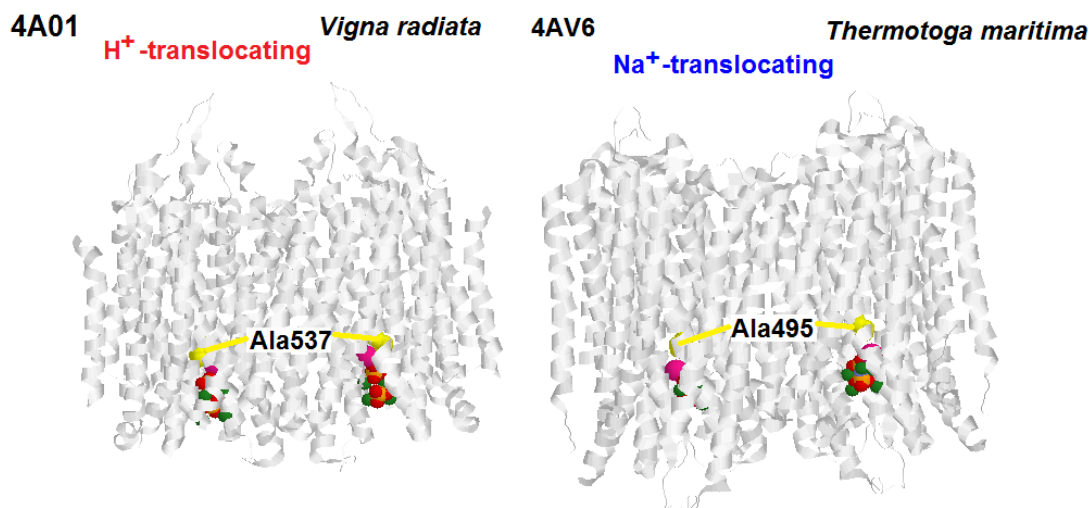
We have constructed a phylogenetic tree of membrane pyrophosphatases (**Figure 3.16**). It is in a general agreement with the previously constructed tree (Luoto *et al.*, 2011). Membrane pyrophosphatases are divided into two major clades on the phylogenetic tree, one bearing all known  $\text{K}^+$ -dependent enzymes with another containing  $\text{K}^+$ -independent enzymes. A single amino acid (K460 in *Carboxydotherrmus hydrogenoformans*) determines the potassium dependence: a substitution which introduces a lysine residue into this position decreases the activity by about 50%, thereby the mutant protein shows no requirement for  $\text{K}^+$  either upon pyrophosphate hydrolysis, or upon cation transport (Belogurov and Lahti, 2002). Indeed, this position has a conserved lysine in potassium-independent enzymes (it could be also arginine, as in *Methylococcus capsulatus*, GI 53804690) and a non-polar alanine residue in the proteins which require potassium (see **Figure 3.14** for the part of the sequence logo and **Figure 3.15** for the two known 3D structures of membrane pyrophosphatases). An interesting feature of our tree, which is absent from the previously published phylogenetic tree of pyrophosphatases (Luoto *et al.*, 2011), is the small clade that is separated by a long branch from the rest of the tree (shown on the bottom of **Figure 3.16**). These proteins can be unambiguously aligned with other membrane pyrophosphatases along full length, but they likely represent a separate family. No member of this clade has been experimentally characterized.

As depicted in **Figure 3.16**, the type of the translocated cation appears to correlate with the  $K^+$ -dependence of the enzyme. Among the  $K^+$ -dependent pyrophosphatases, the majority translocates sodium. Sodium translocation was proposed to be the ancestral function of membrane pyrophosphatases (Luoto *et al.*, 2011), which is in agreement with the previously suggested evolutionary primacy of sodium bioenergetics (Mulkidjanian *et al.*, 2008b). Thus,  $K^+$ -dependence seems to be an ancient trait also in the family of membrane pyrophosphatases.



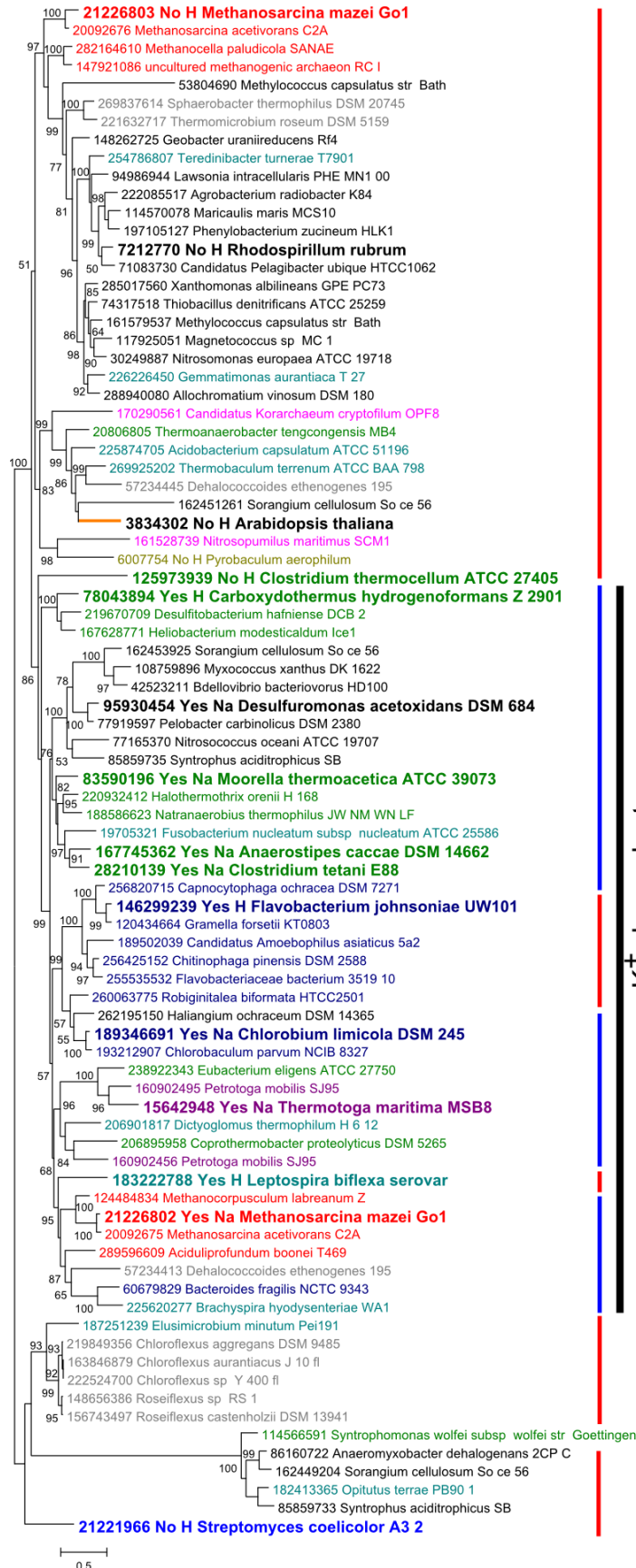
**Figure 3.14.** Part of the sequence logo derived from the seed of the Pfam family PF03030.

Arrow marks the position of the lysine residue which determines potassium independence of the enzymes (position 460 in *C.hydrogenoformans*).



**Figure 3.15.** Crystal structures of two membrane pyrophosphatases from *Vigna radiata* (PDB 4A01 (Lin *et al.*, 2012)) and *Thermotoga maritima* (4AV6 (Kellosalo *et al.*, 2012)).

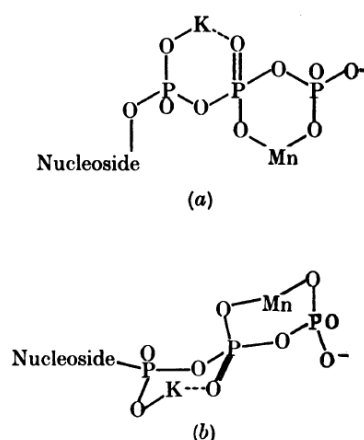
Protein helices are shown in gray. The alanine residue which is replaced by lysine in the  $K^+$ -independent enzymes is shown in yellow. The pyrophosphate molecule and  $Mg^{2+}$  ions are depicted in standard colors as in **Figure 3.10**. The alanine residues marked with a red arrow on the logo in **Figure 3.15** are colored yellow.



**Figure 3.16.** A phylogenetic tree of membrane pyrophosphatases (COG3808) based on the set of 179 genomes chosen from all bacterial and archaeal phyla (Table S2 and Table S3). Experimentally studied sequences, as described in (Luoto *et al.*, 2011), were added to the tree. The known features, namely the K<sup>+</sup>-dependence ("Yes" or "No") and the type of translocated cation ("H" for protons and "Na" for sodium) are given after the numeric ID in the name of these sequences. The predicted K<sup>+</sup>-dependence of one clade is marked with bold black bar. The type of the translocated cation, as could be predicted for the clades on the tree based on the known coupling ion specificity, is shown by thin bars (red for H<sup>+</sup> and blue for Na<sup>+</sup>).

### 3.3. Discussion: Activation by $K^+$ ions could be evolutionarily older than the involvement of lysine or arginine "fingers"

Back in 1960, Lowenstein has shown that  $K^+$  and  $NH_4^+$ , but not  $Na^+$  and  $Li^+$  stimulated the non-enzymatic-transphosphorylation reaction ( $ATP + P_i \rightarrow ADP + PP_i$ ) in the presence of divalent cations (Lowenstein, 1960). Since  $NH_4^+$  – that cannot form coordinating bonds – was as efficient as  $K^+$ , Lowenstein has insightfully suggested electrostatic nature for the catalytic effect, namely that a bulky monovalent cation could compensate electrostatically the negative charges of two neighboring oxygen atoms (see the scheme below in **Figure 3.17** and compare with the aforementioned protein structures). To accelerate hydrolysis of a phosphoester bond, it is needed to electrostatically compensate totally four negative charges. Three are the charges of the phosphate groups and one more charge emerges as the charge of the leaving group that is formed upon the hydrolysis of the phosphoester bond (Cleland and Hengge, 2006). Accordingly, the task of electrostatic compensation could be performed either by two divalent cations (as indeed observed in many hydrolase families, see Section 1.2.2 and (Doublie *et al.*, 1998; Sträter *et al.*, 1996)) or by a divalent cation and two positive charges, as seen in the surveyed structures where usually two positive charges are involved in addition to a  $Mg^{2+}$  ion. Therefore, one more positive charge should be added to the old scheme of Lowenstein that is shown in **Figure 3.17**.



**Figure 3.17.** Tentative positions of  $K^+$  and  $Mg^{2+}$  around an ATP molecule according to Lowenstein (Lowenstein, 1960).

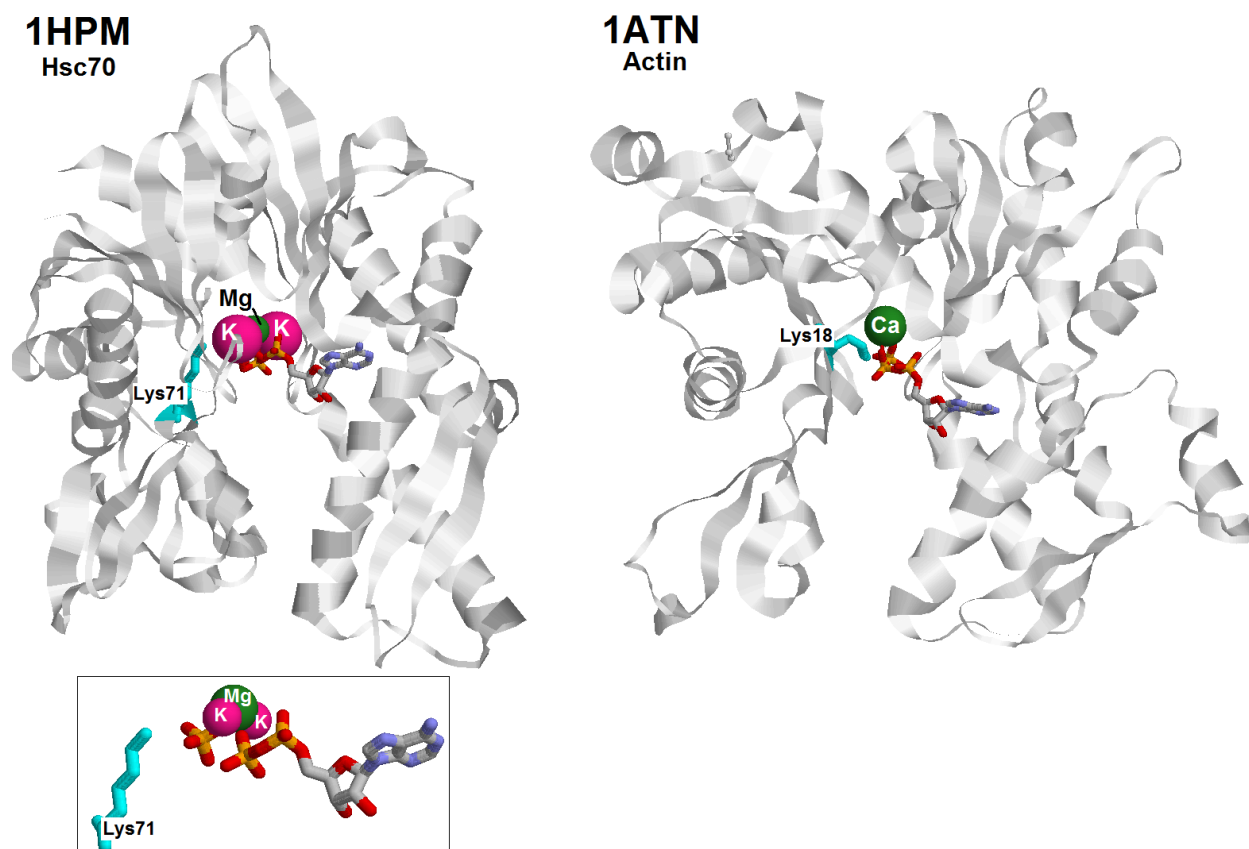
Since the reactions of hydrolysis of nucleoside triphosphates should have been used already by the very first replicating entities, perhaps, before the emergence of peptides, it seems plausible that initially the reaction could be assisted by the  $K^+$  ions of the medium. First proteins could increase their efficiency by providing binding loops for  $K^+$  ions of the medium. It is noteworthy that diverse evidence indicates that amino acids were recruited for the coded translation in a certain order (Jordan *et al.*, 2005; Trifonov, 2000): amino acids with non-polar side chains (Gly, Ala, Val, Leu, Pro), neutrally charged side chains (S, T) and negatively charged side chains (Asp, Glu) are believed to be the first. In other words, the first proteins could lack positively charged amino acid residues. These first proteins could, however, bind  $K^+$  ions by their Asp and Glu residues, so that the aspartate-using  $K^+$ -binding sites might be remnants from the times when the set of coded amino acids was smaller than nowadays. After the recruitment of Lys and Arg by coded translation mechanism, the enzymes got the opportunity to control the reaction of hydrolysis by moving the positively charged residues (their tips, are, in fact,  $NH_4^+$  groups) around the nucleotide. Accordingly, the position of two catalytic  $K^+$  ions could become occupied first by Lys of the P-loop and then by mobile arginine fingers. The possible important advantage of mobile lysine or arginine "fingers" is an increased control over the catalytic reaction. The "finger(s)" can be inserted only during particular step of catalytic cycle. In such enzymes, as RecA or  $\alpha/\beta$ -subunits of ATP synthase, the positively charged residue becomes available only upon proper organization of the protein filament on the nucleic acid or upon the appropriate rotation step of the central stalk within the catalytic hexamer, respectively. In such molecular switches as Ras, an occasional hydrolysis of GTP by, for example, a random binding of a  $K^+$  ion would erroneously activate a cell signaling cascades, which is prevented by providing an arginine finger at the proper time from a specifically bound GAP protein. In gyrases and topoisomerases of the GHKL superfamily, the positively charged residue is provided by another domain (**Figure 3.12**). Such arrangement could allow better domain interconnection and thus regulation of enzyme function.

In the case of the P-loop GTPases family the principle  $K^+$ -binding ligand, Asp or Asn, is present in the most ancestral groups of proteins named in Section 1.3.1 (Leipe *et al.*, 2002), including translational factors. For the RecA/RadA family the evidences are less direct, as pair of lysine residues is observed in all bacterial sequences while  $K^+$ -binding site is

attributed to archaeal and bacterial sequences. However, analysis of paralogous enzymes, namely RadB in archaea and subunits of the rotary membrane ATP synthase, show different possible arrangements of the positively charged "finger", while the  $K^+$ -binding site remains conserved. In the GroEL family the lysine "finger" residue is present in only some groups of bacteria, while other bacterial, archaeal and eukaryotic sequences retain an Asn residue in this position. In the subfamilies of the GHKL superfamily we observed various arrangements of the positively charged residues along the nucleotide, but these residues, while being well-conserved inside the subfamily, were located in different places of the proteins. Finally, for membrane pyrophosphatases, the  $K^+$ -dependence is well-correlated with utilizing  $Na^+$  as a coupling ion, which indicates the  $K^+$ -dependence as an ancient feature, if  $Na^+$ -based bioenergetics preceded  $H^+$ -based (Mulkiđjanian *et al.*, 2008a; Mulkiđjanian *et al.*, 2008b). In Sections 7.1 and 7.2 below we discuss arguments supporting this point of view.

Another example of a possible switch from binding a  $K^+$  ion to hosting a positively charged residue is apparent in the Hsc70/actin superfamily of ATPases. The molecular chaperone Hsc70 is a widespread protein which acts at the levels of folding, assembly and disassembly of protein complexes (see (Hartl *et al.*, 2011) for a recent review on the topic).  $K^+$  ion was observed to bind specifically in the ATPase binding site of the chaperon Hsc70 (Wilbanks and McKay, 1995) and was shown to be required for the optimal ATPase activity (O'Brien and McKay, 1995). Potassium was proposed to stabilize the transition state of the reaction, thus acting in catalysis.

Despite the absence of detectable sequence similarity, the 3D structures of Hsc70 and actin look surprisingly similar ((Kabsch *et al.*, 1990) and **Figure 3.18**). No potassium ions are detected in actins but two lysine residues are located near the active center. One of them with its positively charged termini occupies the same place as one of the  $K^+$  ions in Hsc70. The actin family is in general restricted to eukaryotes (Pfam domain PF00022 corresponding to actins is currently observed in more than 15000 eukaryotic sequences, with only around 20 homologous sequences found in bacteria and archaea), whereas Hsc70 proteins are widely spread in bacteria, archaea and eukaryotes. Thus actins with their Lys (or Arg) residue are likely to be evolutionary younger than to the  $K^+$ -activated Hsc70 chaperons.



**Figure 3.18.** Overall structures of Hsc70 (PDB ID 1HPM) and actin (PDB ID 1ATN).

Color code is the same as in *Figure 3.10*. The ADP molecules and the phosphate groups (or ATP molecules) are shown as wireframes, as well as the lysine residues in the close proximity of the phosphates.

In this chapter we provided evidence for the evolutionary primacy of  $K^+$  activation in widespread protein families belonging to the P-loop GTPases from TRAFAC superfamily, RecA/RadA family, GroEL/Hsp60 family, GHKL ATPase/kinase superfamily and membrane pyrophosphatases. It is a formidable future challenge, which is beyond our current capacities, to trace the replacement of metal cations by basic residues along the phylogenetic trees in all families of different ATPases and GTPases.



## 4. N-ATPases: distinct family of rotary membrane ATPases

Previous phylogenomic analyses of the rotary membrane ATP synthases (see Section 1.3.2.4) have shown that the five Na<sup>+</sup> binding residues in the *c*-subunits are the same in all experimentally studied sodium-pumping ATP synthases (Mulkidjanian *et al.*, 2008a; Mulkidjanian *et al.*, 2008b). Thereby it is possible to predict the ion specificity of the rotary ATP synthase from the sequence of its *c*-subunits: the presence of the full set of Na<sup>+</sup>-binding ligands is a reliable indication of its capability to translocate sodium (Mulkidjanian *et al.*, 2008b).

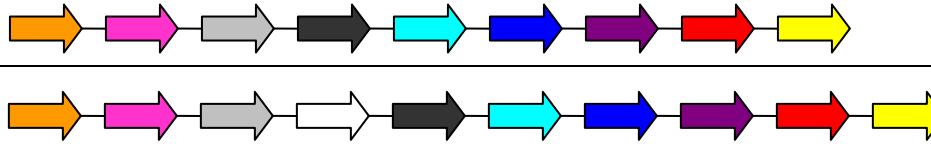
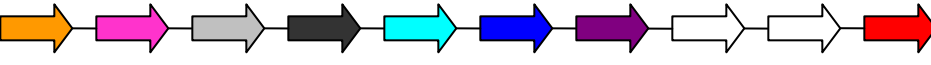
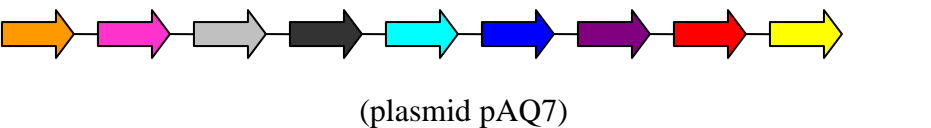
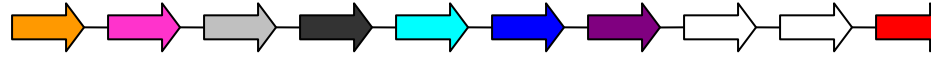
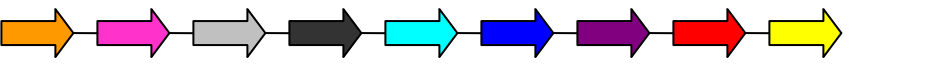
Based on this consideration, we have addressed the issue of the sodium dependence of energy-coupled reactions in cyanobacteria, as tackled in several reports (Pogoryelov *et al.*, 2003; Skulachev, 1988; Willey *et al.*, 1987). A search for Na<sup>+</sup>-translocating ATPases within cyanobacterial genomes has revealed several apparently Na<sup>+</sup>-dependent cyanobacterial ATPases, but always as second ATPases, in addition to the H<sup>+</sup>-translocating ones. These "second" proteins have a number of common properties in addition to their apparent sodium-dependence and comprise a separate subfamily of rotary membrane ATPases with members in other prokaryotic lineages. Since these "second" ATPases, besides forming a new subfamily, are always encoded next to typical rotary ATPases and are predominantly Na<sup>+</sup>-dependent, we suggested to call them N-ATPases; the here presented results on N-ATPases were published in (Dibrova *et al.*, 2010).

Initially, in the RefSeq database we identified 5 cyanobacterial *c*-subunits with a complete set of Na<sup>+</sup>-binding ligands. All these proteins belong to marine cyanobacteria (**Figure 4.2**). The operon encoding the Na<sup>+</sup>-binding subunit was present in each case in the genome along with another ATPase operon, encoding a H<sup>+</sup>-translocating F-type ATPase. **Figure 4.1** shows an alignment of the Na<sup>+</sup>-binding and the H<sup>+</sup>-binding *c* subunits from the same cyanobacterium. This alignment shows an important substitution of the conserved Gln residue (Gln-28 of ATPL\_SYNP2, marked by an asterisk), which serves as a Na<sup>+</sup> ligand in the Na<sup>+</sup>-binding *c* subunits of *Ilyobacter tartaricus* and other bacteria (Meier *et al.*, 2005; Mulkidjanian *et al.*, 2008b). In cyanobacterial Na<sup>+</sup>-binding *c* subunits, this position is occupied by a Glu residue,

which could potentially provide up to two coordination bonds for the bound  $\text{Na}^+$  ion (see below). This type of motif cannot be considered common: among the 4084 *c*-subunit sequences currently (March 2010) listed in the Pfam family ATP-synt\_C (PF00137), only 227 (~5.5%) contain a Glu residue in that position and only ~1% of them combine it with a typical EST motif in the  $\text{Na}^+$ -binding site (see **Figure 4.1** for shortened alignment and **Figure 4.2** for the whole alignment).

The N-ATPase operons in cyanobacteria had similar gene order which distinguished them from the  $\text{H}^+$ -dependent F-ATPase operons. A single gene insertion was observed in *Acaryochloris marina* and a two-gene insertion in the two strains of *Cyanothece sp.* (**Table 4.1**).

**Table 4.1.** Typical operon structure of N-ATPases as observed in cyanobacterial operons (from (Dibrova *et al.*, 2010)).

Organism	N-ATPase operon
	$\beta$ $\epsilon$ Q   R <i>a</i> <i>c</i> <i>b</i> $\alpha$ $\gamma$
<i>Acaryochloris marina</i> MBIC11017	
<i>Cyanothece sp.</i> ATCC 51142	
<i>Synechococcus sp.</i> PCC 7002	
<i>Cyanothece sp.</i> CCY0110	
<i>Nodularia spumigena</i> CCY9414	

B1XRK4\_SYNP2 20 IGSIGPALGEGMAVARALGAIAQQPDKANMITRRLFVGLAMV**ESTAIY**CLV**V**72  
 \* \* \* \* \*  
 ATPL\_SYNP2 18 LAAIGPGIGQGNAAGSAAEGIARQPEAEGKIRG**TLLLSLAFMEALTY**GLV**V**70

Figure 4.1. Amino acid sequences (partial) of the Na<sup>+</sup>-binding c-subunit (above) and the H<sup>+</sup>-binding c-subunit (below) from the genome of *Synechococcus* sp. PCC 7002.

Positions of the membrane helices are shown in grey, Na<sup>+</sup> ligands are marked with asterisks.

**Cyanobacteria**  
 A82NS2\_ACAM1 1 MDN**ALV**GMASIVISGLTTIAIGSIG**PALGE**GRALSQALSALAQQPDEANTITR**LVFVGMALV****ESTAIY**CFVITLILIFAN**PFW**TYVVEQASTNGG 95  
 B4WTA0\_9SYNE 1 MDNVV**LIG**IASIVMGLTTIAVGSIA**PALGE**GRALAQAL**TALA**QQPDEANTITR**LVFVGMALV****ESTAIY**CFVITLILIFAN**PFW**AYVTATAGG--- 92  
 B1WXB3\_CYAA5 1 MDN**LGL**TGVSIITAGLTTIAIGSV**PAIGE**GLALANAL**FALA**QQPDKANTITR**LVFVGLALV****ESTAIY**CFVSMILIFAN**PFW**NYFVQAGG--- 92  
 B1XRK4\_SYNP2 1 MSD**LAL**IGSVAMV**TAGIT**IAIGSIG**PALGE**MAVARALGAIAQQPDKANMITR**LVFVGLAMV****ESTAIY**CLVSMILL**VNPFW**NYFLNQGG--- 91  
 A0ZFR6\_NODSP 3 IDN**LGL**GMVSIIMAGLTTIAIGSIG**PAIAGE**WAVARALGAMAQQPDDQANTITR**LVFVGLAIIE****ESTAIY**CFVSMILIFAN**PFW**YDFFTSS----- 91

**Archaea**  
 Q8TN57\_METAC 4 DTY**TTII**AVASIASITGLTIGIGV**LGPAIGE**GRAVATAL**SSLA**QQPDASATITR**LVFVGLAMIE****ESLSIY**CFVSMILIFAN**PFW**NTATV----- 91  
 Q467F4\_METBF 4 DTY**TTII**AVASIASITAGITIGIGV**LGPAIGE**GRAVATAL**SSLA**QQPDASATITR**LVFVGLAMIE****ESLSIY**CFVSMILIFAN**PFW**NRALT----- 91

**Aquificae**  
 C0QS18\_PERMH 1 MDS**LTI**IAAVSIFTAGLTTIAIGG**MMPSRGE**EALTKAESVA**RQPE**QANNI**IRLFF**IGAAV**VE**SAI**Y**SLVIALI**ILFAN**PF**II**HLFTK----- 88

**Chlorobi**  
 Q3B401\_PELLD 1 MDI**LTI**VAAVSIFTAGITIAVGSIG**PALGE**GRAVASALE**EALA**QQPDASSITR**LVFVGLAMIE****ESVAIY**CFVSMILIFAN**PFW**TMSLQAGG-- 93  
 B3QNH4\_CHLP8 1 MDT**LTI**IAVSIATAGLSIGIGV**LGPAI**GGRAVSSAL**TALA**QQPDAASTITR**LVFVGLAMIE****ESIAIY**CFVISI**ILIFAN**PF**W**NHAI**TAQ**VGGH-- 93

**Planctomycetes**  
 Q7UH07\_RHOBA 1 MDS**TTI**AVASIIIMAGLTTIAIGSIG**PAFAE**GRAVAQALNSIAQQPDSNTITR**LVFVGLAMIE****ESTAIY**CFVSMILL**FAN**PF**W**NQLTQ----- 88  
 A6C217\_9PLAN 1 MDSNTVIAAVSIFTAGITIAIGSIG**PALGE**GRALAQALS**IA**QQPDEASTITR**LVFVGLAMV****ESTAIY**CFVSMILIFAN**PFW**NHFMQAATR--- 92  
 Q1Q5S1\_9BACT 1 MDN**VGL**IGMVSIIVAGFTIAVGSIG**PALGE**RAAAQALS**IA**QQPDEANTITR**LVFVSMAMIE****ESTAIY**CFVAMIV**IFAN**PF**W**NYVITKAGGQ-- 93

**Proteobacteria**  
 B7RJ30\_9RHOB 1 MTD**LAI**IAAISILTAGLTVCFGAIG**PALGE**GR**TASA**AL**TAIA**QQPDAASSIS**RTL**FSVLAMIE**ESTAIY**CFVAMILL**FAN**PF**W**NAAIAAAKATGS 95  
 D0CPG0\_9RHOB 1 MTD**LAI**IAAVSIFTAGLTVSFGAIG**PALGE**GRAAAQALASIAQQPDAAP**TL**SR**TL**FSVLAMIE**ESTAIY**CFVAMILIFAN**PFW**DTALELVRTSGQ 95  
 C6BWK9\_DESAD 1 MET**LGW**IAFGSIIAAGLCMGIG**PAIGE**GMALS**RALSSIA**QQPDE**TNT**IVK**FL**FVGMAM**VE****ESTAIY**CFVLAMILL**FAN**PF**W**SYFLEKAGG--- 92  
 A3IYM7\_9CHRO 1 MDN**LGL**TGVSIITAGLTTIAIGSV**PAIGE**GLALANAL**FALA**QQPDKANTITR**LVFVGLALV****ESTAIY**CFVSMILIFAN**PFW**NYFVQAGG--- 92  
 B4S799\_PROA2 3 **TTII**AVAVASIVT**AGLTT**IAIGCIG**PALGE**GRAVSSAL**TSLA**QQPDA**AA**ITR**TL**FI**GLAMV****ESVAIY**CFVISMILIFAN**PFW**NQVIVQAGG-- 95  
 A8LN44\_DINSH 1 MTD**LAI**IAAVSIFTAGLTTIAIG**PAIGE**GRAASTAL**SAIA**QQPDA**AST**LSR**TL**FSVLAMIE**ESTAIY**CFVAMILIFAN**PFW**EAALEAATV**GAG** 95  
 Q0AEJ1\_NITEC 1 MTD**LAI**IAAISIFTAGLTTIAIG**SLGPAIGE**GRAAAAA**IAAIA**QQPDA**APT**LSR**TL**FSVLAMIE**ESTAIY**CFVAMILIFAN**PFW**NAMQAVAGG--- 92  
 Q62EB2\_BURMA 1 ---M**N**NL**E**EVV**SI**AAAA**AL**AVS**FGAIG****PALAE**GRAVGAAM**DAIAR**QPDAS**GT**VS**RTL**FSVLAMIE**ETMAIY**CLVVAL**LLLFAN**PF**VK**----- 82  
 Q63I**W8**\_BURPS 1 ---M**N**NL**E**EVV**SI**AAAA**AL**AVS**FGAIG****PALAE**GRAVGAAM**DAIAR**QPDAS**GT**VS**RTL**FSVLAMIE**ETMAIY**CLVVAL**LLLFAN**PF**VK**----- 82  
 Q3A078\_PELCD 1 MDY**LAW**AVASISL**SAG**LCIGIG**PAI**GGRA**LQA**L**AALA**QQPDEANTITR**LVFVGLAMV****ESTAIY**CFVSMILIFAN**PFW**RF**FL**VRAGEGG-- 94  
 Q1K2D6\_DESAC 1 MEA**Q**TWIAMISIFTAGITIAIGSIG**SALGE**GRAVASAL**TSLA**QQPDSAS**TITR**LVFVGLAMV**ESTAIY**CFVSMIV**L**FAN**PFW**NHFLEAAGG--- 92  
 C7LQ36\_DESBD 1 MDS**MTI**IAVASII**IAGIT**TFG**TMG****PALAE**GGKAVATAL**TSLA**QQPDSAS**TITR**LVFVGLAMIE**ESTAIY**CFVSMILIFAN**PFW**NYAIAQ**MAGK**-- 93  
 B3QG15\_RHOPT 1 ---M**N**NL**E**EVV**SI**AAAA**AL**AVS**FGAIG****PALAE**GRAVGAAM**DAIAR**QPEA**ANT**IS**RTL**FSVLAMIE**ETMAIY**CLVVAL**LLLFAN**PF**LLK**----- 81  
 Q2Y8G7\_NITMU 1 MDS**LTI**IAAISILTAGLTTISIGV**LGPAI**GGRAVATAL**TSLA**QQP**DVAG**TIAR**TL**FI**GLAIIE****ESLAIY**CFVSMILIFAN**PFW**DH**VI**AHG**VGK**-- 93  
 Q2LY41\_SYNAS 1 MDS**MTI**IAVSIITAGLTTIAIGV**LGPSLGE**GN**AVK**AL**TAIA**QQPDE**RNS**ITR**TL**FI**GLAMIE****ESIAIY**CFVISMILIFAN**PFW**SH**AI**ARAGG--- 92  
 A1SS60\_PSYIN 1 MDS**MTY**IAMVSI**IAGL**TFG**TMG****PALAE**GRAVGAAM**ASLA**QQPDSAS**TITR**LVFVGLAMIE**ESTAIY**CFVSMI**ILFAN**PF**W**DFAV**TQ**AGK-- 93  
 A1VPR0\_POLNA 1 MDS**MTI**IAVASIV**IAGL**TFG**CMG****PAFAE**GRAVATAL**TALA**QQPDSAS**TITR**LVFVGLAMIE**ESTAIY**CFVSMILIFAN**PFW**NFAIAQ**AGK**-- 93  
 Q21ZAI\_RHOFD 1 MDS**LTI**IAVASIV**IAGL**TFG**CMG****PALAE**GRAVATAL**TALS**QQPDSAS**TITR**LVFVGLAMIE**ESTAIY**CFVSMILIFAN**PFW**NH**VI**AQ**TAGK**-- 93  
 Q15SF3\_PSEA6 1 MDS**LTI**IAICSIITAGLTTIGIGV**LGPSLAE**GS**AVAS**AL**KALA**QQPDSAS**TITR**LVFVGLAMIE**ESTAIY**CFVSMILL**FSN**PF**W**NYFISQ**NGG**-- 92  
 A6VWQ4\_MARMS 1 MDS**MTI**IAVSIITAGLTTIAIGV**LGPSLGE**GGKAVATAL**TSLA**QQPDSAS**TITR**LVFVGLAMIE**ESTAIY**CFVSMILL**FAN**PF**W**NYGIAQ**AGG**-- 93  
 Q07YM2\_SHEFN 1 MDS**ITI**IAMVSIITAGFTITIGV**IGPSLGE**GGKAVATAL**SSLA**QQPDSAS**TITR**LVFVGLAMIE**ESTAIY**CFVSMI**LLFAN**PF**W**NYVIDQ**AGG**-- 93  
 A3JAV4\_9ALTE 1 MTD**LAI**IAAISIFTAGFTIAIGCIG**PSLAE**GRAAAAA**IAAIA**QQPDA**APT**LSR**TL**FSVLAMIE**ESTAIY**CFVAMILIFAN**PFW**DA**AV**Q**AAT**VTAG 95  
 C1DEJ6\_AZOV**D** 1 ---M**N**PL**E**IVS**IL**GAALAVS**FGALG****PALAE**GRAVGAAM**DAIAR**QPEA**AG**TL**SR**TL**FSVLAMIE****ETMAIY**CLVVAL**LLLFAN**PF**W**H----- 82

<-----MEMBRANE-----> <-----CYTOPLASM-----> <-----MEMBRANE----->  
 ATPL\_ILYTA 3 ML**F**AK**TV**LAASAVGAGTAM**IAGIG**PGV**GQGYAAGKAVESV**AR**QPEAK**GD**II**ST**MTVLGQAVAE**ST**GIYS**LVIAL**ILLYAN**PF**V**GLL**G**----- 89  
 ATPL\_PROMO 3 MV**L**AK**TV**LAASAVGAGAM**IAGIG**PGV**GQGYAAGKAVESV**AR**QPEAK**GD**II**ST**MTVLGQAVAE**ST**GIYS**LVIAL**ILLYAN**PF**V**GLL**G**----- 89

\* \* \* \* \*  
 A1T0Z4\_PSYIN 1 ---ME**II**IS**TA**IAVALMIG**LAA**FGTAVGF**AILGGKFL**ES**ASAR**Q**PE**MG**PAL**Q**TKM**FIVAG**LDAIS**MI**AVG**VAL**FFV**FAN**PF**L**SQ**LA----- 85  
 A1VIU7\_POLNA 1 ---MEN**IL**GLVALACGL**IV**GLG**AI**GS**IG**IALMGG**KFL**ES**ASAR**Q**PE**LM**N**LQ**TKM**FILAG**LIDAA**FLIG**V**AI**ALL**FAN**PF**VL**R**----- 82  
 Q223D1\_RHOFD 1 ---MEN**IL**GLVALACGL**IV**GLG**AI**GS**IG**IALMGG**KFL**ES**ASAR**Q**PE**LM**N**LQ**TKM**FILAG**LIDAA**FLIG**V**AI**ALL**FAN**PF**VL**R**----- 81  
 Q15MT9\_PSEA6 1 ---ME**IL**GNLYTAVG**IL**IGL**CA**LG**PAI**GF**LLGGKFL**ES**ASAR**Q**PE**LA**P**LQ**V**K**M**FIVAG**LIDAI**AMIG**V**AI**ALL**FL**VE**SA**K**FA**V**----- 82  
 A6W3T3\_MARMS 1 ---ME**TV**VLG**TA**IAV**ALL**IGL**GA**LG**TAI**GF**LLGGKFL**ES**ASAR**Q**PE**MA**P**LQ**V**K**M**FIVAG**LIDAV**TMIG**V**GI**AL**FF**T**FAN**PF**VAL**VAG**----- 85  
 Q07VT9\_SHEFN 1 ---ME**TV**LG**MTA**IAV**ALL**IGM**GA**LG**TAI**GF**LLGGKFL**ES**ASAR**Q**PE**MA**P**LQ**V**K**M**FIVAG**LIDAV**TMIG**V**GI**AL**FF**T**FAN**PF**L**AM**L----- 83  
 A3JGY5\_9ALTE 1 -----M**TA**IAV**ALL**IGL**GA**LG**TAI**GF**LLGGKFL**ES**ASAR**Q**PE**MI**P**LQ**V**K**M**FIVAG**LIDAV**TMIG**V**GI**AL**FF**T**FAN**PF**V**G**LAG----- 79  
 C1DND8\_AZOV**D** 1 -----ME**LV**II**AA**SIMIG**L**GA**L**STIG**F**AL**GGKFL**ES**TA**R**Q**PE**L**AP**L**Q**TK**FL**M**AG**L**L**D**AV**PM**IG**V**GI**AM**Y**LI**F**VV**AP**S**AA----- 78

Figure 4.2. Full-length sequence alignment of N-ATPase c-subunits with c-subunits of the Na<sup>+</sup>-translocating F-ATPases from *Ilyobacter tartaricus* (ATPL\_ILYTA) and *Propionigenium modestum* (ATPL\_PROMO) and with H<sup>+</sup>-translocating F-ATPases from several N-ATPase-encoding organisms (from (Dibrova et al., 2010)). Amino acid residues that form the Na<sup>+</sup>-binding site of N-ATPases, Na<sup>+</sup>-translocating F-ATPases from *Ilyobacter tartaricus* and *Propionigenium modestum*, as well as the H<sup>+</sup>-binding site of other F-type ATPases are colored and shown in bold typeface and marked with asterisks. Uncharged residues in the membrane hydrophobic core are shaded yellow. The proteins are listed under their UniProt identifiers.

#### 4.1. Unusual operon structure of N-ATPases. Additional genes and their possible functions

A search of the complete genome database identified homologous N-ATPase operons in certain representatives of the bacterial phyla *Cyanobacteria*, *Aquificae*, *Chlorobi*, *Planctomycetes* and *Verrucomicrobia*, in some members of  $\alpha$ -,  $\beta$ -,  $\gamma$ -, and  $\delta$ -subdivisions of *Proteobacteria*, as well as in two archaea, *Methanosarcina barkeri* and *M. acetivorans*. All the operons had a typical structure as shown on example of cyanobacteria in **Table 4.1**. The gene order in the operons was *atpDCI-urf2-atpBEFAG*, encoding, respectively,  $\beta$ ,  $\epsilon$ , 1, Urf2,  $\alpha$ ,  $c$ ,  $b$ ,  $\alpha$  and  $\gamma$  subunits of this particular ATPase, which was first described in *M. barkeri* as an "archaeal F-type ATPase" more than 15 years ago (Sumi *et al.*, 1997). In the genome of *Persephonella marina* (phylum *Aquificae*) the exception from of the above mentioned rule was observed: the N-ATPase operon had an unusual gene order *urf2-atpBEFAGDCI* and additional operons for F-type and A-type ATPases were coded in the genome.

Identification of the N-ATPase operons in phylogenetically diverse microorganisms was greatly simplified by the fact that these operons were always present in the genomes as second copies alongside operons that encoded typical  $H^+$ -transporting  $F_1F_0$ -type ATPases (in *M. barkeri* and *M. acetivorans*,  $A_1A_0$ -type ATPases, respectively). In contrast, the N-ATPase *c* subunits contained full sets of  $Na^+$  ligands, indicating that respective enzymes were specific for  $Na^+$  ions (**Figure 4.2**). Just like the cyanobacterial  $Na^+$ -binding *c* subunits, *c* subunits of all N-ATPases had Glu residues in the middle of both transmembrane helices (see also below).

A distinct trait of the N-ATPase operons was the presence of an extra gene, *urf2* (Sumi *et al.*, 1997). In the UniProt database these proteins had not yet received annotation and were called "Putative uncharacterized protein". The *urf2* gene (hereafter referred to as *atpR*) was found in every N-ATPase operon and actually only in these operons, thus it could be used as a tell-tale sign of these operons. Its product contains three predicted transmembrane segments, two of them containing highly conserved Arg residues (Arg-49 and Arg-79 marked in **Figure 4.3**). The Presence of charged residues in the hydrophobic core of the membrane is very rare and should be an indicator of a functional role.

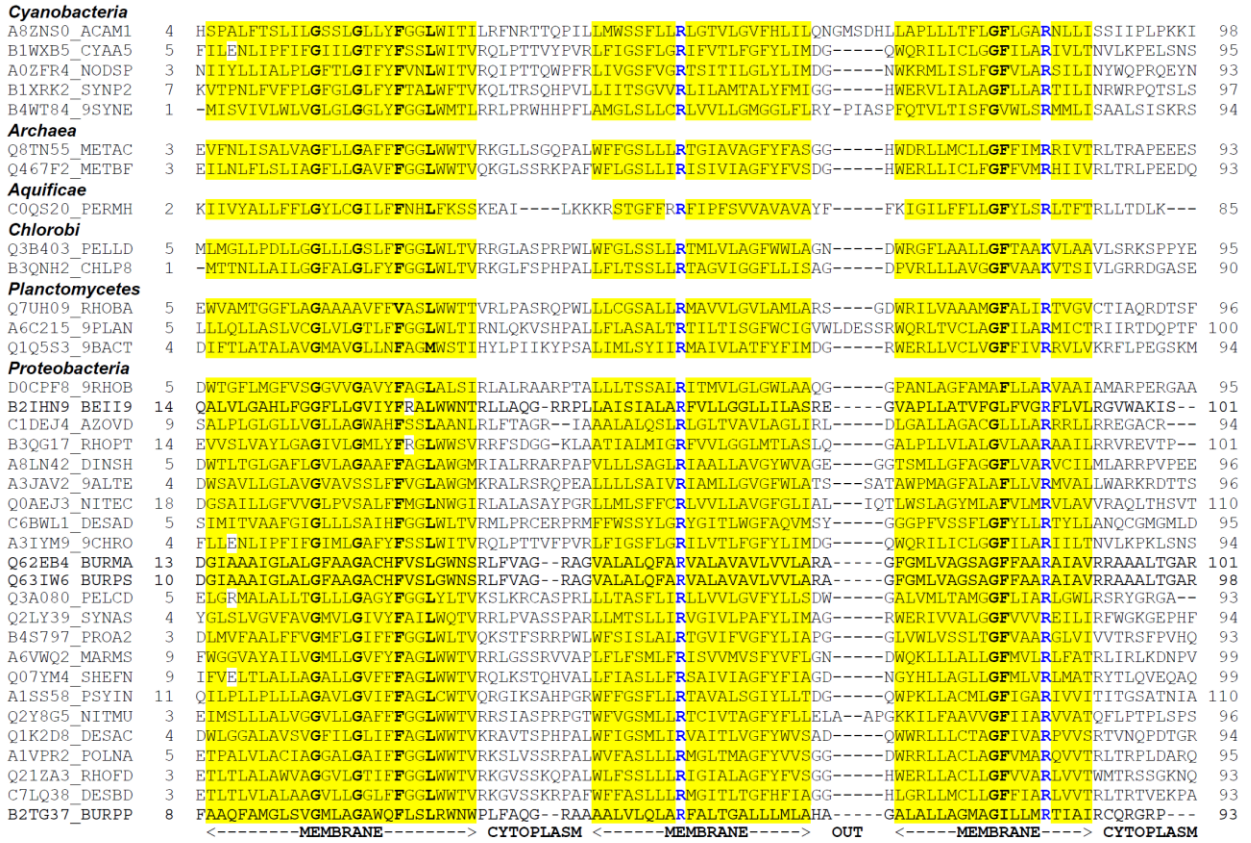


Figure 4.3. Alignment of the AtpR subunits of the N-ATPase (from (Dibrova *et al.*, 2010)).

The last line shows transmembrane topology prediction from TMHMM and Phobius (Kall *et al.*, 2007) servers. The characteristic Arg residues located inside the 2<sup>nd</sup> and the 3<sup>rd</sup> TM segments are colored blue. The proteins are listed under their UniProt identifiers.

The N-ATPase operons also include so-called "gene 1" (hereafter *atpQ*) that encodes a second distinct membrane protein. Although this gene was originally marked as *atpI* (Sumi *et al.*, 1997) and is still labelled this way occasionally (Saum *et al.*, 2009), its product shows no statistically significant sequence similarity to AtpI proteins previously described in the genomes of *E. coli* and other model organisms. In accordance with their distinct sequences, AtpI and AtpQ proteins are assigned to two different Pfam families, PF03899 and PF09527, respectively, and to two different COGs, COG3312 and COG5536. The *E. coli* AtpI (UncI) protein has been characterized as a chaperone that assists the assembly of *c*-subunit monomers into the membrane *c*-ring (Brandt *et al.*, 2012; Suzuki *et al.*, 2007). AtpQ definitely lacks any highly conserved positively charged (Arg or Lys) residue located in the middle of the transmembrane hydrophobic segment while AtpI possesses it. In AtpI this

residue is suggested to be involved in the interaction with the Glu/Asp residue of AtpE. Thus the function of AtpQ remains unknown. In contrast to *atpR*, the *atpQ* gene is not limited to the N-ATPase operons.

The initial description of the *Methanosarcina* N-ATPase operons (Sumi *et al.*, 1997) revealed an apparent absence of the *atpH* gene encoding the  $\delta$ -subunit of the F-type ATPase. The authors considered the idea that the unusually long *atpF* gene of the N-ATPase (*atpF<sub>N</sub>*) could encode a fusion of *b*- and  $\delta$ - subunits but rejected it because of the apparent lack of similarity between AtpF<sub>N</sub> and the AtpH gene products in the C-terminal part. Our analysis, however, shows that the coiled-coil C-terminal regions of these sequences can be aligned (**Figure 4.4**), which implies that the N-ATPase contains at least part of the  $\delta$ -subunit, in agreement with the earlier analyses (Pallen *et al.*, 2006; Saum *et al.*, 2009).

```

Q8TN58_METAC MA2435      175 ASSPVLVRFHAFDLPQAQRELTKKTKETL----GIEVQVCFETVPDLVSGIHELTANGQKVAVWSIHAGYLSLSEEGIDELLQEQPRSEARTEP 261
Q467F5_METBF Mbar_A3100 175 SPGQVLRHIRAFDLPQAQRDSIKKTKETL----GIEIQPRFETAPDLVSGIHELTTDGGQKVAVWSIHADYLSMSQKSIDELLNEQPKENKARTEP 261
C0QS17_PERMH PERMA_1699 174 ERNTVTVEHTYAPLSEEEIESSKSKKDMF----GIDVTKTEERKDLIAGVHKIHTASKMIDASLEGQISVFEETLIRDKIETS----- 251
ATPF2_ACAM1 AM1_D0164    176 RGEELVLSHSTFFMPEEDRKAQILAVLHQYAPA--MESSVQFVTIPDLICGIEHLKIPGYKLAWTIEQYLEEQLEVKNQVWDGMEKSWVEQG-- 263
ATPF1_CYAA5 cce_1506     175 TDNGLIIRHSHFEISPESSRNRLLSSLQOQTHI---YQGDNVQFMNSDLICGIEHLQASDYKIAWNLKDYVEALEI----- 244
A0ZFR7_NODSP N9414_04310 172 SATEIIIIHYSFDIPQTQRQETIKILQSQOI---INSKNKFTTSPDLICGIEHLQISNYQISWTFDDYLQTLSEQSTTIKQTTNKP---- 254
ATPF1_PELLD Plut_1069      170 SGGTIVLRHSGFEMGEEEEKELVRKTLADRF----GYGRIDFMTEESYRGGIHALEQGGRSIEWNSRLEATDEASSALLDGPDDMENEEEG 257
ATPF1_PROA2 Paes_0894    175 KSLNELVLRHSTFFLSSSLQERIQAVVRDVF---SIDTQLRFEEGDDMVGGIHELSMNGHSVSWVRSYLDSEKMTAELLEAGAE----- 253
ATPF1_RHOBA RB4915       175 SGNPVLVRHSAKGLDSSDQNOIRDAIHRVFE---ENKVEVRFSEPALIAGIHEMDAGGYSIPWNAERTLKTMEANVA----- 246
AGC218_9PLAN PM8797T_02744 175 SDPQMLVLRHSAFELTQPERKNTIDLIHDYI---SREVTIRFRVNHELICGIEHLDHLAGYKISWNLQESLEEELEEEFVRSLNDVISTDSEPGI 261
ATPF1_DINSH Dshi_0441       170 DARQAVLRHTRDPLPEEVQARLRSDLDGVDL---EGVSRFRFVDPGQSPGLHNLRLGGAQLGWTVDHSYLNGLEATIAEAAGRPARRGHDA 255
ATPF2_NITMU Nmul_A1655     199 GIAVILVLRHSAFELPAEQQEKISTVIRETL---KEEVRFRNFGEVPLISGIEHMTAGGHKVAWSVAGYLASLEEEVSRLLNAKAVGEGEAK 285
ATPF2_RHOVD Rfer_1167      175 ASEPALVLRHSAFELPADQRAAIQNAINETF---SADIPHFATAPEVVGIEHLSTNGQKVWGSITDYLASLEKGVDELKERDKAEPQSEL 261
C1DEJ7_AZOV Avin_19730   171 ESAAALVLRHVPRTESAGELEACRQALARAL---GREPELKVAVDPGLIAGIHLEEGSNVAVRNSFRADLARLHRELHSHESA----- 246
ATPF1_PSEA6 Patl_2674      175 SGNQIVVLRHSAQPLAEAQKKQLLACLQOYL-20-APSIKLSSEIVPRLINGIHELTMGGKWLAWSTDNYLAEQLQEDVEAEFFPFTETLLGLPE 281
Q07YMI_SHEFN Sfri_3055     179 AAQPILLIRHAFALPTPQCSSLIAAAIKKIL---SHDIFIQFSIEPDVLSGIEHLSIKGQKISWSTINEYLSLAKRVDEELLQSPMTEQANDT 265
ATPF2_PELCD Pcar_2995     173 SERGVLVLRHSPFTLPELDRDITQGVQRAL---GEEIDMQQDRADMLGIEHLTVGGLKLSWGVDSYFQELRDVATLYDAQAATVSEGGSP 259
ATPF2_SYNAS SYN_02103     175 SEEGIVLRHSAFALFPEDRQRITDTVRDLIG---KPAVIRYQESSDLIGGIEHLASGHRVAWSISDYLEHLEQDDRVLVHEEVRLTKPKPS 262

pdb|2WSS_S      EEEEE EEEEE HHHHHH
ATPD_ECOLI     105 ATAENVHISAALSEQQLAKISAAMEKRI---SRVKINKCKIDKSVMAHGVITRAGDMVTDGSVHRGRTERLADVHQS----- 177
ATPO_BOVIN     137 GEVPCVTHASALDETTTLTELKTVLKSFLSK--GQVLKLEVKIDPSIMHGMIVIRIGEKEYDMSAKTKIQKLSRAMREIL----- 213
ATPD_ACAM1    AM1_0895     110 NTVLAEVHSTVELTDEQRHAITDKVKHMSQ---AAQVDLETSIDPDLIGHVIIKIGSQVLDASIRGQLRRMNSITSL----- 185
ATPD_NITMU    Nmul_A0307  106 GVLDKAVHISAFANSDAQKLDVTDLEAKF---KRKIEAKVSVADDLIGHVIVEIGDEVLDASVHRKLEAMAVAKS----- 178
ATPD_RHOVD    Rfer_0109   104 GSSDAVVYHSAFALDSTALAEALATLQREF---ARKLNVSQLQPELIGHIRVVVGGDEVLDSSVHKARLEQMKMALTA----- 176
ATPD_AZOV Avin_52190  106 KTIDVELEHYAYELSAEQLETIAAALSHKRL---DRSVNFRQVNPALIGHGLVIRAGDLVVDGSVHRKLSQLAELS----- 178
ATPD_PSEA6    Patl_4298   105 KEIEADVHSAEINAKQQADISAALKRI---ARKVKINCSDVPTLIAGHVKIRAGDTVIDGSVHRKSNRLTDAIQA----- 177
ATPD_SHEFN    Sfri_4048   105 KEVEADVHSAEISSEQQQQIVSVLEKRI---ARKVKINCSDVPTLIAGHVKIRAGDTVIDGSVHRKSNRLTDAIQA----- 177

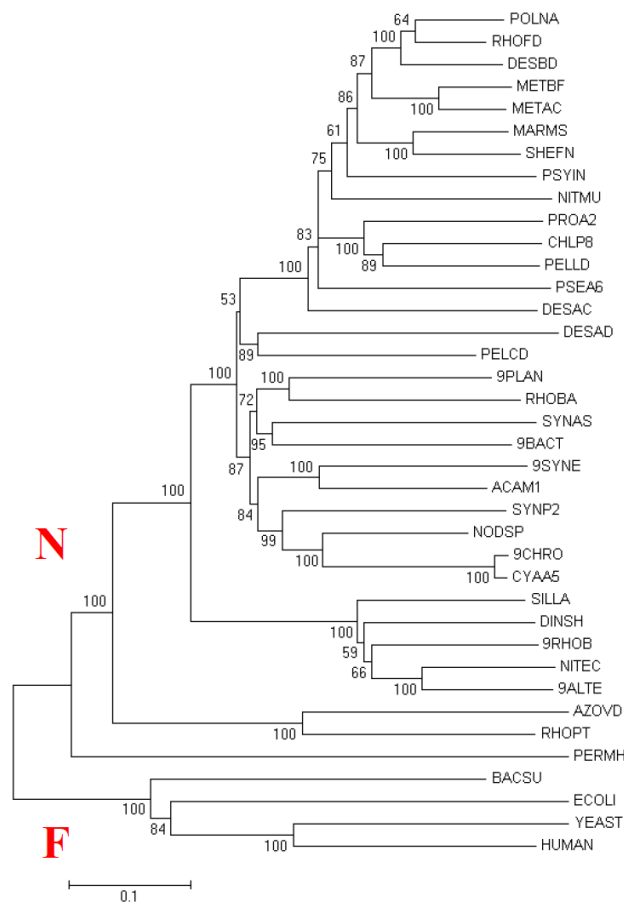
```

**Figure 4.4.** Alignment of the C-terminal fragments of the N-ATPase *b*-subunits (top) with  $\delta$ -subunits of F-ATPases from *Escherichia coli* (ATPD\_ECOLI), bovine mitochondria (ATPO\_BOVIN), and of H<sup>+</sup>-translocating F-ATPases from several organisms containing N-ATPases (bottom) (from (Dibrova *et al.*, 2010)).

Structural elements from the bovine oligomycin-sensitivity conferring protein (OSCP,  $\delta$ -subunit) structure (PDB entry: 2WSS\_S), reported by Rees *et al.* (Rees *et al.*, 2009), are indicated as follows: E,  $\beta$ -strand; H,  $\alpha$ -helix. Conserved Gly and Ser residues are shown in bold typeface, conserved hydrophobic amino acid residues are shaded yellow. The proteins are listed under their UniProt identifiers and genome locus names.

## 4.2. Phylogenomic analysis indicates that N-ATPases are a separate family of rotary membrane ATPases

The equivalent subunits of the N-ATPases from diverse bacteria were closely related and formed distinct branches on the phylogenetic trees, well separated from the corresponding subunits of the F-type ATPases (**Figure 4.5**).



**Figure 4.5.** Phylogenetic tree of concatenated  $\alpha$ -,  $\beta$ -,  $\epsilon$ -,  $c$ - and  $b$ -subunits from N-type ATPase operons (top) and F-type ATPase operons from *Bacillus subtilis*, *Escherichia coli*, yeast and human (bottom) (from (Dibrova *et al.*, 2010)).

The N-ATPase-encoding organisms are as follows: POLNA, *Polaromonas naphthalenivorans*; RHOFD, *Rhodoferrum ferrireducens*; DESBD, *Desulfomicrobium baculatum*; METAC, *Methanosarcina acetivorans*; METBF, *Methanosarcina barkeri*; NITMU, *Nitrosospira multiformis*; MARMS, *Marinomonas sp. strain MWYL1*; SHEFN, *Shewanella frigidimarina*; PSYIN, *Psychromonas ingrahamii*; PROA2, *Prosthecochloris aestuarii*; CHLP8, *Chlorobaculum parvum*; PELLD, *Pelodictyon luteolum*; DESAC, *Desulfuromonas acetoxidans*; PSEA6, *Pseudoalteromonas atlantica*; PELCD, *Pelobacter carbinolicus*;

ACAM1, *Acaryochloris marina*; 9SYNE *Synechococcus sp.* PCC 7335; SYN2, *Synechococcus sp.* PCC 7002; NODSP, *Nodularia spumigena*; CYAA5, *Cyanothece sp.* ATCC 51142; 9CHRO, *Cyanothece sp.* CCY0110; 9BACT, *Kuenenia stuttgartiensis*; RHOB, *Rhodospirillum rubrum*; 9PLAN, *Planctomyces maris*; DESAD, *Desulfovibrio salexigens*; 9RHOB (D0CPF5), *Silicibacter lacuscaerulensis*; DINSH, *Dinoroseobacter shibae*; 9RHOB (B7RJ35), *Roseobacter sp.* GAI101; 9ALTE, *Marinobacter sp.* ELB17; NITEC, *Nitrosomonas eutropha*; AZOVD, *Azotobacter vinelandii*; RHOPT, *Rhodopseudomonas palustris* strain TIE-1; PERMH, *Persephonella marina*; SYNAS, *Syntrophus aciditrophicus*.

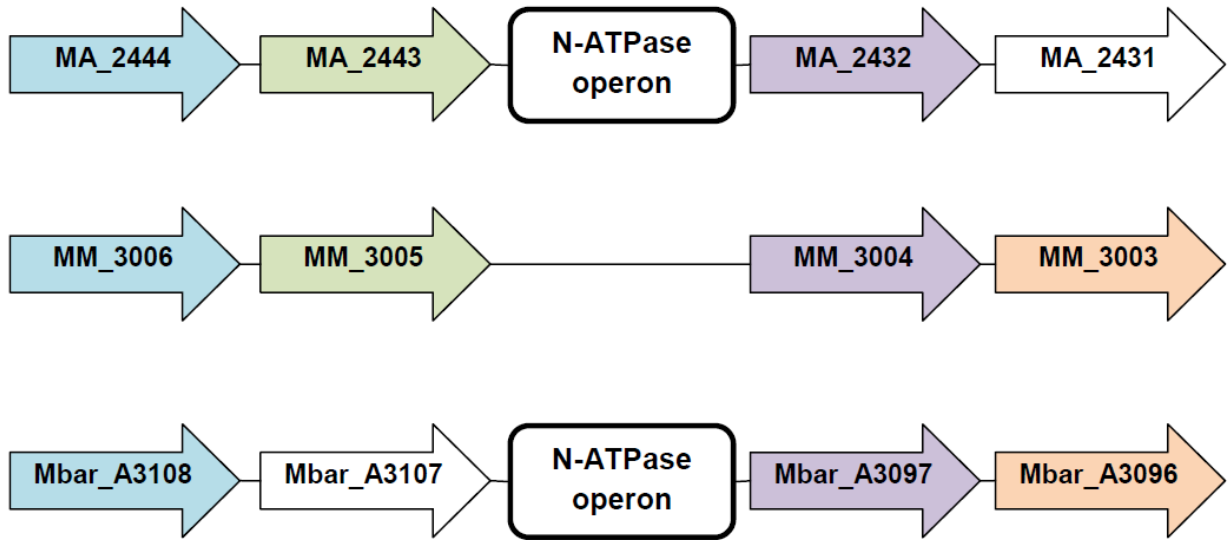
### 4.3. Evidences for the spreading of N-ATPases via the lateral gene transfer

Rotary membrane ATPases are among proteins which are only occasionally involved in the lateral gene transfer (LGT) (Hilario and Gogarten, 1993). However, the patchy distribution of N-ATPases in different taxons combined with their clearly separate position on the phylogenetic tree (Section 4.2) and distinctive features (Section 4.1) suggest the possibility of their LGT. In this section more specific evidence for the LGT of N-ATPases is presented.

Presence of the N-ATPase operon in the genomes of *M. barkeri* and *M. acetivorans*, but not in the closely related *M. mazei*, suggested that this operon had been acquired *via* lateral gene transfer by the two former archaea. This suggestion is fully consistent with the gene neighborhoods of the *atpDCQRBEFG* operons in *M. barkeri* and *M. acetivorans* (**Figure 4.6**) and the absence of these operons in any other archaeal genomes sequenced so far. Gene neighborhoods of the N-ATPase operons in various bacteria are also consistent with the insertion of this operon.

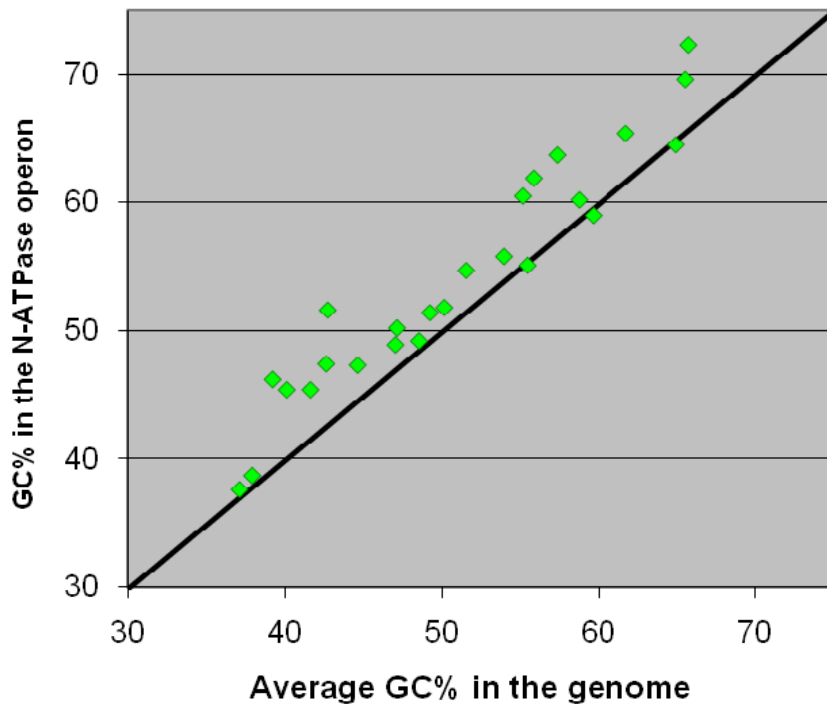
Generally, the spread of the N-ATPases among a plethora of diverse bacteria discounts the historical designation of these enzymes as "archaeobacterial F<sub>1</sub>F<sub>0</sub>-ATPases" (Sumi *et al.*, 1997). However, the GC content of the N-ATPase operons showed a good correlation with the average GC content of the host genomes (**Figure 4.7**), indicating either a relatively ancient gene transfers, or a rapid adaptation of the N-ATPase genes to their host environment. The strict conservation of the gene order and co-linearity of the phylogenetic trees for distinct N-ATPase subunits suggests that the whole operon was transferred as a single unit. An additional indication of the lateral mobility of the N-ATPase operon is its presence on plasmids, pREB4 in *A. marina* and pAQ7 in *Synechococcus* sp. PCC 7002. These data might be related to the earlier functional evidence of the presence of two ion-translocating ATPases in *M. mazei* Gö1. While the A-type ATP synthase was apparently H<sup>+</sup>-dependent, the second, F-type ATPase appeared to be Na<sup>+</sup>-translocating (Becher and Muller, 1994; Pisa *et al.*, 2007). Although this second ATPase has not been found in the sequenced genome of *M. mazei* (**Figure 4.6**), the respective genes could be plasmid-borne, as is the case of N-ATPases in at least two cyanobacteria.





**Figure 4.6.** Conserved gene neighborhoods in *M. acetivorans* (top), *M. mazei* (center), and *M. barkeri* (bottom) indicating the points of insertion of the N-ATPase operons.

Orthologous genes are indicated by the same colors. The arrows do not reflect the lengths of the genes.



**Figure 4.7.** Correlation between the GC content of the N-ATPase genes and of the total genome (from reference (Dibrova *et al.*, 2010)).

GC content is calculated as a percentage of guanine and cytosine nucleobases. Each dot corresponds to one N-ATPase operon.

#### 4.4. Discussion: specific features of N-ATPases as possible ancient features of rotary membrane ATPases

Following the first description of an "archaeobacterial F<sub>1</sub>F<sub>0</sub>-ATPase" in *M. barkeri* (Sumi *et al.*, 1997), the presence of a "*Methanosarcina*-like" F<sub>1</sub>F<sub>0</sub>-ATPase operon was repeatedly noted in bacterial genomes (Glockner *et al.*, 2003; McInerney *et al.*, 2007; Swingley *et al.*, 2008), although the exact function(s) of these enzymes and their cation specificity remained obscure. McInerney *et al.* (McInerney *et al.*, 2007) while discussing the energy metabolism of *Syntrophus aciditrophicus*, noted two F-type ATPases encoded in its genome and suggested that both of them were Na<sup>+</sup>-translocating (based on the present analysis, the F-type ATPase of this organism is definitely H<sup>+</sup>-specific, whereas the cation specificity of its N-ATPase is unknown; it might be specific for Na<sup>+</sup>, see below). Analysis of the *A. marina* genome by Swingley *et al.* (Swingley *et al.*, 2008) noted the presence of a plasmid-encoded "set of ATP synthase genes that were arranged into a unique operon, ... conserved with full synteny in a remarkable array of organisms, including cyanobacteria, archaea, planctomycetes, chlorobi, and proteobacteria". (Swingley *et al.*, 2008). However, these authors slightly overstated their case by claiming that the "individual proteins do not clearly fit into any of the described families". The same authors also disputed the idea that these enzymes were Na<sup>+</sup>-translocating. Several recent genome papers noted the presence of the second F-type ATP synthase (ATPase) operon with an unusual phylogenetic affinity (Davidsen *et al.*, 2010; Hou *et al.*, 2008; Setubal *et al.*, 2009).

**Figure 4.2** shows that, despite the previous doubts (Swingley *et al.*, 2008), the *c*-subunit of the *A. marina* N-ATPase has a full set of Na<sup>+</sup>-binding ligands, including the recently recognized additional Thr residue (Meier *et al.*, 2009; Mulkidjanian *et al.*, 2008b). This residue, however, is missing in *c* subunits of several N-ATPases, including the one from *S. aciditrophicus*, where a Glu residue is invariably present instead of the Na<sup>+</sup>-coordinating Gln residue in the first transmembrane helix of the *c*-subunit of all N-ATPases (**Figure 4.2**). As noted previously (Meier *et al.*, 2009; Saum *et al.*, 2009), this Glu residue could potentially provide two ligands for the Na<sup>+</sup> ion and thereby complete the Na<sup>+</sup> coordination shell. If so, all N-ATPases would end up being Na<sup>+</sup>-dependent enzymes.

A hallmark of the N-ATPase operons is the presence of the *atpR* gene. Because of the low dielectric permittivity of the membrane, the strategic positioning of two Arg residues of AtpR in the hydrophobic core of the membrane implies the presence of negatively charged residues somewhere nearby. Given the absence in the N-ATPase operons of the *atpI* gene, whose product was recently shown to serve as a chaperone that assists the assembly of *c*-subunit monomers in the membrane into the *c*-ring (Suzuki *et al.*, 2007), essentially the same function can be suggested for the product of the AtpR gene. Just like AtpI assists *c*-ring assembly by directly interacting with the *c* subunits, AtpR could do that through the interaction of its two Arg residues with N-ATPase-specific *c*-subunits, which all carry two Glu residues in the middle of their transmembrane helices (compare **Figure 4.2** and **Figure 4.3**).

The fact that N-ATPases are always found alongside typical F- or A-type ATPases may suggest that the N-ATPases cannot fully replace those enzymes in their role as ATP synthases. Indeed, in *M. acetivorans*, deletion of the N-ATPase operon had no visible effect on cell growth or ATP synthesis, whereas a mutant lacking the A-ATP synthase genes could not be obtained (Saum *et al.*, 2009). Thus, ATP-driven ion outpumping could be suggested as the most plausible function for these enzymes (Dibrova *et al.*, 2010).

Acquisition of an operon capable of carrying out this function would be beneficial to the organisms living in high-salt environments, for example, marine bacteria. The extrusion of Na<sup>+</sup> ions is an important part of maintaining the ionic balance in the cytoplasm, particularly under stress conditions. Accordingly, despite the phylogenetic diversity of the N-ATPase-encoding bacteria, many of them are either marine organisms (all representatives of *Cyanobacteria*, *Planctomycetes*, *Persephonella marina*, *Dinoroseobacter shibae*, *Rhodoferrax ferrireducens*, all representatives of  $\gamma$ -Proteobacteria, except for *Azotobacter vinelandii*) or tend to grow in the presence of salt (e.g., *Chlorobaculum parvum*, *Chlorobium luteolum*, *Desulfovibrio salexigens*, *Silicibacter lacuscaerulensis*).

The N-ATPases are encoded in an apparently highly mobile operon that, most likely, confers on (at least some of) its hosts the ability of ATP-driven outward pumping of Na<sup>+</sup> ions, which complements the H<sup>+</sup> specificity of the native chromosome-encoded F-ATPase (or A-ATPase). Therefore we have speculated that, similarly to the eukaryotic V-ATPases, the N-ATPases do not catalyze ATP synthesis, which is why an N-ATPase is never found alone in

a genome, but rather only as the second enzyme in cells that already encode an F-type or a V-type ATP synthase (Dibrova *et al.*, 2010).

After the appearance of our article with the characterization of N-ATPases, several reports provided factual data on this enzyme. The operon of N-ATPase was shown to be transcribed and regulated (the level of transcription depended on the conditions) in the cyanobacterium *Synechococcus sp.* PCC 7002. In the normal growth conditions, when the cells were exposed to light, its transcription level was around 5% of the transcription level of F-ATPase, but it was increased 3-fold during the incubation in the dark (Ludwig and Bryant, 2011). In another experiment, the transcription level of the operon of the N-ATPase in *Rhodospseudomonas palustris* was increased upon adding  $\text{Cu}^{2+}$ , but not  $\text{Fe}^{2+}$  salts (Bird *et al.*, 2013); authors suggested that the N-ATPase could play a role in the response to the metal toxicity. Expression study for the proteobacterium *Gluconobacter oxydans* has detected 2-3-fold increase in transcription levels of all genes in the N-ATPase operon under the conditions of oxygen limitation, while transcription levels of genes coding for a normal  $\text{F}_1\text{F}_0$ -ATPase decreased by 50% (Hanke *et al.*, 2012).

A direct experimental verification of our predictions was performed for the N-ATPase of *Aphanothece halophytica*, a halotolerant cyanobacteria isolated from the Dead Sea. The whole N-ATPase operon from *A. halophytica* genome was cut and expressed in the  $\Delta atp$  strain of *E. coli* which did not contain a native ATP synthase (Soontharapirakkul *et al.*, 2011). Inverted membrane vesicles were prepared by cell disruption in a French pressure cell. The ATPase activity in the fraction of inverted membrane vesicles increased almost linearly with an increase in  $\text{Na}^+$  concentration until 10 mM NaCl, after which the activity remained constant. In contrast, the ATPase activity in vesicles from control *E. coli* cells, which were transferred with an empty vector, was very low and did not increase with the increase in  $\text{Na}^+$  concentration. Inhibitor analysis suggested that the ATPase activity was due to the action of a  $\text{Na}^+$ -dependent rotary membrane ATPase (Soontharapirakkul *et al.*, 2011). An inhibitor of  $\text{F}_1$ , azide, decreased the activity by 55%, the inhibitors of  $\text{F}_0$ , DCCD (N,N'-dicyclohexylcarbodiimide) and tributyltin chloride, were 70% and 80% effective, respectively. A similar inhibitory effect has been observed with a dissipator of  $\text{Na}^+$  gradient, monensin (70% of inhibition), but not with the protonophore CCCP (carbonyl cyanide m-chlorophenyl hydrazone). Addition of  $\text{KNO}_3$  (a permeant anion) also did not affect the

ATPase activity. An ATP synthase activity was detected with luciferin/luciferase reaction on the inverted membrane vesicles from the cells which obtained a vector with N-ATPase genes, while no activity was seen for the vesicles from the control cells. When vesicles were pre-loaded with  $\text{Na}^+$ , DCCD and monensin inhibited the ATP synthase activity, while CCCP did not. Also, the ATP synthesis rate decreased upon addition of external  $\text{Na}^+$  (which decreased the SMF) (Soontharapirakkul *et al.*, 2011). Transformation of the freshwater cyanobacterium *Synechococcus sp.* PCC 7942 with the vector that carried an N-ATPase operon gave the cells an advantage during increase in NaCl concentration when compared to the cells that were transformed with an empty vector. The cells transformed with an N-ATPase operon still grew in the presence of 0.5 M NaCl, whereas growth of the cells that lacked the N-ATPase was strongly suppressed. In the presence of 0.3 M NaCl, cells with the N-ATPase grew faster than cell that lacked it. When membrane vesicles of *Synechococcus sp.* PCC 7942 transformed with N-ATPase operon were analyzed by immunoblot analysis, an increase in production of  $\gamma$ -subunit (antibodies were recognizing a His-tag added to it) upon increasing external NaCl was observed (Soontharapirakkul *et al.*, 2011). This experiment, taking into account the possibility of ATP hydrolysis by the N-ATPase described above, suggested that, in agreement with our earlier suggestion (Dibrova *et al.*, 2010), N-ATPase could be used by the organisms living in high salt concentration to pump excess sodium out of the cells.

Based on a larger set of sequences than the one used by Swingley *et al.* (Swingley *et al.*, 2008), the present investigation has confirmed that N-ATPases branch separately from other F-type ATPases (**Figure 4.5**), which suggests a possible early divergence of the N-ATPases. They may be the missing "molecular fossil" from the time when all membrane bioenergetics was  $\text{Na}^+$ -dependent, as suggested before (Mulkidjanian *et al.*, 2008a; Mulkidjanian *et al.*, 2008b). This is compatible with the following, supposedly ancient traits of these enzymes:

- Outward pumping of  $\text{Na}^+$  ions, the predicted function of N-ATPases, appears to be an ancient trait and has been previously suggested as a possible function of the common ancestor of the F- and V-type ATPases, which likely had a  $\text{Na}^+$ -binding *c*-oligomer (Mulkidjanian *et al.*, 2008a; Mulkidjanian *et al.*, 2008b; Mulkidjanian *et al.*, 2007).
- While in the typical F-type ATPases the peripheral stalk is built from separate  $\delta$  and *b* subunits, in the V/A-type ATPases the equivalent E subunit of the peripheral stalk contains stretches that correspond to the fused *b* and  $\delta$  subunits of typical F-ATPases,

respectively (Pallen *et al.*, 2006), similarly to the "long" *b* subunit AtpF<sub>N</sub> of the N-ATPase.

- The *c*-subunit is believed to evolve from a duplication of an amphiphilic helix that contained a Glu residue in the middle (Davis, 2002). The supposedly ancient two-Glu architecture, besides N-ATPases, is found in the *c*-subunits of the A-type ATPases of some euryarchaea (e.g. methanogens), as well as in the F-ATPase *c*-subunits of some early branching bacteria, such as *Thermotogae*.

All these features, which N-ATPases share either with V/A-ATPases – to the exclusion of most F-ATPases – or with the putative common ancestor of all rotary ATPases suggest that N-ATPases represent a distinct early-diverging family of rotating ATPases. Further experimental verification of the predicted functions of the N-ATPases would be useful to help understanding the evolution of energy conservation mechanisms.

## 5. Phylogenomic analysis of the cytochrome *bc* complex

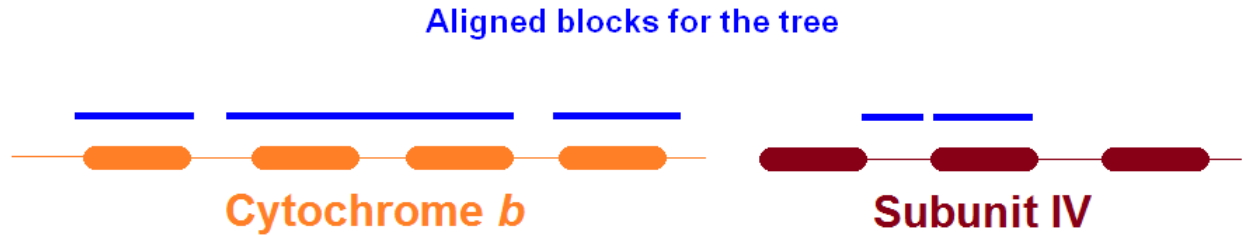
This chapter, which contains results published in (Dibrova *et al.*, 2013), is devoted to the evolutionary relations among various cytochrome *bc* complexes (reviewed in Section 1.5.3.3). We attempted to answer the following questions:

1. Could cytochrome *bc* complexes be present already in the LUCA?
2. Which evolutionary scenario (fusion of short cytochrome *b* and subunit IV or fission of the long ancestral protein) is more plausible for cytochrome *bc* complexes?
3. What could have driven the separation into the two major families, the *bc*<sub>1</sub>-type complexes and the *b<sub>6f</sub>*-type complexes, respectively?

### 5.1. Analysis of the phylogenetic tree of the cytochromes *b*

As noted in Section 1.5.3.3, the Rieske iron-sulfur protein and the cytochrome *b* are the only components that are shared by the *bc*<sub>1</sub>-type and the *b<sub>6f</sub>*-type complexes. The redox carriers (usually, *c*-type cytochromes) which accept electrons from the FeS domain of the Rieske protein are not homologous in different lineages and therefore were not included into the analysis. The Rieske protein, although ubiquitously present, was previously shown to carry weak phylogenetic signal (Lebrun *et al.*, 2006); the tests performed in this study have also shown that its inclusion into the analysis did not significantly influence the tree topology. For these reasons the phylogenetic tree has been built only on the sequences of "long" cytochromes *b* of the *bc*<sub>1</sub>-type complexes and the corresponding sequences of "short" cytochromes *b* plus subunit IV of the *b<sub>6f</sub>*-type complexes.

We have constructed a multiple alignment of cytochromes *b*, as described in Section 2.7.3. Positions of the blocks are shown schematically in the **Figure 5.1**. The resulting tree is shown in **Figure 5.2**. Its alternative view is given in **Figure 5.3**.



**Figure 5.1.** Mapping of the conserved blocks on the schematic representation of cytochrome *b* and subunit IV sequences from cyanobacterial cytochrome *b<sub>6f</sub>*-complex.

Thick segments are showing transmembrane helices. Blue lines show well-aligned regions (so-called "blocks") that were used for the tree construction.

**Review of the tree topology.** As could be seen in *Figure 5.2*, the cytochrome *b* sequences separate into several subfamilies (clades of the tree), the members of which not only show sequence similarity, but also, in many cases, share specific functionally relevant traits. Both the bootstrap values supporting major branches of the tree and some specific features of proteins analyzed are depicted in *Figure 5.2*. It is noteworthy that, in a number of genomes, several homologues of the cytochrome *b* could be identified. An analysis of the respective operons showed the genes of Rieske proteins upstream of the cytochrome *b* gene(s) in most of them, so these multiple operons most likely code for cytochrome *bc* complexes, as already noted earlier (Baymann *et al.*, 2012). The organisms with several operons of cytochrome *bc* complexes are found both within bacteria (e.g. among *Planctomycetes* *Table 5.2*) and among archaea (e.g. in *Halobacteria* *Table 5.3*). The phylogenetic tree in *Figure 5.2* shows that the multiple copies of the cytochrome *bc* complex-encoding operons do not result from duplication within these phyla. In the case of *Planctomycetes*, the sequences of cytochromes *b* do not cluster together but belong to separate clades (D and G). Moreover, sequences from clade D are full-length cytochromes *b* while sequences from clade G are split cytochromes *b* of the *b<sub>6f</sub>* type. In the case of archaeal cytochrome *bc* complexes, the split cytochrome *b* branches next to the branch G, while the "fused" archaeal cytochromes *b* belong to the branches E and F.

The sequence of *Thermodesulfator indicus* cytochrome *b* is a strongly diverged one. While bearing most of the conserved motifs including heme-binding histidine residues, it is truncated in several places. This fact is represented with a long branch on the phylogenetic tree, and the real position of this sequence on the tree is actually unknown. However, on the



full-length tree it was grouped with the archaeal sequences from clade E which also have relatively long branches. One can assume that this observation can be explained if the whole clade E is suffering from the long-branch attraction (for a review on this phenomenon see (Philippe *et al.*, 2005)). Long-branch attraction is an erroneous grouping of two or more long branches as sister groups due to methodological artifacts (Bergsten, 2005). Briefly, this could happen because divergent proteins could be treated by the algorithm of tree construction as having a common feature, namely the clear difference from other sequences. This feature, obviously, is due not to the direct relation between the proteins, but to the algorithms' failure to detect the exact positioning of the divergent branches and the tendency to place them together. Cases of long-branch attraction are not easy to prove, as the real tree topology is not known. However, cases of long branch attraction were identified on different phylogenetic "scales" that ranged from species-level phylogenies of invertebrates to the entire tree of life (Bergsten, 2005).

**Table 5.1. Color codes for the Figure 5.2.**

The terms "poorly sampled" and "well-sampled" refer to the number of full genome sequences available for the respective phyla.

**Archaea:**

Crenarchaeota  
 Euryarchaeota  
 Korarchaeota  
 Nanoarchaeota  
 Thaumarchaeota

**Well-sampled bacterial phyla:**

Actinobacteria  
 Aquificae  
 Bacteroidetes/Chlorobi group  
 Chlamydiae/Verrucomicrobia group  
 Chloroflexi  
 Cyanobacteria  
 Deinococcus-Thermus  
 Firmicutes  
 Proteobacteria  
 Thermotogae

**Poorly sampled bacterial phyla:**

Chrysiogenetes  
 Deferribacteres  
 Dictyoglomi  
 Elusimicrobia  
 Fibrobacteres/Acidobacteria group  
 Fusobacteria  
 Gemmatimonadetes  
 Nitrospirae  
 Planctomycetes  
 Spirochaetes  
 Synergistetes  
 Thermodesulfobacteria  
 unclassified Bacteria

**Table 5.2. Planctomycetal operons containing homologs of cytochrome *b*.**

Arrows represent the genes. Arrows of the same color correspond to homologous genes. Color code:  $\Rightarrow$  —

Rieske protein,  $\Rightarrow$  — cytochrome *b* ( $b_6$ -like part),  $\Rightarrow$  — cytochrome *b* (subunit IV-like part),  $\Rightarrow$  — uncharacterized

"oxidoreductase",  $\Rightarrow$  — proteins with unknown function.

#	ID ( <i>cyt. b</i> )	Organism	Operon structure
1	87308057	<i>Blastopirellula marina</i> DSM 3645	
2	873089 76-77		
3	87312209		
4	91200800	<i>Candidatus Kuenenia stuttgartiensis</i>	
5	912022 73-74		
6	912025 87-90		
7	91204467		
8	2837807 06-05	<i>Pirellula staleyi</i> DSM 6068	
9	283781609		
10	296120695	<i>Planctomyces limnophilus</i> DSM 3776	
11	149175601	<i>Planctomyces maris</i> DSM 8797	
12	149178012		
13	159901864	uncultured planctomycete 6N14	

**Table 5.3. Halobacterial operons containing homologs of cytochrome *b*.**See legend to **Table 5.2** for the details.

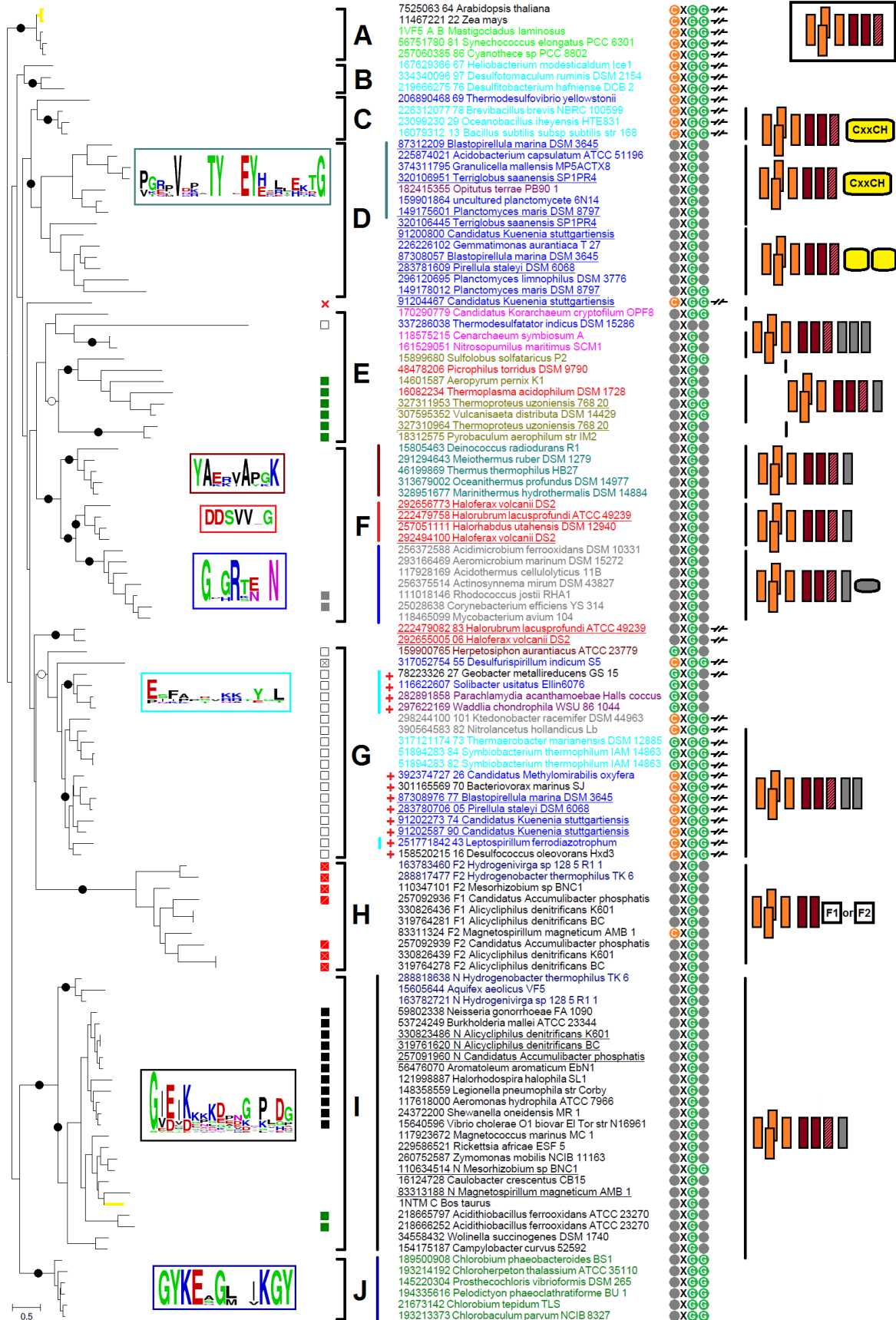
#	ID ( <i>cyt. b</i> )	Organism	Operon structure
1	222479082	<i>Halorubrum lacusprofundi</i> ATCC 49239	
2	222479758		
3	292655005	<i>Haloferax volcanii</i> DS2	
4	292656773		
5	292494100		 (plasmid pHV3)

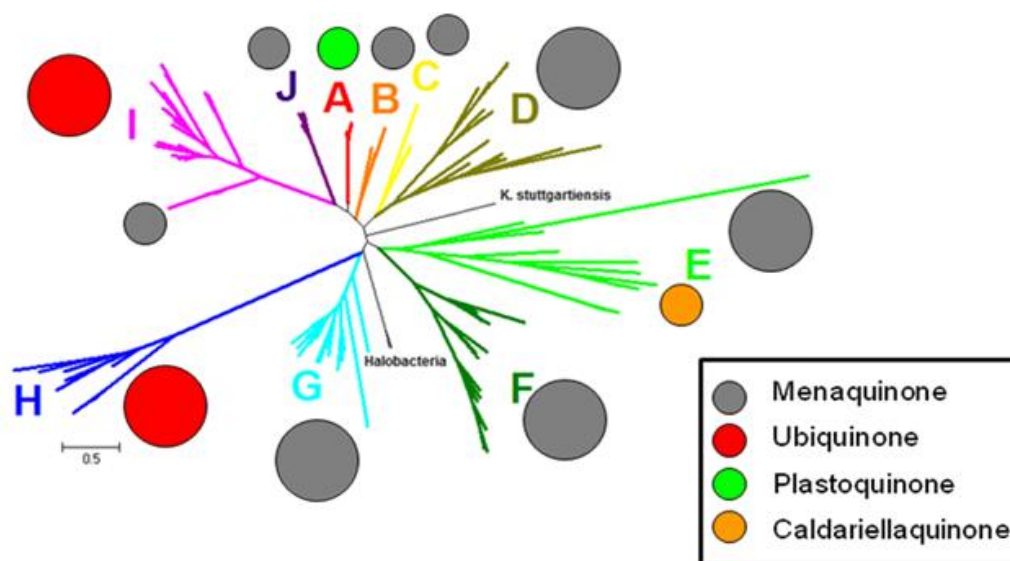
**Figure 5.2 (on the next page). Phylogenetic tree of cytochromes *b*.**

Each protein is indicated by its NCBI's GenInfo identifier (gi number), followed by the name of the source organism; in two instances, the proteins are labeled by their PDB codes (1VF5 and 1NTM). The colors of protein names indicate the taxonomical positions of the respective species. The detailed correspondence between colors and taxons is provided in **Table 5.1**. An alternative schematic representation of the same tree with indicated chemical nature of the predominant pool quinone is given as **Figure 5.3**. Black circles mark branches supported by more than 75% by bootstrap test, open circles mark branches with support below 75% but above 50%.

Clades marked on the trees are as follows: **A**, cyanobacterial and plant clade; eukaryotic sequences are marked with bold yellow branches; **B**, clade with cytochrome *b* of *Heliobacterium modesticaldum* and related proteins; **C**, clade of *Bacilli* members and *Thermodesulfovibrio yellowstonii*; **D** and **G**, unusual clades, see the main text; **E**, mostly archaeal clade; **F**, clade with sequences from *Deinococcus-Thermus* bacteria, actinobacteria and haloarchaea; **H**, fusions between cytochrome *b* and different sets of redox domains; **I**, proteobacterial clade with mitochondrial cytochromes *b* and proteins from *Aquificae* (the mitochondrial cytochrome is indicated by an orange-colored branch), **J**, *Chlorobi* clade. Specific marks are as follows:

- 1) Circled motif belongs to the beginning of the first transmembrane helix, which in cyanobacterial complexes is responsible for binding the third heme. Consensus sequence of the functional motif is CxGG.
- 2) Squares with different filling show the following deviations from the typical quinone-binding motif P[DE]W[FY] in the subunit IV (or in the C-terminal part of the "long" cytochrome *b*): black square, PVW[FY], green square, PPW[FY], gray square, PDIY, white square, P–W[FY], white square with a cross, G–WF, red square with a cross, [LV]DW[FY], red square with a slash, FDW[FY].
- 3) The red cross symbol indicates the absence of subunit IV.
- 4) Complexes with subunit IV as a separate protein are marked with symbol –/ /–. The gi's of the pairs "cytochrome *b* - subunit IV" are separated with a space (for instance, "7525063 64" means that cytochrome *b*<sub>6</sub>-like and subunit IV-like proteins have gi's 7525063 and 7525064, respectively).
- 5) If the genome contains sequences coding for unusual cytochromes *c* (see main text), proteins from this genome are marked with the "red plus" sign.
- 6) Lines of different colors before the names of the proteins indicate different conservative linkers between the cytochrome *b*<sub>6</sub>-like parts and subunit IV-like parts of the full-length cytochromes *b*. Frames of the same color to the left show sequence logo diagrams for these linkers.
- 7) Figures after the species names depict divergence from the typical structure shown in a rectangle in the top right corner. Four orange rectangles correspond to a 4-helical bundle (cytochrome *b*<sub>6</sub>-like part), two dark red rectangles depict two well-aligned helices of the subunit IV, hatched dark rectangles indicate an unaligned helix of subunit IV. Vertical grey rectangles correspond to the additional helices after subunit IV, yellow rectangles with round edges signify domains with heme-binding sites (cytochrome *c*-like domains), the small grey rectangle with round edges indicates a small domain conserved in actinobacteria. Finally, rectangles with marks "F1" and "F2" in the clade **H** indicate two types of fusions with different sets of domains.





**Figure 5.3. Alternative schematic representation of the phylogenetic tree of Figure 5.2 with indicated chemical nature of the predominant pool quinone.**

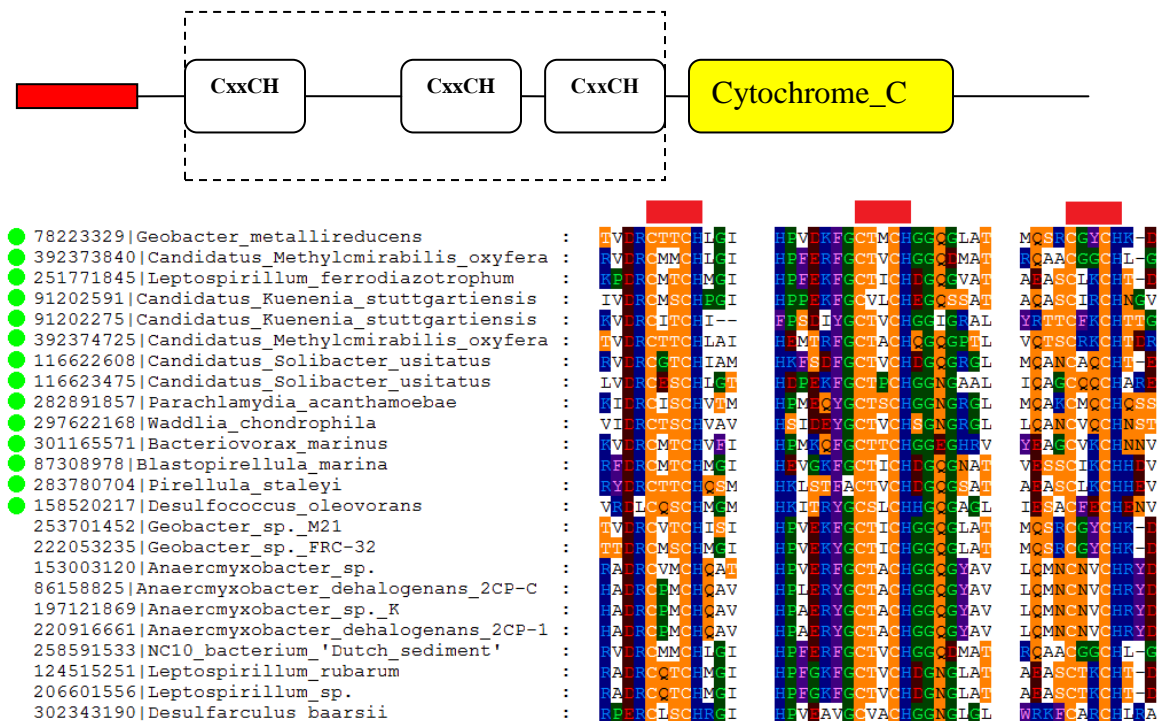
The type of the pool quinone in the organisms corresponding to each clade is depicted with the colored circle; the legend for each color is given in the box.

**Existence of at least two groups of split cytochromes *b*: distinct properties of the clade G.** The main feature of the phylogenetic trees in *Figure 5.2* and *Figure 5.3* is that the split cytochromes *b* of the  $b_6f$ -type complexes cluster together in two different parts of the tree. The first clade of split cytochromes *b* contains sequences from *Cyanobacteria* (and chloroplasts), *Firmicutes* (*Clostridia* and *Bacilli*), and *Thermodesulfobacteria*. These are clades A, B, and C in the upper part of the tree (these three groups, interestingly, do not form a single group). The second clade that contains split cytochromes *b* is the clade G that is located on the tree far away from clades A, B, and C. It contains sequences from organisms belonging to a number of bacterial phyla: *Chloroflexi*, *Chrysiogenetes*, *Proteobacteria*, *Acidobacteria*, *Chlamydiae*, *Firmicutes*, *Nitrospirae*, *Planctomycetes* and even a member of a newly proposed candidate division NC10, *Candidatus Methylomirabilis oxyfera* (Ettwig *et al.*, 2010).

Most of the cytochrome *b* sequences contain a quinol-binding motif P[DE]W[FY] either in the cytochrome *b* ("long" cytochromes *b*) or in the subunit IV (in case of split cytochromes *b*). Still, some clades show replacements in this motif (see square marks in *Figure 5.2*). Specifically, the members of the clade G share the P–W[FY] motif. The short cytochromes *b*, which form clades A, B, C and G, almost all have a conserved cysteine residue close to the

N-terminus, just before the first transmembrane helix. This residue was shown to bind the third heme ( $c_n$ ) in the  $b_f$ -type complexes of plants and cyanobacteria (Kurisu *et al.*, 2003; Stroebel *et al.*, 2003), as well as in the cytochrome  $bc$  complexes of *Firmicutes* (Baymann and Nitschke, 2010; Yu and Le Brun, 1998). The rather strict conservation of this cysteine residue within clade G might indicate the presence of the third heme also in the cytochromes  $b$  belonging to this clade, which, unfortunately, has not yet been experimentally characterized.

The split cytochromes  $b$ , which belong to the clade G, co-occur with large proteins possessing a NADPH-binding domain and many conservative cysteine residues capable of binding FeS-clusters. In prokaryotic genomes, these proteins share operons with ferredoxin-NADPH oxidoreductases and are mostly annotated as subunits of glutamate synthase. We failed to find experimental data on any of the proteins belonging to this group. In some cases, the genes that encode proteins of this family are found within the operons of cytochrome  $bc$  complexes (*Table 5.2*), as noted earlier (Baymann *et al.*, 2012).



**Figure 5.4.** Scheme of the functional parts in unusual  $c$ -type cytochromes found in clade G (above) and three heme-binding regions in the respective alignment (below).

Sequences from the genomes with cytochromes  $b$  from clade G are marked with green circle to the left.








In addition, the presence of cytochromes *b* that belong to the clade G often correlates with the presence of a protein that shows similarity (e-value  $\sim 10^{-10}$ ) to a typical cytochrome *c* Pfam domain (PF00034, Cytochrome\_C) and contains two-three heme binding motifs, see **Figure 5.4**. Since neither typical cytochrome *c*<sub>1</sub> nor cytochrome *f* were found in the respective genomes it is tempting to suggest that these *c*-type cytochromes are functional analogues of cytochrome *c*<sub>1</sub>/*f* and accept electrons from the FeS cluster of the Rieske protein. The occurrence of the specific *c*-type cytochrome and a putative dehydrogenase within the operons prompts us to consider the *b*<sub>6</sub>*f*-type complexes forming the clade G as a separate subfamily of cytochrome *bc* complexes.

**Fusions of cytochromes *b* with other domains: clade H.** A number of bacteria have cytochrome *b*-containing, fused genes (**Table 5.5**) along with the "classical" *bc* complex operons (**Table 5.4**) in their genomes. Two types of fusions have been identified in this work so far, labeled F1 (strings 3b, 4b and 5b in **Table 5.5**) and F2. Fusions were observed with the domains which presumably bind FeS-clusters (light green and red color in **Table 5.5** show domains with binding sites for Fe<sub>2</sub>S<sub>2</sub> and Fe<sub>4</sub>S<sub>4</sub> clusters, respectively) and other cofactors (FAD and NAD<sup>+</sup>-binding domains in the F1-type fusions). F1- and F2-type cytochromes *b* form a separate clade on the phylogenetic tree (clade H in **Figure 5.2**) while all cytochromes *b* from "classical" operons of the same organisms fall within the clade I. Operons with fused cytochromes *b* do not contain Rieske proteins and have only very short linkers between the *b*<sub>6</sub>-like and the subunit IV-like parts of cytochrome *b*. The function of these cytochromes *b* should be established yet.

Genes with the F1 type of fusion (rows #3b, #4b and #5b) could be found only in an operon downstream of genes with F2 type of fusion, whereas the latter can occur in the genomes without F1 type fusions.

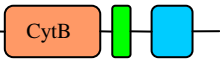
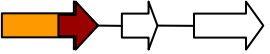
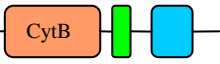

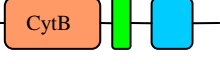




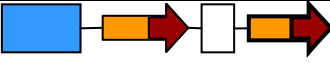
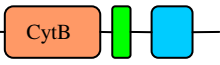
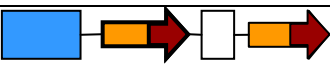

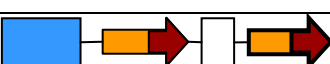

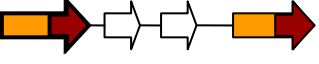

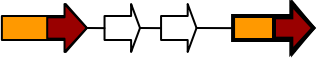


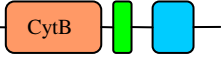
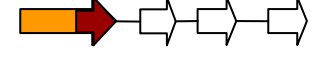


**Table 5.4.** "Typical" cytochrome *bc* complexes identified in the genomes that also contain "fused" cytochromes *b*, as shown in Table 5.5.

#	ID ( <i>cyt. b</i> )	Organism	Operon structure
1	163782721	<i>Hydrogenivirga</i> sp. 128-5-R1-1	
2	288818638	<i>Hydrogenobacter thermophilus</i> TK-6	
3	330823486	<i>Alicyclophilus denitrificans</i> K601	
4	319761620	<i>Alicyclophilus denitrificans</i> K601	
5	257091960	<i>Candidatus Accumulibacter phosphatis</i> clade IIA str. UW-1	
6	83313188	<i>Magnetospirillum magneticum</i> AMB-1	
7	110634514	<i>Mesorhizobium</i> sp. BNC1 ( <i>Chelativorans</i> sp. BNC1)	

**Table 5.5. Proteins with a fusion between cytochrome *b* and different redox-domains.**

In operons with two cytochromes *b* the one with shown domain structure is outlined by a bold line. Colors of the domains – **green**, Fe4\_7 (PF12838); **blue**, FlpD (PF02662); **dark green**, Fer2 (PF00111); **red**, FAD\_binding\_6 (PF00970); **violet**, NAD\_binding\_1 (PF00175); **yellow**, *c*-type cytochrome.

#	ID (cyt. <i>b</i> )	Organism	Domain structure (cyt. <i>b</i> )	Operon structure
1	163783460	<i>Hydrogenivirga</i> sp. 128-5-R1-1		
2	288817477	<i>Hydrogenobacter thermophilus</i> TK-6		
3a	330826439	<i>Alicyclophilus denitrificans</i> K601	 (F2)	 <ul style="list-style-type: none"> <li> 5 genes: hypothetical, 2 cytochrome <i>c</i> oxidase subunits, protoheme IX farnesyltransferase, <math>\bar{e}</math> transfer protein SenC</li> <li> 2 genes: hypothetical protein and cytochrome oxidase subunit</li> </ul>
3b	330826436		 (F1)	
4a	319764278	<i>Alicyclophilus denitrificans</i> BC		
4b	319764281			
5a	257092939	<i>Candidatus Accumulibacter phosphatis</i> clade IIA str. UW-1		
5b	257092936			
6	83311324	<i>Magnetospirillum magneticum</i> AMB-1		
7	110347101	<i>Mesorhizobium</i> sp. BNC1 ( <i>Chelativorans</i> sp. BNC1)		

**Archaeal cytochromes *b* and the lateral gene transfer.** In the phylogenetic tree in *Figure 5.2*, the archaeal sequences belong mostly to the clades E and F. Only a pair of haloarchaeal sequences is found outside these clades, and they have rather uncommon operon structure

(**Table 5.3**, #1 and #3); these sequences do not group with other archaeal proteins and appear to be a result of a separate lateral gene transfer event.

The clade F, besides archaeal sequences, contains also the sequences of cytochromes *b* from *Actinobacteria* and the *Deinococcus-Thermus* group. It seems plausible that haloarchaeal sequences from clade F were obtained by LGT from bacteria (Baymann *et al.*, 2012). The opposite, less likely hypothesis would imply two independent LGTs from a halobacterial ancestor into *Actinobacteria* and *Deinococcus-Thermus* phyla.

The branches in the archaeal clade E are remarkably longer than most other individual leaf branches, whereas the branch that separates clade E from other clades is short. Vertical inheritance from a common ancestor would imply a reciprocal relation, namely a long separating branch and short branches within a clade. The observed pattern is compatible with a suggestion of a LGT from bacteria to archaea: a bacterial *bc* complex, after being transferred to archaea, finds itself in quite a different physico-chemical environment (as archaeal membranes differ fundamentally from the bacterial ones (Pereto *et al.*, 2004)), not to mention a new genome environment. Such abrupt changes could prompt fast adaptation of the laterally transferred cytochrome *bc* complexes to the new environments and thus the long branches within the clade. Similar argument was suggested by Hemp and Gennis for the evolution of the cytochrome oxidases. To explain the observed sequence divergence within archaea they have argued that archaeal membranes could select for different sequence characteristics than do bacterial membranes, which would lead to rapid divergence (Hemp and Gennis, 2008).

Furthermore, the presence of cytochrome *bc* complexes in archaea correlates in most cases (except for *Candidatus Korarchaeum cryptofilum* and *Thermoplasma acidophilum*) with the presence of the cytochrome oxidase genes, another likely LGT into Archaea (see Section 1.4.8 and (Hemp and Gennis, 2008)). As will be argued elsewhere, it appears that archaea have obtained their respiratory enzymes via LGT in a "package" which included not only the genes of energy-converting enzymes, but also the genes related to the biosynthesis of membrane components which enabled the usage of bacterial enzymes in archaeal membranes (D.V. Dibrova *et al.*, manuscript in preparation).

Hence, multiple operons of cytochrome *bc* complexes in many bacterial and archaeal genomes (**Tables 5.2-5.5**), as well as the affiliation of these operons with different clades in

**Figure 5.2**, indicate that cytochrome *bc* complexes are subjects to the LGT. One of the *bc*<sub>1</sub> operons in *Haloferax volcanii* DS2 (#5 in **Table 5.3**) is located on a plasmid, which points out the possible mechanism of the LGT of the cytochrome *bc* complexes.

In sum, the results of our phylogenomic analysis do not support the claim of vertical inheritance of the cytochrome *bc* complexes (Baymann *et al.*, 2012; Nitschke *et al.*, 2010; Schutz *et al.*, 2000). Apparently, these enzyme complexes underwent numerous lateral gene transfers between prokaryotic clades.

## **5.2. Discussion: implications from the lateral transfer of the cytochrome *bc* complexes for their evolution**

### **5.2.1. Discrepancies caused by the hypothesis on the presence of the cytochrome *bc* complexes in the LUCA**

The very presence of two distinct types of cytochrome *bc* complexes, namely of the *bc*<sub>1</sub>-type complexes and the *b<sub>6</sub>f*-type complexes, has prompted questions on the evolutionary relations between them as discussed in Section 1.5.3.3. Since initially only *bc*<sub>1</sub>-type complexes were found in archaea, it has been speculated that the LUCA, as the common ancestor of bacteria and archaea, already contained a cytochrome *bc* complex of the *bc*<sub>1</sub> type, whereas the *b<sub>6</sub>f*-type enzymes may have emerged in one of bacterial lineages after the separation of Bacteria from Archaea (Baymann *et al.*, 2012; Nitschke *et al.*, 2010; Schutz *et al.*, 2000). However, the presence of a certain enzyme in bacteria and archaea cannot be alone considered as an ultimate evidence for the presence of its ancestor in the LUCA because of the possibility of lateral gene transfer (LGT) between the domains (Koonin *et al.*, 2001). This possibility was considered, among others, by Cramer and co-workers (see Section 1.5.3.3 for more details). Later, however, the possibility of a significant LGT in the case of cytochrome *bc* complexes was ruled out based on a claim that the genes of cytochrome *bc* complexes are mostly vertically inherited, because their phylogeny, with few exceptions, corresponds to the phylogeny based on the rRNA sequences (Baymann *et al.*, 2012; Nitschke *et al.*, 2010; Schutz *et al.*, 2000).

Our analysis shows that this is not the case. As noted earlier, *Halobacteria* have obtained the cytochrome *bc* complexes from bacteria (Baymann *et al.*, 2012) and on two independent occasions, as they contain two distinct types of cytochromes *b* (**Table 5.3**). Euryarchaeal proteins from species belonging to the same class *Thermoplasmata* do not group together on the phylogenetic tree. The clade G in **Figure 5.2** is clearly formed by enzymes from various taxons, and several planctomycetes have more than one operon coding for the cytochrome *bc* complex in the genome (**Table 5.2**). In our analysis, as well as in previous studies, the high-order branches of the phylogenetic tree are not supported by confiding bootstrap values (Baymann *et al.*, 2012; Nitschke *et al.*, 2010) and thus the tree topology is unclear. The most striking dissimilarity between the tree of cytochromes *b* and trees of the acknowledged phylogenetic markers is in a surprisingly short branch separating the archaeal clade E from the other part of the tree: this could be observed both in **Figure 5.2** and in **Figure 5.3**. Similar short branch lengths were observed in previous studies (Baymann *et al.*, 2012) so that this is not a feature of our particular tree. The trees that were constructed for such acknowledged phylogenetic markers as 16S rRNA (Woese *et al.*, 1990), ribosomal proteins (Yutin *et al.*, 2008) or GroEL (Archibald *et al.*, 2000) (see **Figure 3.9** in Section 3.2.3 for our schematic representation of the phylogenetic tree for GroEL) clearly show the largest difference between the bacterial and archaeal groups of proteins. The absence of a clear separation between the archaeal and bacterial sequences is one more indication of the emergence of the former from the LGT.

In addition, the suggestion that the primordial cytochrome *bc* complex was a *bc*<sub>1</sub>-type enzyme and was present already in the LUCA, leads to a certain conundrum. First, according to the results of phylogenomic analysis, the LUCA could have had a Na<sup>+</sup>-dependent energetics, but not a H<sup>+</sup>-dependent energetics, which, apparently, has developed independently in bacteria and archaea after their separation from the LUCA (discussed in more detail in Section 1.3.2.4, see also (Mulkidjanian *et al.*, 2009; Mulkidjanian *et al.*, 2008b)). The cytochrome *bc* complex, which by its mechanism is strictly a proton translocator, should have been of little use in the context of Na<sup>+</sup>-dependent energetics. Second, the cytochrome *bc* complexes require high-potential electron acceptors for proper functioning. Before the "invention" of oxygenic photosynthesis by the ancestors of cyanobacteria, the Earth environments were highly reduced and there should have been no

abiotic electron acceptors for the cytochrome *bc* complexes (Williams and Frausto da Silva, 2006).

Hence, on the one hand, the presence of a *bc*<sub>1</sub>-type complex in the LUCA is unlikely because there was no apparent functional niche for a cytochrome *bc* complex at the stage of the LUCA. On the other hand, the results of phylogenomic analysis are compatible with the lateral transfer of the *bc*<sub>1</sub>-type and *b*<sub>6f</sub>-type complexes from bacteria to archaea, specifically to those that possess enzymes of fatty acid metabolism (see Section 5.1).

### **5.2.2. Fusion of the cytochrome *b* subunits is a more probable evolutionary scenario for cytochrome *bc* complexes than their fission**

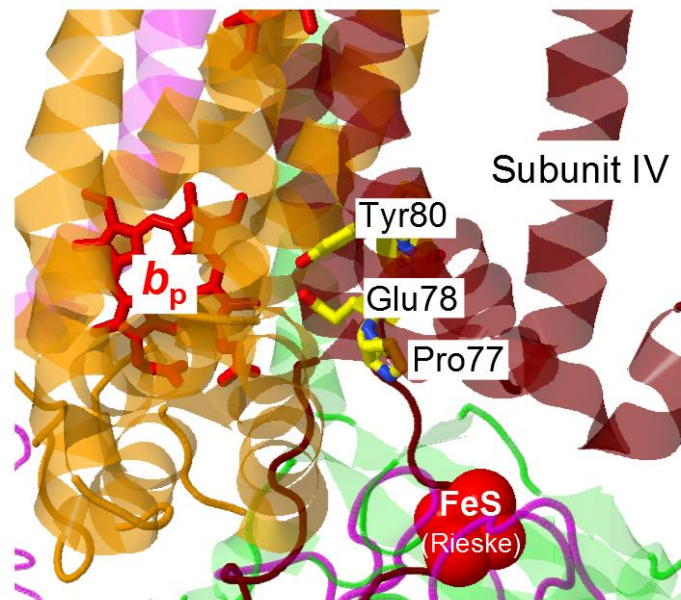
An important question in the evolutionary scenario for the cytochrome *bc* complex is whether its ancestral form could have a long cytochrome *b* as modern cytochrome *bc*<sub>1</sub>-complexes of *Proteobacteria*, *Chlorobi*, mitochondria and other species or a shorter cytochrome *b* with subunit IV as cyanobacterial *b*<sub>6f</sub>-complexes.

The phylogenetic tree in **Figure 5.2** contains several clades of the *bc*<sub>1</sub>-type complexes with "long" cytochromes *b* and four clades of the *b*<sub>6f</sub>-type complexes, namely the groups A, B, C and G. Hence, we should choose between two possibilities: **(1)** the fusion scenario under which the ancestral version of the cytochrome *bc* complex was a *b*<sub>6f</sub>-type complex with a split cytochrome *b*, so that the fusion of a small cytochrome *b* with subunit IV took place several times in different lineages and **(2)** the fission scenario, where the ancestral version contained a long cytochrome *b* that has split at least twice in the course of evolution. The fusion scenario seems more probable than the fission scenario because of several reasons:

1. The fission scenario implies independent splits of the cytochrome *b* gene at the same point in several lineages, followed by independent appearances of the conserved, heme *c*<sub>n</sub>-binding motif CxGG in exactly the same positions within these lineages. These events seem extremely unlikely.
2. The superposition of the structures of the *bc*<sub>1</sub>-type and *b*<sub>6f</sub>-type complexes shows that the Q<sub>N</sub> quinone in *bc*<sub>1</sub> is bound in the position that is occupied by heme *c*<sub>n</sub> in the *b*<sub>6f</sub>-type complex (Stroebel *et al.*, 2003). Generally, a loss of a porphyrin ring and its

- functional replacement by a smaller quinone ring is much easier to envision than an insertion of a porphyrin ring into a cavity that had been occupied by a quinone ring (Mulkiđjanian and Junge, 1997). Larger unit (porphyrin) cannot be inserted into originally smaller cavity without major disturbing of the structure of the protein.
3. The clade-specific conservation of the linkers between the cytochrome *b*<sub>6</sub>-like and subunit IV-like parts of full-length cytochromes *b* and the absence of any conservation in this region between different clades (**Figure 5.2**) are consistent with the hypothesis of several independent fusions. If the long version of cytochrome *b* were present in the common ancestor of the cytochrome *bc* complexes, one would observe a similar conservation pattern in this region in all branches of the tree (or an absence of any sequence conservation). Archaeal sequences from clade E do not show any conservation in this region, and *Thaumarchaeota* (*Caldiararchaeum subterraneum* and *Nitrosopumilis maritimus*) even have a truncation in this region.
  4. The number of residues between the two heme-binding histidines in the helix 4 of cytochrome *b* could be either 13 or 14. Early argument in favor of the long cytochrome *b* as ancestral version was based on the observation that archaeal and mitochondrial (i.e. proteobacterial) sequences both have 13 residues in the helix 4 while the *b<sub>6</sub>f* complexes of cyanobacteria with 14 residues between the histidine residues were considered an exception (Widger *et al.*, 1984). However, with more genomes sequenced, it now appears that the vast majority of the cytochrome *b* sequences actually have 14 residues between these histidines. This fact has been already noted in (Baymann *et al.*, 2012) but no evolutionary consequences were drawn. This trait speaks against evolutionary oldness of archaeal *bc*<sub>1</sub> complexes and might indicate their acquisition via the LGT from bacteria.
  5. The components of cytochrome *bc* complexes tend to fuse together. These fusions are particularly widespread among Gram-positive bacteria (Yu *et al.*, 1995); the organisms in the clade H provide further examples of fused cytochromes *b*, see **Table 5.4**. Fusions, generally are more popular in evolution than fissions (Enright *et al.*, 1999; Marcotte *et al.*, 1999). Generally, the fusion between two interacting proteins can sufficiently decrease  $\Delta S$  of dissociation of protein complex, thus decreasing the free energy of association (Marcotte *et al.*, 1999). This reduction in entropy is often

expressed as an increase in the effective concentration of one interacting protein to another, thus fusions between interacting proteins are thermodynamically favourable. Fusion between previously non-interacting proteins could occur through the mutations in the stop-codon of one protein, thus, in the absence of frameshift, the long protein coded by both genes would be produced (Hawkins and Lamb, 1995). A "Rosetta Stone" hypothesis was suggested to predict functional cooperation between proteins that are found as domains of the multisubunit protein in some genomes (Marcotte *et al.*, 1999). In principle, this hypothesis takes into account possible gene fissions, with fissioned proteins maintaining functions of the separate domains. But for the cytochrome *bc* complex the fission scenario implies a fission within the key catalytic site (*Figure 5.5*).



**Figure 5.5. Residues from the conserved PEWY motif directly face the *b<sub>p</sub>* heme.**

The catalytic *P*-site in the structure of cytochrome *b<sub>6</sub>f* complex from *Nostoc sp.* PCC 7120 (PDB ID 4H44 (Hasan *et al.*, 2013)) is shown.

Subunit IV in cytochrome *b<sub>6</sub>f* complexes of cyanobacterial and plants contains a conserved motif PEWY, in which the Glu residue was suggested to be involved in the abstraction of the proton from the semiquinone radical in the *P*-side (Widger *et al.*, 1984). Recently structural data have confirmed this; in the structure of cytochrome *b<sub>6</sub>f* complex from *Nostoc* the PEWY motif faces the heme *b<sub>p</sub>*, as shown on *Figure 5.5*



(Hasan *et al.*, 2013). We are not aware of any example of a protein fission through the catalytic site in the course of evolution, and therefore the fission scenario seems highly unlikely.

### 5.2.3. Possible scenario of the emergence of cytochrome *bc* complexes

In the currently popular scenario of the evolution of the cytochrome *bc* complexes, the ancestral *bc*<sub>1</sub>-type form of the complex has been suggested to emerge at the stage of the LUCA from an interaction between a Rieske-type iron sulfur protein and a large cytochrome *b* (Baymann *et al.*, 2003; Baymann *et al.*, 2012; Nitschke *et al.*, 2010; Schutz *et al.*, 2000). The same authors have suggested that the primordial "construction kit" of protein "modules" has contained two different cytochrome *b* modules, namely a four-helical model to be used in dehydrogenases/oxidoreductases and a large, 8-helical module to be used only in the ancestral cytochrome *bc* complex (Baymann *et al.*, 2003). This suggestion, however, makes this set redundant, and thus the origin of the 8-helical cytochrome *b* and, more importantly, its possible function in the LUCA before being recruited into the cytochrome *bc*<sub>1</sub> complex has remained enigmatic in this scenario.

The evolutionary primacy of the *b<sub>6f</sub>*-type complexes, as suggested in the current work, implies that the ancestral form of the cytochrome *bc* complex contained a four-helical cytochrome *b* (see **Figure 5.6**). This assumption helps to reduce redundancy in the "construction kit" (Baymann *et al.*, 2003). A bundle formed of 4 alpha-helices represents one of the widespread protein folds; this is one of the few folds which are found both in water-soluble and in membrane proteins (Neumann *et al.*, 2010). In the SCOP database (Murzin *et al.*, 1995) the fold "heme binding four helical bundle" comprises three superfamilies; thereby the four-helix cytochrome *b* of the cytochrome *bc* complex, together with the membrane cytochrome of the formate dehydrogenase make the superfamily of "transmembrane di-heme cytochromes". Binding of two hemes has been shown to stabilize the fold (Choma *et al.*, 1994); it has been shown that the half of the *de novo* four-helical proteins from designed combinatorial libraries could bind the heme (Rojas *et al.*, 1997). Membrane cytochromes with such a fold usually serve as membrane anchors for large, protruding subunits where a distal substrate-binding site is connected by an electron-transferring "wire" of iron sulfur

clusters with the membrane, as e.g. in formate dehydrogenase (Jormakka *et al.*, 2003) or [NiFe] hydrogenase (Pandelia *et al.*, 2012).

In some of such enzymes the protruding parts are facing the exterior of the prokaryotic cell (formate dehydrogenase, Ni-Fe hydrogenase), while in others they look into the cytoplasm (e.g. fumarate reductase). Functionally, the enzymes that protrude out of the cell interact with simple electron donors, such as formate and hydrogen, and reduce quinones within the membrane, while the enzymes which protrude into the cytoplasm connect the membrane quinone pool with cellular metabolites such as e.g. succinate or fumarate. Acting together, such enzymes accomplish a quinone-mediated translocation of reducing equivalents across the cellular membrane. Depending on the metabolic situation, the cell could benefit from either a sink for excess electrons or from electrons for biosynthesis, so that a system of reversible, differently oriented dehydrogenases/oxidoreductases could catalyze electron transfer in both directions. The presence of such enzymes in the LUCA – in a capacity of electron carriers across the membrane - is rather easy to imagine. Primitive membranes could already represent a significant barrier for reducing equivalents (electrons). By invoking large porphyrin rings as electron carriers, the desolvation penalty for electrons could be decreased and the transfer of electrons across the membrane could be accelerated, see e.g. (Krishtalik, 1996). Further acceleration could be achieved by translocating a proton together with an electron; such phenomenon, the mechanism of which is not quite clear, has been described for the menaquinol:fumarate reductase of *Wolinella succinogenes* (Madej *et al.*, 2009; Madej *et al.*, 2006). The recruitment of two hemes, which seems to happen independently in several protein families, enables, by providing two electron vacancies, the electronic coupling with quinols, which are two-electron carriers. After the cell membranes became proton tight (Mulkidjanian *et al.*, 2009), a joint action of several, differently oriented membrane dehydrogenases/oxidoreductases, which released protons into periplasm upon oxidation of electron donors such as hydrogen and formate, and trapped cytosolic protons upon reducing the biochemical substrates such as e.g. fumarate, could pave the way to proton-dependent bioenergetics by completing Mitchell's proton loop, as suggested Jormakka and co-workers (Jormakka *et al.*, 2003).

It is tempting to speculate that the transition from a membrane electron translocase to a primordial cytochrome *bc* complex could be driven by the appearance of high-potential

electron acceptors. Since the redox potential of the environment was low before the oxygenation of earth some 2.5 Gyr ago (Hazen *et al.*, 2011), potential electron acceptors for cytochrome *bc* complexes should have been absent from the environment (Williams and Frausto da Silva, 2006). Under these conditions, high-potential electron acceptors could be, however, produced upon photosynthesis (A. Bogachev, personal communication). Indeed, the essence of (bacterio)chlorophyll-based photosynthesis is using the energy of light quanta for separating electric charges at the so called "special" pair of (bacterio)chlorophyll molecules within photochemical reaction center (PRC), see Section 1.5.4 and (Crofts and Wraight, 1983; Govindjee *et al.*, 1982) for reviews. As a result of such separation, an electron is removed from the special pair to reduce low-potential acceptors, such as NAD(P)<sup>+</sup> or quinones, while a high-potential electron vacancy (hole) remains at the (bacterio)chlorophyll moiety. In most modern phototrophic organisms, the *bc*<sub>1</sub>-type and *b<sub>6</sub>f*-type complexes are involved in re-reducing these oxidized (bacterio)chlorophyll molecules. It is tempting to speculate that this function could have been the initial function of the first menaquinol-oxidizing cytochrome *bc* complexes of the *b<sub>6</sub>f* type. This suggestion is consistent with the evolutionary primacy of the *b<sub>6</sub>f*-type complexes, as inferred from phylogenomic analysis (see Section 5.2.2), and with the affiliation of many *b<sub>6</sub>f*-type complexes with photosynthetic reaction centers (Nitschke *et al.*, 2010). The emergence of first cytochrome *bc* complexes within phototrophic membranes can also explain the involvement of a chlorophyll molecule and a carotenoid molecule as structural elements of the *b<sub>6</sub>f*-type complexes of green plants and cyanobacteria (Kurisu *et al.*, 2003; Stroebel *et al.*, 2003). While it is difficult to envision an insertion of large chlorophyll and carotenoid molecules into a pre-formed, tightly folded membrane protein, their recruitment upon the very formation of the protein seems quite plausible.

As shown in **Figure 5.6**, the ancestor of the first *b<sub>6</sub>f*-type complex could have been a membrane oxidoreductase that possibly interacted with NAD(P)H, with its membrane subunit belonging to the "transmembrane di-heme cytochrome" fold. It is noteworthy that unlike the *bc*<sub>1</sub>-type complexes, the *b<sub>6</sub>f*-type complexes seem to be functionally coupled to oxidoreductases. Specifically, the ferredoxin:NADP<sup>+</sup> oxidoreductase (FNR) is a functional counterpart of the plant cytochrome *b<sub>6</sub>f*-type complex (Baniulis *et al.*, 2011; Iwai *et al.*, 2010; Szymanska *et al.*, 2011; Zhang *et al.*, 2001), whereas the *b<sub>6</sub>f*-type complexes which belong to

the clade G co-occur with a gene that codes for a large, NAD(P)-binding oxidoreductase subunit and, in some cases, have this gene within the operon of their *b<sub>6</sub>f*-type complexes.

The transition from a membrane oxidoreductase to the cytochrome *bc* complex should have included recruitment of a three-helix protein (future subunit IV) and a Rieske protein. The three-helical subunit IV, the evolutionary origin of which could not so far be traced, provided the quinol-binding P[DE]W[FY] motif enabling the bifurcated oxidation of quinol. Since the subunit IV is also involved in binding of the *c<sub>n</sub>* heme (Kurisu *et al.*, 2003; Stroebel *et al.*, 2003), the recruitment of both may have occurred simultaneously. The recruitment of a mobile FeS domain-possessing Rieske protein, which is present also in other enzymes, e.g. arsenite oxidase (Lebrun *et al.*, 2006), should have enabled the delivery of electrons to the high-potential electron vacancies at primordial photochemical reaction centers.

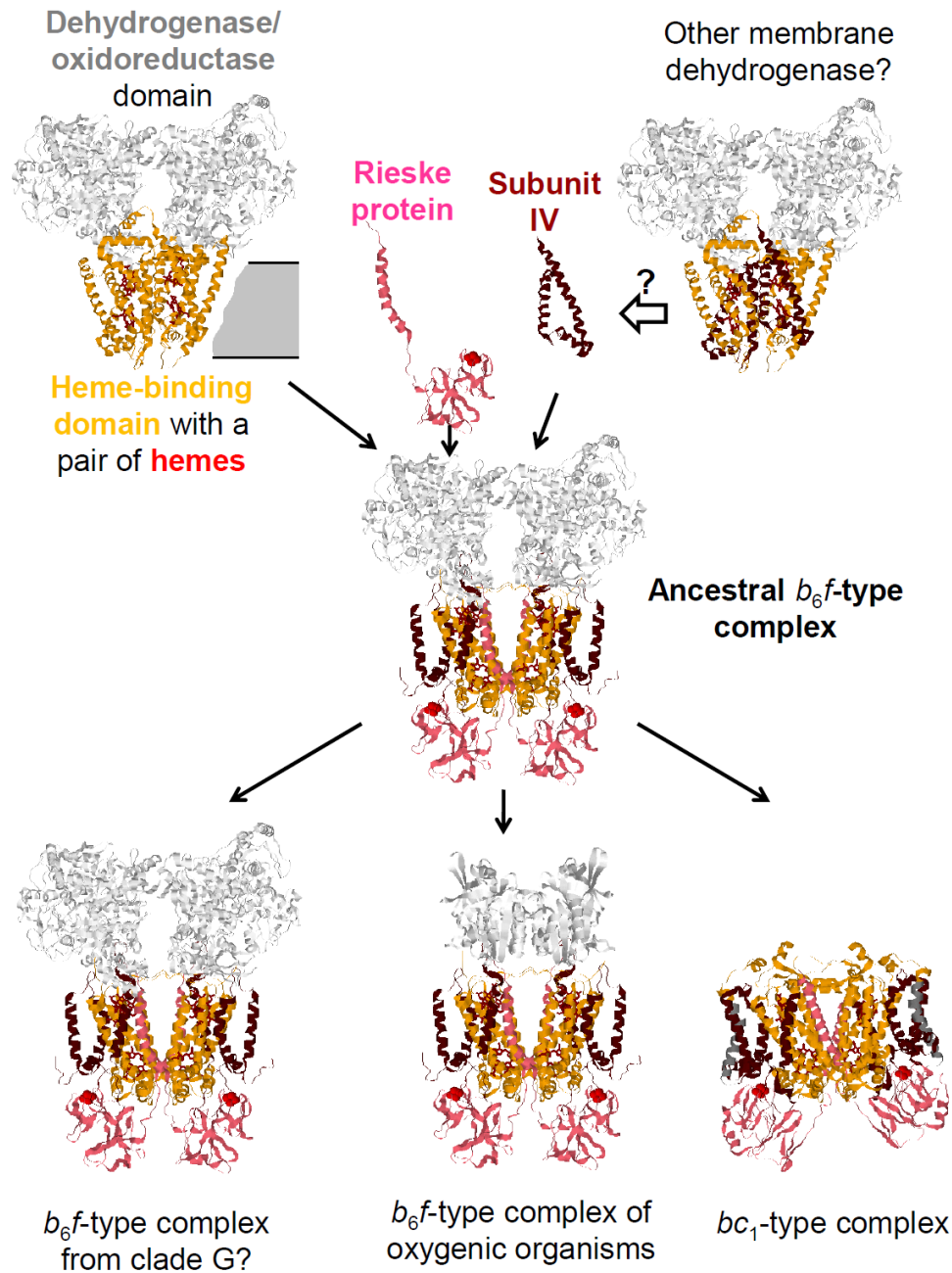
Hence, as shown in **Figure 5.6**, the ancestral form of the cytochrome *bc* complex could structurally and functionally resemble the *b<sub>6</sub>f*-type complexes of anaerobic organisms such as still unexplored enzymes from clade G organisms or the *b<sub>6</sub>f*-type complex from the heliobacterial clade B in **Figure 5.2**. Specifically, the photosynthetic apparatus of *Heliobacillus mobilis* and *Heliobacterium modesticaldum* are harbored on large operons (Sattley *et al.*, 2008; Xiong *et al.*, 1998), potentially capable of LGT. Since *Heliobacteriaceae* are the only phototrophs among *Firmicutes* (Gram-positive bacteria), it was argued that heliobacteria most likely obtained their photosynthetic genes via the LGT from now extinct phototrophic, anoxygenic ancestors of cyanobacteria (Mulkidjanian and Galperin, 2013; Mulkidjanian *et al.*, 2006). However, while the cyanobacteria should have undergone dramatic changes in response to the oxygenation (Mulkidjanian *et al.*, 2006; Raymond and Blankenship, 2004; Rutherford *et al.*, 2012), which they could not evade, the strictly anaerobic *Heliobacteria* retained not only the ancestral version of the homodimeric photosynthetic RC, but apparently, an ancestral version of the *b<sub>6</sub>f*-type complex, which is coded by the same operon as the ancient PRC (Xiong *et al.*, 1998). It is noteworthy that a separate operon in the genome of *Heliobacterium modesticaldum* codes for a tandem of a large NAD(P)-binding oxidoreductase, which is found within the operons of the *b<sub>6</sub>f*-type complexes of the clade G (see **Table 5.3**), and a ferredoxin-NADP<sup>+</sup> reductase.

Based on the available data on the properties of the *b<sub>6</sub>f*-type complex of *Heliobacteria* (Baymann and Nitschke, 2010; Yue *et al.*, 2012) and on the sequence data for the *b<sub>6</sub>f*-type

complexes of clade G, it is possible to infer that the ancestral cytochrome *bc* complex should have possessed, in addition to a split, two-subunit cytochrome *b*, a low-potential heme  $c_n$  with a  $E_m$  value of  $\sim -100$  mV, a low-potential version of the Rieske FeS cluster with a  $E_m$  value of  $\sim 150$  mV, a multiheme cytochrome *c* as an acceptor of electrons from the Rieske protein, and, perhaps, a further FeS-cluster(s)-containing subunit (NADPH oxidoreductase?) localized on the cytoplasmic, *n*-side of the membrane. Most likely, the ancestral  $b_6f$ -type complex had a conserved P[DE]W[FY] motif in its subunit IV. This motif is found in clades A, B, and C, as well as in the majority of long cytochromes *b*, see **Figure 5.2**. As it also follows from this figure, the cytochrome *b* of the ancestral enzyme most likely contained seven transmembrane helices, four of cytochrome *b* and three of subunit IV. The two additional helices in the  $b_6f$ -type complexes from clade G do not show significant sequence similarity with the additional, eighth helix of long cytochromes *b* of the  $bc_1$ -type complexes and seem to be later independent acquisitions.

#### **5.2.4. Adaptation of the cytochrome *bc* complexes to oxygenation of the atmosphere**

The menaquinol-oxidizing  $b_6f$ -type complexes of modern anaerobes differ substantially both from the  $bc_1$ -type complexes of aerobic organisms and from the  $b_6f$ -type complexes of oxygenic plants and cyanobacteria. It has been argued that the appearance of oxygen in the atmosphere some 2.5 Gyr ago, followed by the replacement of the low-potential menaquinone by the high-potential ubiquinone in several proteobacterial clades and by plastoquinone in cyanobacteria, should have prompted major modifications in the cytochrome *bc* complexes of aerobic organisms (Baymann *et al.*, 2012; Schoepp-Cothenet *et al.*, 2009). First, the  $E_m$  values of the redox components involved should have been adjusted to the  $\sim 150$  mV increase in the redox potential of the pool quinone. Second, the electron escape from the redox components of the cytochrome *bc* complexes to oxygen (leading to the formation of the potentially deleterious ROS) should have been prevented, which could be achieved by minimizing the number of auto-oxidizable components in the electron transport chain.



**Figure 5.6. Evolutionary scenario for cytochrome  $bc$  complexes.**

The cytochrome  $b_6$ -like parts (the 4-helical bundle) are colored orange, the subunit IV-like parts are colored dark red, and the Rieske proteins are colored pink. The three-helix subunit IV is arbitrarily suggested to be recruited from a membrane dehydrogenase.

The modifications of cytochrome  $bc$  complex could have involved independent replacement of the low-potential multiheme cytochrome (which is present in the  $b_6f$ -type complexes of anaerobes (Baymann and Nitschke, 2010; Baymann *et al.*, 2012; Blankenship, 1992;

Nitschke *et al.*, 2010)) with the cytochrome *f* with a high-potential, not auto-oxidizable heme in cyanobacteria and with a single-heme cytochrome  $c_1$  in the  $bc_1$ -type complexes of certain proteobacteria. Evolutional origin of cytochrome *f* remains enigmatic (Al-Attar and de Vries, 2013; Cramer *et al.*, 2006), but the origin of cytochrome  $c_1$  could be traced to a two-heme  $c_4$ -type proteobacterial cytochrome (Baymann *et al.*, 2004). As argued elsewhere (Dibrova *et al.*, 2013), either increase of the  $E_m$  value for the heme  $c_n$  (in cytochrome  $b_6f$ -complex involved in oxygenic photosynthesis) or its complete loss in the cytochrome  $bc_1$  complex of aerobic bacteria could have allowed to prevent its oxidation under aerobic conditions. Altogether these modifications could have been performed to diminish the formation of ROS by cytochrome *bc* complexes. In the next chapter we will discuss further possible adaptations of the system to aerobic conditions on the example of mitochondrial cytochrome  $bc_1$ -complex.

## 6. Evolution of apoptosis as a strategy to diminish the oxygen-caused damage to consortia of cells

### 6.1. Introduction: Apoptosis as a mechanism to kill a cell with a broken mitochondrial cytochrome *bc*<sub>1</sub>-complex

As discussed in Section 2.2.4, the evolution of the cytochrome *bc* complexes both in aerobic prokaryotes and in oxygenic phototrophs was accompanied by deactivating the potential sources of ROS (sites where oxygen molecules could undergo single-electron reduction). Still, one source of ROS could not be deactivated completely. The oxidation of a quinol molecule in the centre *P* of all cytochrome *bc* complexes is accompanied by a transient formation of a low-potential unstable ubisemiquinone that quickly reduces the low-potential heme *b*<sub>p</sub>, but also can reduce oxygen, see **Figure 1.5.22** and (Mulkidjanian, 2005; Rutherford *et al.*, 2012) for reviews.

As discussed in Section 1.5.5 in more detail, mitochondria, the descendants of endosymbiotically obtained  $\alpha$ -proteobacteria, were shown to serve as triggers of the so-called intrinsic apoptotic pathway (Petit *et al.*, 1996; Skulachev, 1996; Wang and Youle, 2009; Zamzami *et al.*, 1996). Since mitochondria could start producing ROS under certain conditions, specifically in the cytochrome *bc*<sub>1</sub> complex (see Section 1.5.5 for details), the ROS-induced damage to other cells could be diminished by triggering the apoptosis as soon as possible, i.e. even *before* the disruption of the affected mitochondria. This is achieved by a signal amplification cascade within mitochondria, as depicted in **Figure 1.5.23** and discussed in section 1.5.5. It has been shown that molecules of a four-tail lipid cardiolipin (CL) that is particularly susceptible to the ROS-induced peroxidation, when oxidized, can interact with the molecules of cytochrome *c* at the outer surface of the inner mitochondrial membrane and cause conformational change in the latter (Huttemann *et al.*, 2011; Kagan *et al.*, 2009). The affected molecules of cytochrome *c* attain peroxidase activity and start to produce additional ROS (including singlet oxygen). This, apparently, accelerates the formation of a pore in the outer mitochondrial membrane and the release of cytochrome *c* into the cytoplasm (Huttemann *et al.*, 2011; Kagan *et al.*, 2009; Miyamoto *et al.*, 2012).



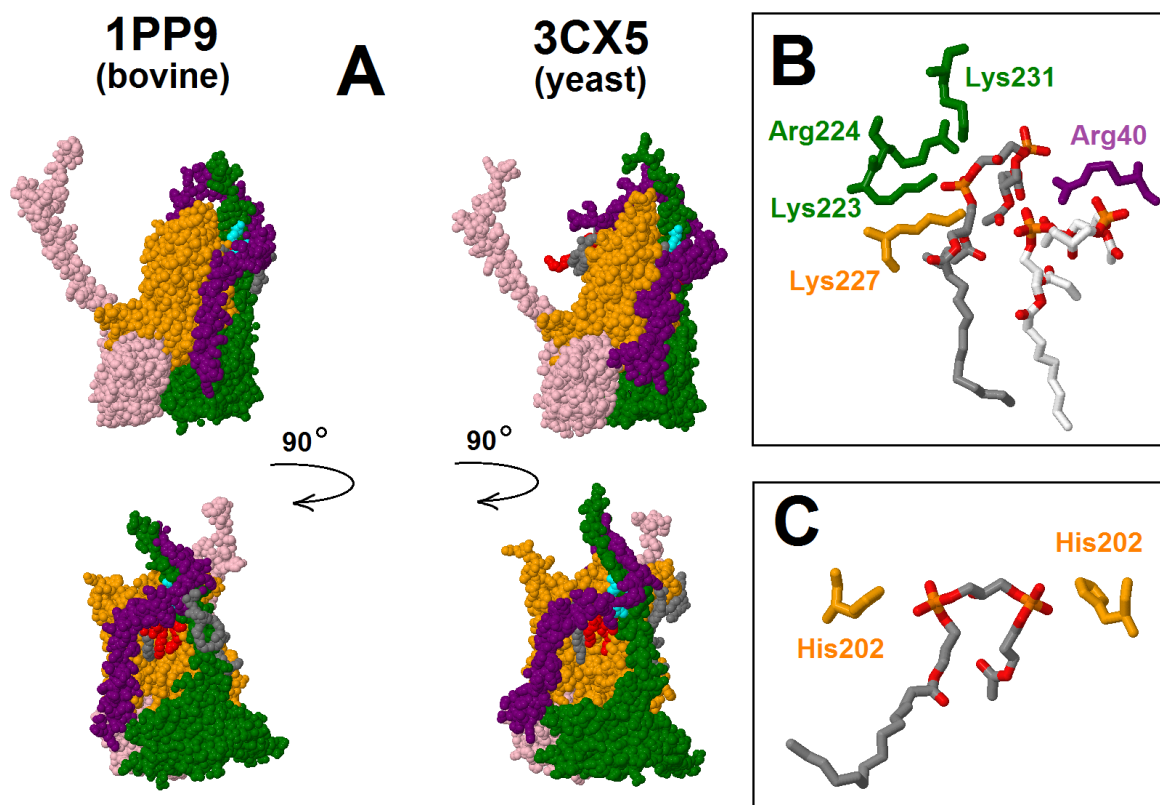
Obviously, the fastest way to accelerate the apoptosis is to use a ROS-generating cytochrome  $bc_1$  complex proper as a trigger of a signal amplification cascade that involves oxidized CL molecules. Apparently, exactly this strategy is realized in vertebrates, as discussed in section 1.5.5. In this chapter, we have checked how the elements of the signal amplification cascade have changed in course of evolution from prokaryotes to eukaryotes; the results obtained have been published in (Dibrova *et al.*, 2013).

## 6.2. Ligands for cardiolipin, a possible early detector of the ROS production, in the components of the cytochrome $bc_1$ complex

Cardiolipin molecules have highly negatively charged lipid heads compared to other types of lipids as they contain two phosphate groups. They are bound to the  $bc_1$  complex by positively charged residues. It is noteworthy that in all inspected CL-containing structures, with the exception of bovine cytochrome  $bc_1$  complex (complex III), CL is bound at the protein/membrane interface. In the bovine complex III, however, the CL binding site is deeply buried within the protein (**Figure 6.2A**, bottom). Specifically, in the 1PP9 structure (Huang *et al.*, 2005) the site contains two CL molecules and one molecule of phosphatidylcholine. The protein completely encases the lipid molecules like a belt. On the contrary, in the structure of yeast enzyme (**Figure 6.2B**, bottom), there is an opening in the "belt" which should permit fast exchange of cardiolipin molecules with those in the lipid phase. We have checked how the cardiolipin-binding residues in the cytochrome  $bc_1$  complex have changed in course of evolution from prokaryotes to eukaryotes.

**Ligands of cardiolipin that are provided by cytochrome  $b$ .** The ligands of the cardiolipin molecule that is located between the monomers are positioned at the end of the transmembrane helix 4 of cytochrome  $b$  (**Figure 6.3D**, blue line). The side chains of two histidine residues from both monomers are positioned near the phosphate groups of cardiolipin (**Figure 6.2C**). This His residue (#202 in yeast, #201 in bovine sequence) is absolutely conserved in eukaryotic complex III and is mostly conserved in cytochrome  $bc_1$  complexes from the proteobacterial branch I (**Figure 5.2**). However, the sequences from *Campylobacter curvus*, *Wolinella succinogenes* and *Aquifex aeolicus* have arginine in this position. Arginine is also present in cyanobacterial, heliobacterial and plant  $b_6f$ -complexes as

well as in most sequences from clade G and some sequences from clade D (*Figure 6.4*). Other sequences mostly have non-polar replacement in this position.

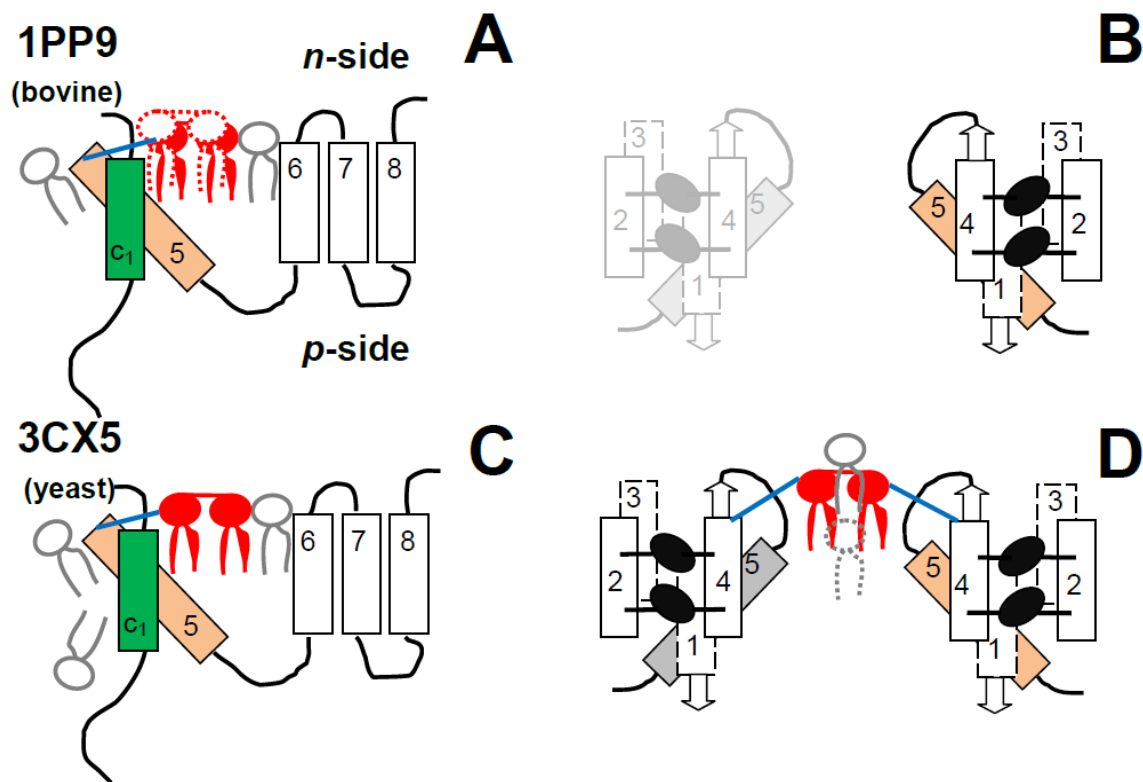


**Figure 6.2.** Cardiolipin molecules in the bovine complex III (PDB ID 1PP9) and the yeast complex III (PDB ID 3CX5).

(A) Only several chains of one monomer are shown for each complex, namely the cytochrome *b* (orange), the iron-sulfur Rieske protein (pink), the cytochrome *c*<sub>1</sub> (green), and the 9.5 kDa subunit (purple). Cardiolipin molecules or their analogues are colored red, other lipids are colored grey. Positively charged residues in contact with the cardiolipin molecule are colored cyan.

(B) Ligands of the cardiolipin dimer in the bovine structure. The colors of the depicted residues correspond to the colors of respective protein chains.

(C) Ligands of the cardiolipin molecule that is located between the two monomers in the yeast structure.

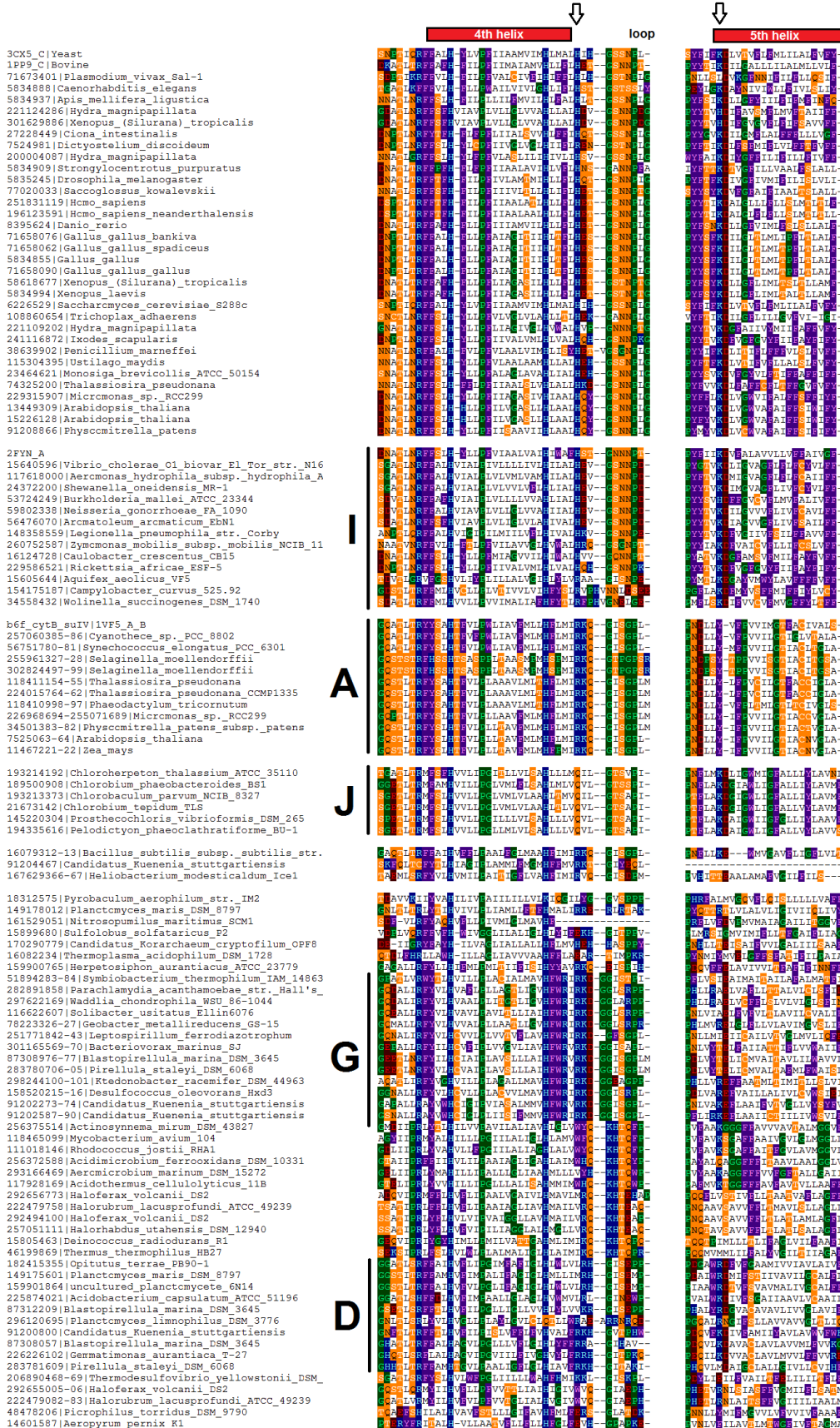


**Figure 6.3.** Scheme of the binding sites of cardiolipin molecules as provided by cytochrome *b* and cytochrome *c*<sub>1</sub>.

Binding sites of cardiolipin (red four-tailed entity) on the flanks of the cytochrome *bc*<sub>1</sub> complex (A), (C) and in the middle of a cytochrome *bc*<sub>1</sub> complex dimer (B), (D). The helices of cytochrome *b* are numbered. Hemes are shown as black circles. The arrows at the helices 1 and 4 show the direction from the N-terminus to the C-terminus of the protein.

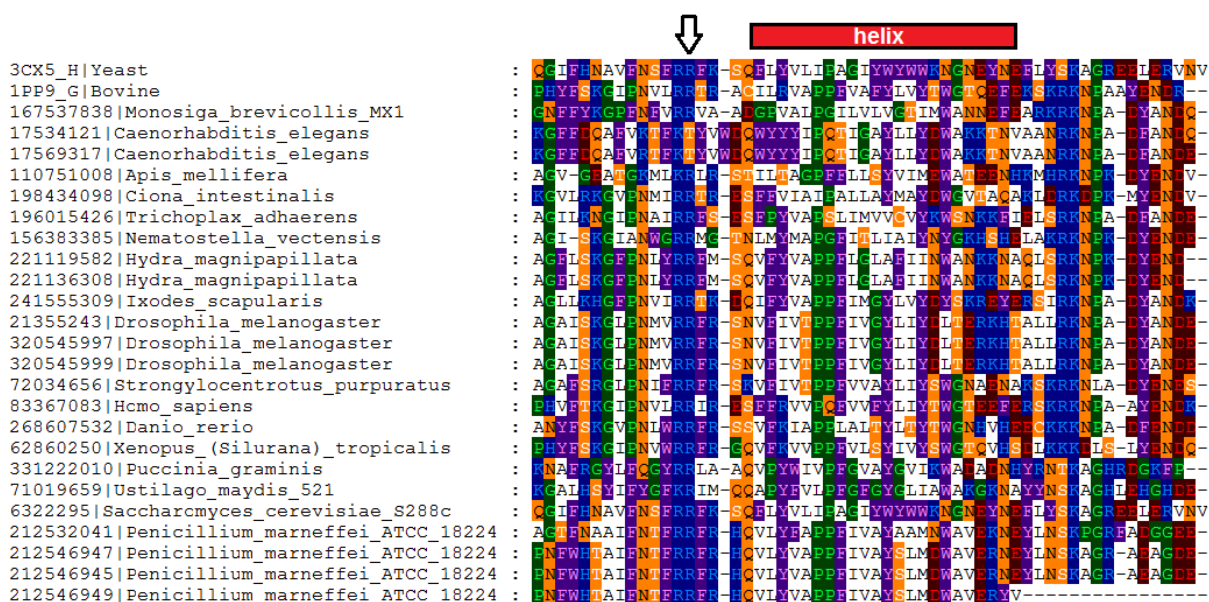
**Figure 6.4 (on the next page).** Part of multiple alignment of the cytochromes *b* in the region of helices 4 and 5.

Names of selected clades of prokaryotic sequences are given according to **Figure 5.2** in the main text. Arrows mark positions of the cardiolipin ligands.



The ligand of the second cardiolipin is positioned in the beginning of the transmembrane helix 5 of cytochrome *b* (**Figure 6.3A** and **C**, blue line). This Lys residue (#228 in yeast, #227 in bovine) in the complex III of eukaryotes is mostly conserved (with the exception of *Plasmodium vivax*). It is the same in *Proteobacteria* and *Chlorobi*. Actinobacterial sequences contain either lysine or arginine in this position, however, other members of the clade F have hydrophobic replacements. Two unclustered halobacterial sequences as well as most clade D and some clade G sequences also have either Lys or Arg.

**Ligand of cardiolipin as provided by the subunit 8 (9.5 kDa subunit).** This subunit (termed H in yeast and G in mammals) is specific for the eukaryotic complex III. Arginine residue (#40 in bovine) is conserved in all eukaryotic enzymes with the exception of *Caenorhabditis* species (**Figure 6.5**).

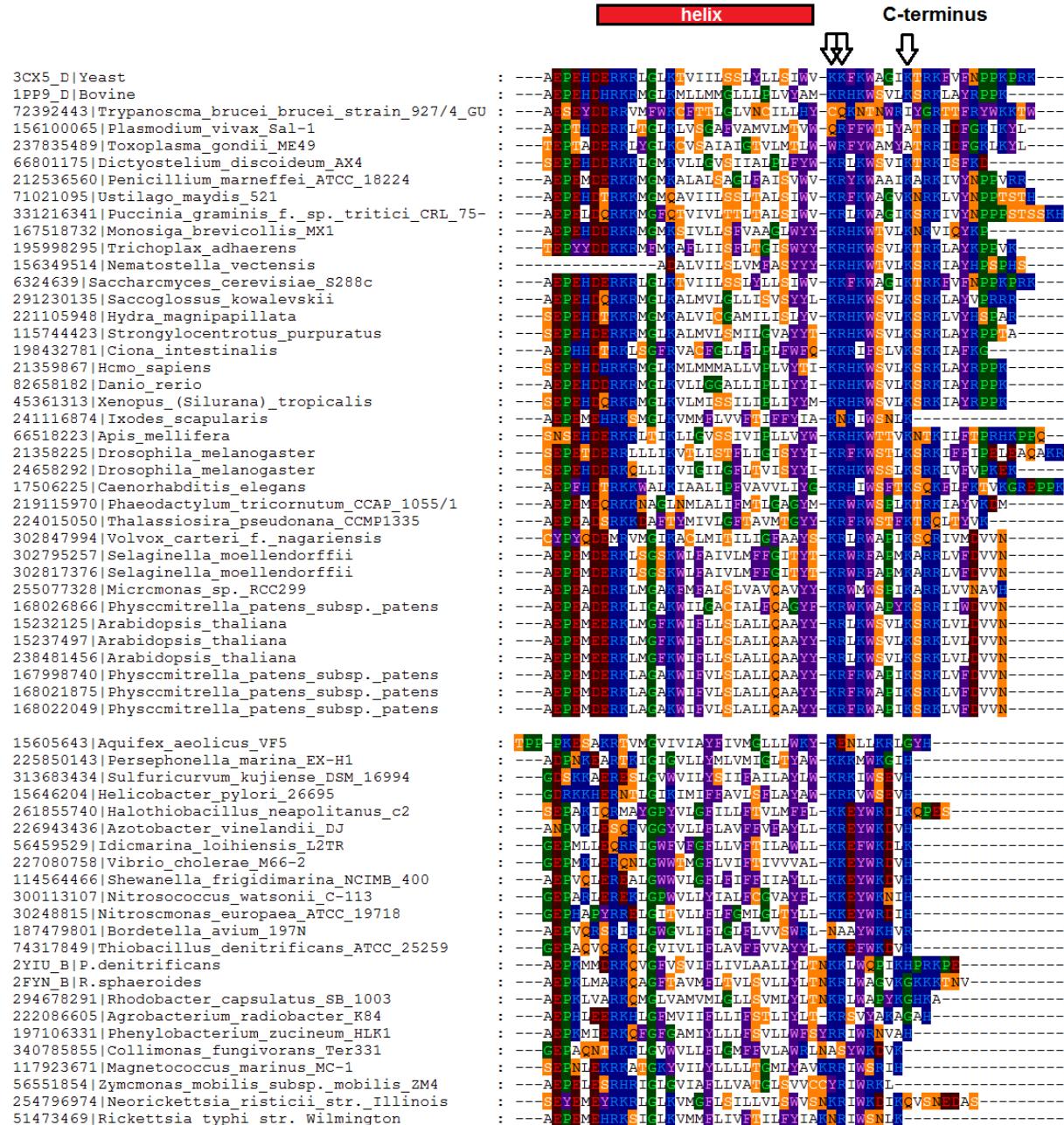


**Figure 6.5.** Part of multiple alignment of subunits 8 from eukaryotes.

The arrow shows the position of the cardiolipin ligand.

**Ligands of cardiolipin as provided by cytochrome *c*<sub>1</sub>.** Three residues of cytochrome *c*<sub>1</sub> are positioned near the phosphates of cardiolipin dimer. They are located at the end of the C-terminal helix (**Figure 6.6**). The C-terminal region in eukaryotes is longer; all three residues are mostly conserved, and only in parasitic species as *Trypanosoma brucei*, *Plasmodium vivax* and *Toxoplasma gondii* the motif is absent. Bacterial cytochromes *c*<sub>1</sub> mostly preserve

the pair of Lys-Arg (#223-224 in beef). Since bacterial sequences are shorter, they do not have the third residue (Lys231 in beef), but have a number of positively charged residues in this region.



**Figure 6.6. Multiple alignment of C-terminal parts of cytochromes *c*<sub>1</sub>.**

The eukaryotic sequences are shown in the top, the prokaryotic sequences are shown on the bottom. The arrows mark positions of the cardiolipin ligands.

Our structural analysis of the residues that bind the CL molecules within the cytochrome *bc*<sub>1</sub> complex showed that the number of charged residues, which bind the phosphate groups of the CL molecules, has increased upon the evolution from the *b<sub>6</sub>f*-type complexes, via the *bc*<sub>1</sub>-type complexes of  $\alpha$ -proteobacteria, the predecessors of mitochondria, to mitochondrial *bc*<sub>1</sub>-type complexes. Thus, upon the evolution from bacteria to vertebrates, a very special CL-binding sites has evolved within the cytochrome *bc*<sub>1</sub> complex where the CL molecules are tightly bound close to the major source of ROS.

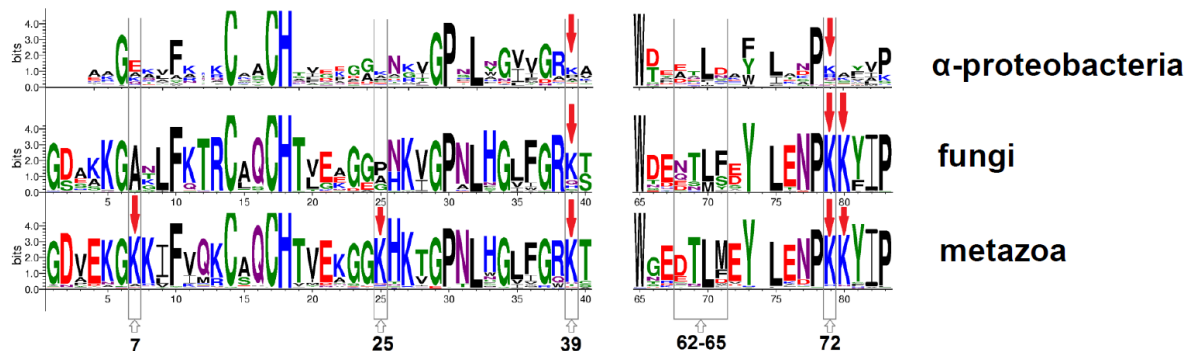
### **6.3. Evolution of the interaction between cytochrome *c* and the components of the apoptosome**

A comparative analysis of intrinsic apoptotic pathways in different multicellular organisms (Wang and Youle, 2009) shows that they have some common properties but also some differences. The common feature is that the apoptotic machinery can be triggered by the interaction of one of its components with a mitochondria-located protein. Under normal conditions, mitochondrial proteins stay within mitochondria, and their appearance in the cytoplasm reports to the cell that the mitochondrion has been broken and the time has come to commit a suicide. The nature of mitochondrial proteins that trigger the apoptosis, however, is different in different organisms. It was shown that in vertebrates the apoptotic cascade in the cytosol is triggered by the release of cytochrome *c* from mitochondria and its interaction with the tryptophane and aspartate-rich, WD (also called WD40) domains of Apaf-1 (Liu *et al.*, 1996; Skulachev, 1998).

The comparative analysis of the cytochrome *c* interactions with the cytosolic apoptotic machinery in different organisms shows differences between them. Specifically, the yeast cytochrome *c*, although being released after disruption of mitochondria, did not trigger the apoptotic cascade either in yeast system or when added to the vertebrate enzymes (Kluck *et al.*, 2000; Sharonov *et al.*, 2005). The cytochrome *c* of *Drosophila* could not trigger the formation of the fly apoptosome, but could still activate the formation of apoptosome when added to the vertebrate enzyme system (Rodriguez *et al.*, 1999). It appears that specific differences between the cytochromes *c* of yeast, flies, and vertebrates affect their interactions with apoptotic enzymes.

An interesting feature of cytochrome *c* is the presence of a positively charged patch of lysine residues which interacts with the negatively charged "docking" patches at the surface of its functional partners (Pettigrew and Moore, 1987), particularly, the cytochrome oxidase (Witt *et al.*, 1998). We have checked how this pivotal patch has evolved. As shown in **Figure 6.7** by arrows, the number of lysine residues in the patch has increased in the course of evolution from proteobacteria to vertebrates.

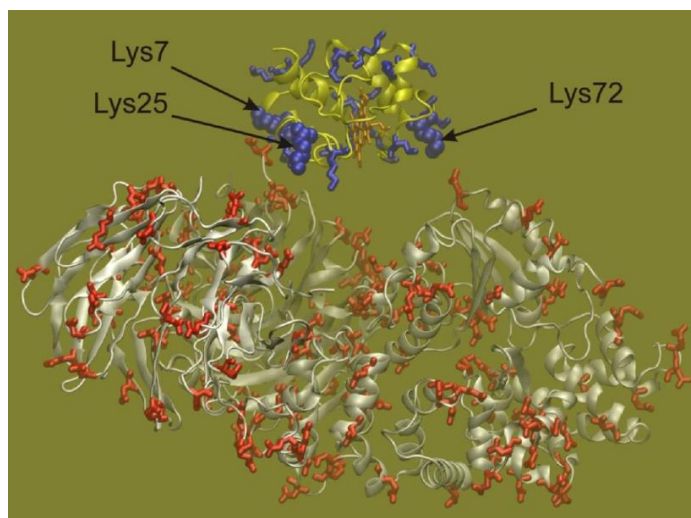
Apparently, the higher number of lysine residues should favor binding of cytochrome *c* to its targets. It is noteworthy that the same set of lysine residues is involved in the interaction with Apaf-1 (Yu *et al.*, 2001) upon triggering the apoptosis. Upon binding to the Apaf-1 protein, cytochrome *c* induces a conformational change of the protein, namely a movement of one of its WD-domains towards another so that the oligomerization domain becomes exposed (**Figure 6.8**). The "open" conformation of the enzyme forms an oligomeric, active apoptosome (Reubold *et al.*, 2011).



**Figure 6.7. Conservation of positively charged residues in the sequences of cytochromes  $c_2$ .**

Multiple alignment of bacterial and eukaryotic cytochromes *c* from fully sequenced genomes (RefSeq release 45) was used to produce the sequence logo. The used numeration corresponds to the horse cytochrome *c*. Each position in the logo corresponds to the position in the alignment while the size of letters in the position represents the relative frequency of corresponding amino acid in this position. Logos were generated with the WebLogo 3 tool from multiple alignments of 168 proteobacterial, 56 fungal, and 209 metazoan sequences.





**Figure 6.8.** A scheme of possible interaction between cytochrome *c* and Apaf-1: patches of charged residues on the surfaces of interacting proteins.

The figure taken from (Dibrova *et al.*, 2013). The structures were taken from the PDB IDs: 1J3S (human cytochrome *c*) and 3SFZ (murine Apaf-1 (Reubold *et al.*, 2011)). Negatively charged residues (Glu and Asp) in the WD-domains of Apaf-1 (residues 600-900 and 901-1200, only the main chain atoms are resolved in the structure) are colored red, while positively charged residues in cytochrome *c* are colored blue. Murine Apaf-1 is more than 87% identical to the human protein and positively charged residues in WD-domains are mostly conserved between them. The figure was produced with the help of the VMD software package (Humphrey *et al.*, 1996).

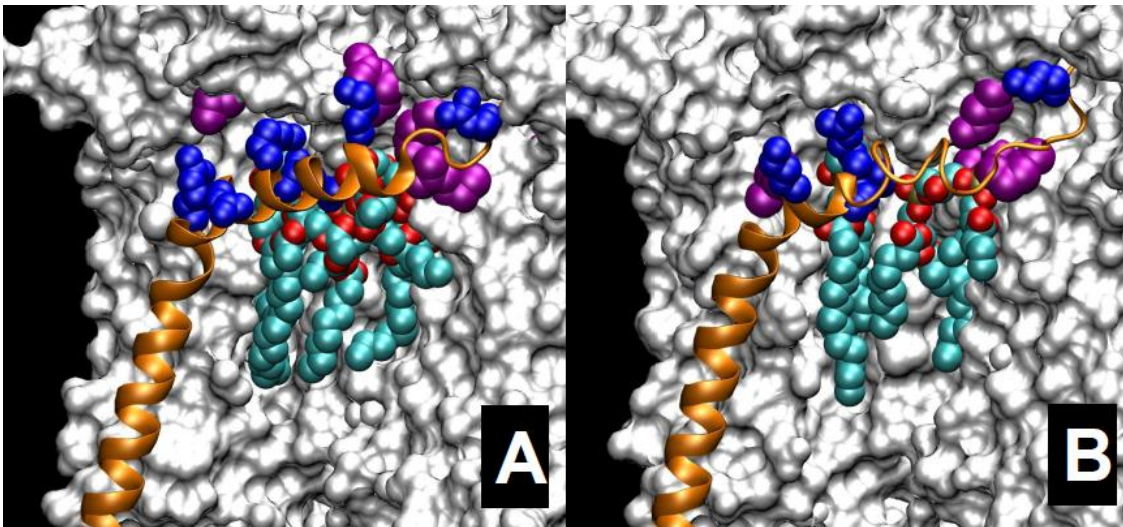
### **6.3. Discussion: The further evolution of components of cytochrome $bc_1$ complex in aerobic environment could have been driven by the optimization of the apoptotic cascade**

The components of the cytochrome  $bc_1$  complex, as suggested by comparison of their sequences from different eukaryotes and prokaryotes, could have evolved a set of ligands for cardiolipin molecules, thus favoring their tight binding to the complex.

While in prokaryotes the ligands for these CL molecules are provided by subunits of cytochromes *b* and *c*<sub>1</sub>, the eukaryotic organisms use one more 9.5 kD subunit subunit to fix the CL molecules patch. With help of this subunit, protein completely encases the lipid molecules by a kind of a "belt", as shown in **Figure 6.9**. In mammals, the  $\alpha$ -helical part of this subunit fully encases the CL patch and additionally stabilizes it by providing positively charged residues (**Figure 6.9A**). CL molecules are tightly bound close to the major source of ROS: the *P*-site of cytochrome  $bc_1$  complex.

In response to the generation of ROS by a mammalian cytochrome  $bc_1$  complex, this patch of bound CL molecules, where eight linoleate chains are tightly packed in the site, would be oxidized pretty soon with the following consequences (see (Dibrova *et al.*, 2013) for more details):

- The conformation of CL molecules would change, and following it the conformation of the cytochrome  $bc_1$  complex will also change. This, according to the data of Yin *et al.* (Yin *et al.*, 2010) should lead to the further increase in the ROS production.
- The peroxidized CL molecules bound either to the cytochrome  $bc_1$  complex proper or to other proteins in the vicinity of this complex would "slip out" out of the bilayer, so that their fatty acid tails could convert molecules of cytochrome  $c$  into peroxidases. The resulting increase in generation of ROS would then trigger the formation of the inner membrane pore, swelling of mitochondrial matrix, disruption of outer mitochondrial membrane, the release of cytochrome  $c$  molecules to cytosol, their interaction with the Apaf-1 protein, and activation of apoptotic caspases (Reubold *et al.*, 2011; Skulachev, 1996; Skulachev, 1998).



**Figure 6.9. Clusters of occluded cardiolipin molecules in the cytochrome  $bc_1$  complexes.**

The figure is taken from (Dibrova *et al.*, 2013). Cardiolipin molecules are colored by element: carbon – cyan, oxygen – red and phosphorus – yellow. The 9.5 kD subunit (subunit G in the bovine cytochrome  $bc_1$  complex and subunit H in the yeast cytochrome  $bc_1$  complex) is colored orange, its positively charged residues are colored blue. The positively charged residues, provided by cytochromes  $b$  and  $c_1$ , are colored violet. (A), bovine cytochrome  $bc_1$  complex (PDB ID 1PP9 (Huang *et al.*, 2005)); (B), yeast cytochrome  $bc_1$  complex (PDB ID 3CX5 (Solmaz and Hunte, 2008)). The figure was produced with the help of the VMD software package (Humphrey *et al.*, 1996).

The water-soluble cytochrome *c*, in the course of evolution from proteobacteria to metazoan, has developed specific traits that made him a trigger of the apoptotic cascade in modern vertebrates (Green and Reed, 1998; Zamzami *et al.*, 1996). As suggested by Kroemer, apoptosis may have evolved together with the endosymbiotic incorporation of bacterial precursor of mitochondria into ancestral unicellular eukaryote (Kroemer, 1997). The increase in number of lysine residues interacting with the Glu- and Asp-rich WD-domains of Apaf-1, which is shown in **Figure 6.7**, might be an example of such co-evolution.

This prompting of a conformational change of the Apaf-1 protein should be apparently possible only if there are enough lysine residues to make bonds with each of the two WD domains. Therefore, all lysine residues are required to activate the Apaf-1, and deletion, by mutation, of at least one of them precludes the ability of cytochrome *c* to activate apoptosis (Yu *et al.*, 2001). Accordingly, the aforementioned data on the interactions between cytochrome *c* and Apaf-1 in yeast and *Drosophila* can be rationalized, based on the sequence analysis, in the following way. The yeast cytochrome *c* cannot activate the vertebrate Apaf-1 (Kluck *et al.*, 2000; Sharonov *et al.*, 2005) because it lacks the lysine residues in positions 7 and 25. In contrast, the fly cytochrome *c* has the whole complement of lysine residues and therefore can activate the vertebrate Apaf-1 protein. However, the homologue of the Apaf-1 in flies has only one WD domain and therefore the vertebrate mechanism of activation is not possible in *Drosophila* (Rodriguez *et al.*, 1999). Only in vertebrates the numerous lysine residues of cytochrome *c* find enough binding partners at the surfaces of the two WD domains to induce conformational transition of Apaf-1. Further studies would be required to reconstruct the evolution of Apaf-1 into the "partner" of cytochrome *c*. Hence, under circumstances when production of ROS could not be avoided, the mitochondrial energy converting enzymes developed an ability to start an apoptotic cascade which begins with a patch of cardiolipin molecules tightly packed in the complex and ends up with the release of cytochrome *c* molecules to cytosol. The amplification of a signal (the increased production of ROS) is likely accomplished by the malfunctioning cytochrome *bc*<sub>1</sub> complex itself (Yin *et al.*, 2010) and by peroxidase activity of a conformationally changed cytochrome *c* (see Section 1.5.5. for details)(Yin *et al.*, 2010). Thus, the key stages in this triggering of the apoptotic cascade seem to be performed by energy converting enzymes (Skulachev, 1998; Skulachev, 2006).

## **7. Outlook: Evolution of biological energy conversion as inferred from phylogenomic analysis of energy-converting enzymes**

In this chapter the results of our phylogenomic analysis are discussed from the viewpoint of the overall evolution process. Based on the data obtained, we put forward a tentative scenario of the evolution of biological energy conversion. In more detail, the scenario is presented in (Dibrova *et al.*, 2012).

### **7.1. K<sup>+</sup> ions as catalysts of the primordial phosphate transfer reactions**

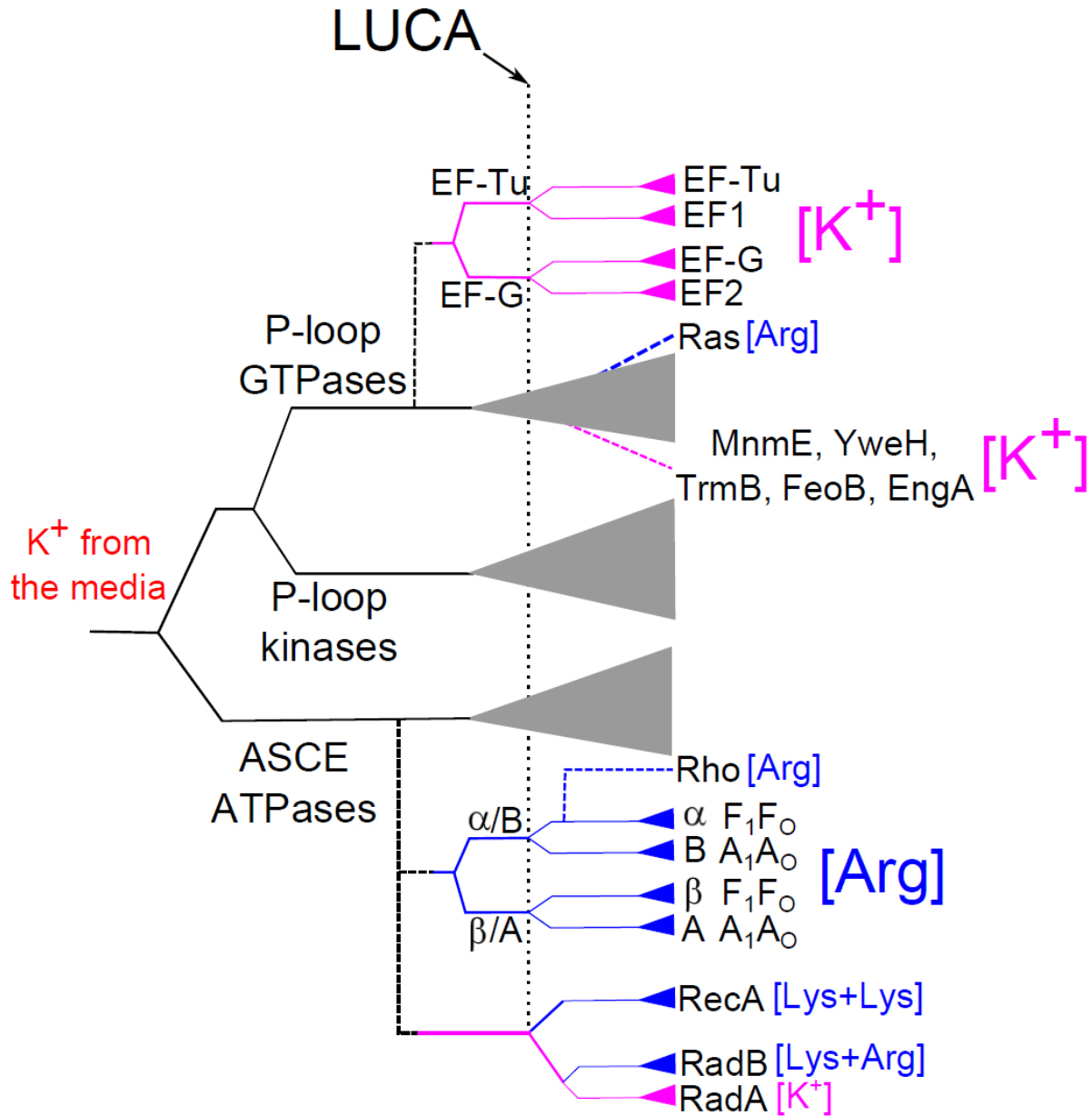
Generally, the absence of any enzymes related to autotrophy in the ubiquitous protein set (see *Table 3.1* and *TableS1*) suggests that the LUCA and its predecessors were heterotrophs, i.e. their growth depended on the supply of abiotically produced nutrients/metabolites as proposed previously from biochemical reasoning (Lazcano and Miller, 1999; Mansy *et al.*, 2008; Miller and Cleaves, 2006; Oparin, 1924). Most metabolites are less complex than nucleotides; the concentrations of metabolites should have been much higher than the concentrations of nucleotides in the habitats of first organisms (Horowitz, 1945). The energetics of the first cells, which can be inferred from the inspection of the ubiquitous protein set, must have been based on phosphate transfer reactions and specifically on hydrolysis of NTPs. That phosphate-based metabolism was ancestral in cellular life follows also from the results of a recent global phylogenomic analysis (David and Alm, 2011).

As shown in chapter 3, several ubiquitous proteins, which apparently belong to the LUCA gene set, show a dependence on K<sup>+</sup> ions. For the inspected enzymes (P-loop GTPase family, members of the RecA/RadA family, molecular chaperons of the GroEL family, ATPases from the GHKL superfamily, membrane pyrophosphatases) we provided evidence for the K<sup>+</sup>-dependence of their ancestral forms. The earlier seminal results of Lowenstein concerning influence of monovalent cations on the rate of non-enzymatic transphosphorylation reaction (Lowenstein, 1960) suggest that large monovalent cations, such as as K<sup>+</sup> and NH<sub>4</sub><sup>+</sup>, are capable of enhancing the transphosphorylation reaction, unlike the Na<sup>+</sup> ions. Thus, utilizing K<sup>+</sup> ion in the reactions involving phosphate group transfer could precede the utilization of

amino acid side chains (Arg or Lys "fingers") in the same capacity. In **Figure 7.1** we present a putative scheme of the evolution of the catalytic site of P-loop NTPases.

It is noteworthy that for the studied  $K^+$  binding sites, the selectivity for  $K^+$  ions as compared to  $Na^+$  ions is low (Page and Di Cera, 2006). Therefore the presence of  $K^+$  in these sites could be secured only in media with  $K^+/Na^+ > 1$ . In modern cells this is achieved by keeping the cytoplasmic  $K^+/Na^+$  ratio  $\gg 1.0$ . However, it was repeatedly argued that at the stage of the LUCA the cells were unlikely to have ion-tight membranes (Deamer, 1997; Deamer, 2008; Mansy *et al.*, 2008; Mulkidjanian and Galperin, 2010; Mulkidjanian *et al.*, 2009). Thus, the ancient organisms were unlikely to maintain disequilibrium, as concerns small monovalent cations, between the inside and outside. These two observations lead to the idea that the first organisms likely lived in the conditions with naturally high  $K^+/Na^+$  ratio, which means that the first forms of life are unlikely to have originated in the sea water. From the geological reconstructions and based on the geological data, the terrestrial anoxic geothermal fields were proposed as the best candidates for the environment of the first life forms (Mulkidjanian *et al.*, 2012). Under these conditions, relatively high potassium concentration together with  $Mg^{2+}$  would catalyze the reactions of phosphate group transfer.

The major open question in the field of the pre-LUCA energetics - as we see it - is the mechanism of recycling of NTPs, i.e. the mechanism of driving the NDP to NTP transitions. It could proceed via diverse enzyme-catalyzed transphosphorylation reactions or even abiotically. The abiotic energy-dependent phosphorylation of NMPs and NDPs to NTPs (re-charging) could proceed either photochemically, given that ATP photorecovery has been shown to occur with high yield on clays (Kritsky *et al.*, 2007) or via phosphite-dependent reactions (Bryant *et al.*, 2010; Pasek *et al.*, 2008). The reconstruction of the primordial recycling mechanisms for ATP and GTP seems to be a major challenge for the future research.



**Figure 7.1. A possible scheme of evolution of the catalytic site of P-loop NTPases.**

The solid lines show the general classification of the P-loop NTPases superfamily, the dashed lines show putative events of the emergence of particular protein families. The vertical dotted line marks the presumed position of the LUCA on the scheme; the length of horizontal lines on the scheme does not represent a real time scale. The magenta color in the branch shows appearance and further spreading of the specific K<sup>+</sup>-binding site in the protein family while the blue color shows appearance and further spreading of lysine and arginine finger(s).

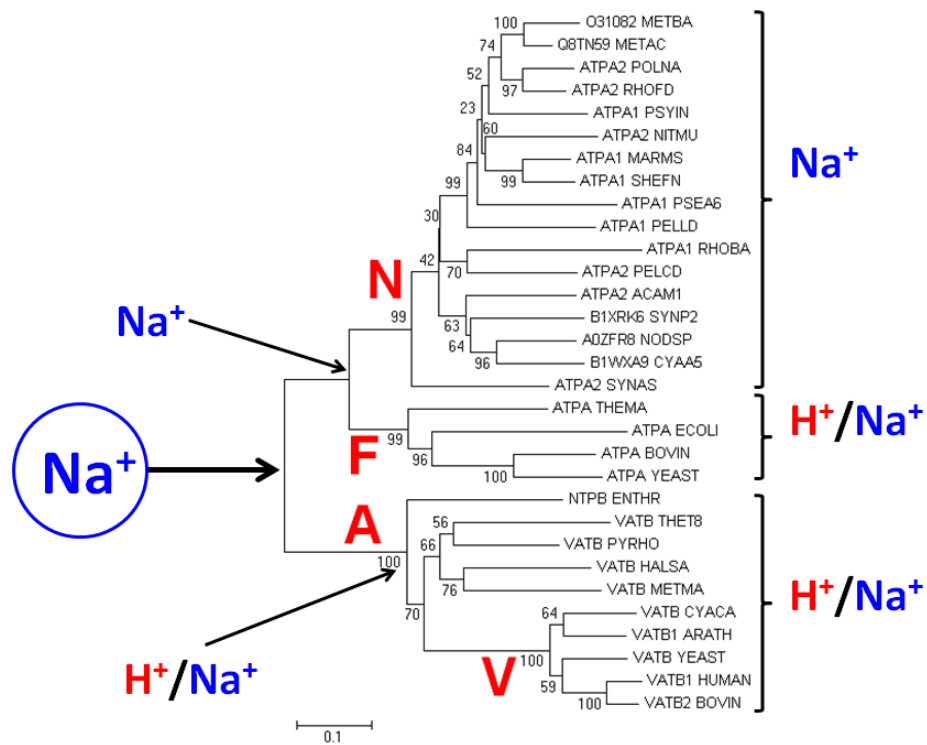
## 7.2. Sodium-dependent membrane bioenergetics: The ancestral form of rotary ATPases was a sodium-dependent enzyme

After the emergence of ion-tight membranes, followed by the emergence of membrane ion pumps, the cells could invade terrestrial water basins with low  $K^+/Na^+$  ratios. Spreading further, they would finally reach the ocean and would have been severely challenged by the high sodium levels. The translation systems of all organisms need high  $K^+/Na^+$  ratio for functioning, see (Mulkiđjanian *et al.*, 2012) and references therein. Therefore, ancient cells would require ion pumps capable of pumping  $Na^+$  ions out of the cell against large concentration backpressure.

Conservation of the  $Na^+$ -binding site in all the  $Na^+$ -translocating ATP synthases, despite them being scattered all over the phylogenetic tree, indicated the presence of the same  $Na^+$ -binding site in the ancestral form of rotary ATP synthase (Mulkiđjanian *et al.*, 2008b). Our analysis of the N-ATPases, which are positioned on the phylogenetic tree as a sister group to the F-ATPases, provides further evidence for this hypothesis. The family of N-ATPases comprises only  $Na^+$ -translocating enzyme complexes (see section 4.4). Since among the F-ATPases there are both  $Na^+$ - and  $H^+$ -translocating forms, the common ancestor of N- and F-ATPases should have been a  $Na^+$ -translocating enzyme (**Figure 7.2**). Following the same rationale and juxtaposing the  $Na^+$ -binding common ancestor of the F/N ATPases with the A/V-type ATPases, which can be either proton- or sodium-dependent, we should conclude that the ancestor of the whole family of rotary ATPases was, in all likelihood, a  $Na^+$ -binding ATPase.

Hence, the phylogenomic analysis shows that the ancestral form of the rotary ATP synthase was binding sodium. Earlier it was argued, based on phylogenomic and structural analysis, that the common ancestor of the rotary ATP synthases, an ATP-driven protein translocase, could give rise to an ATP-driven sodium outpump, needed in sodium-rich environments (Mulkiđjanian and Galperin, 2007; Mulkiđjanian *et al.*, 2009). Unlike the other  $Na^+$  outpumps, such as  $Na^+$ -transporting pyrophosphatase (Luoto *et al.*, 2011) and  $Na^+$ -transporting decarboxylase (Dimroth, 1997), the common ancestor of the rotary ATP synthases, because of its rotating scaffold, would be able to translocate  $Na^+$  ions in both directions, so that, at high external salinity (of  $\sim 1$  M of  $Na^+$  in the Archean ocean (Pinti, 2005)), the reversal of the rotation could result in  $Na^+$ -driven synthesis of ATP. This event,

which could take place close to the stage of the LUCA, would be the beginning of membrane bioenergetics: together with the ancient  $\text{Na}^+$  pumps, the  $\text{Na}^+$ -driven ATP synthase would complete the first, sodium-dependent bioenergetic cycle in a cell membrane (Mulkidjanian *et al.*, 2009). The emergence of a rotary ATP synthase was a major breakthrough: this rotary enzyme could accumulate the energy of several sequentially translocated cations to yield one ATP molecule, this enzyme could harvest "small" energy quanta and use them for ATP synthesis (Pascal and Boiteau, 2011). In the case of sodium ions, such small portions of energy could be obtained already from co-translocation of anionic metabolic products and  $\text{Na}^+$  ions out of the cell.



**Figure 7.2. Phylogenetic tree of the  $\alpha$ -subunits of N- and F-ATPases and B-subunits of A- and V-type ATPases.**

Proteins are shown under their Uniprot identifiers.

### 7.3. Separation of bacteria and archaea and the transition to the oxygenated atmosphere

It is noteworthy that iron-dependent proteins are nearly absent among the ubiquitous protein set, see **Table 3.1** and **Table S1**.  $\text{Fe}^{2+}$  ions and FeS clusters are efficient cleavage agents for both RNA and DNA because these ions trigger the production of active hydroxyl radicals,



especially under illumination (Anbar and Holland, 1992; Cohn *et al.*, 2004; Mulkidjanian *et al.*, 2009). Therefore it has been suggested that the early organisms dwelled in geothermal puddles and ponds that were enriched in Zn and Mn salts but depleted in iron (Mulkidjanian *et al.*, 2012; Mulkidjanian *et al.*, 2009). The late recruitment of iron follows also from the results of a recent global phylogenomic analysis (David and Alm, 2011).

At the reduced, anoxic Earth, redox reaction could hardly be used for gaining energy because of the absence of notable redox gradients. Redox enzymes still should have been needed to maintain the redox equilibria, e.g. by catalyzing the electron exchange within the cell (see thioredoxin and thioredoxin reductase in Table 3.1).

Only after the emergence of ion-tight membranes that could protect RNA and DNA from the damaging action of iron compounds, cells could invade iron-rich environments and iron ions could be recruited, in a controlled way, as redox-active cofactors, e.g. in enzymes which brought redox equivalents, as required for biosynthetic reactions, across – now tight – cellular membranes. Such enzymes could catalyze the redox exchange between the cell interior and its environment, with the involvement of quinones as intermediate, membrane-soluble, hydrophobic electron carriers.

We suggest that only the emergence of photosynthesis within the bacterial clade prompted the appearance of first high-potential electron acceptors, namely, photo-oxidized molecules of bacteriochlorophyll which, in turn, could prompt the emergence of first *b<sub>6</sub>f*-type complexes as enzymes that could obtain free energy from cycling electrons around photosynthetic reaction centers. Only after the oxygenation of the atmosphere, the functional coupling of cytochrome *bc* complexes with diverse oxidases became possible. In the framework of this scenario, archaea, which do not use bacteriochlorophyll-based photosynthesis, had no need for cytochrome *bc* complexes and could obtain them, usually in a "package" with oxidases (the bacterial origin of the cytochrome oxidase has been discussed elsewhere (Hemp and Gennis, 2008; Hemp *et al.*, 2012)), on several different occasions, via the LGT. Koonin and co-workers have argued as early as 2001 that the gene flow from bacteria to archaea was much stronger than in the reverse direction (Koonin *et al.*, 2001). A recent paper suggested that the ancestor of *Halobacteria*, for example, acquired around 1000 genes from bacteria (Nelson-Sathi *et al.*, 2012).

The same oxygenation of the atmosphere led to the formation of ROS as byproducts of several energy converting reactions. Animal tissues have developed apoptosis as a mechanism that allows them to get rid of damaged cells, including the ROS-producing cells. By performing the evolutionary analysis of several energy converting enzymes involved in the apoptotic cascade, we show that the techniques that have been applied in the previous sections to prokaryotic enzymes, which evolved for billions of years, could be also applied on shorter time scales and could help to trace the evolution of functionally important traits within the *Metazoa*.

In sum, evolution of biological energy conversion seems to be driven by the tendency of living organisms to invade new environments (the invasion of the ocean), their ability to change the chemistry of these environments (the oxygenation of the atmosphere) and their ability to adapt to the continuously changing habitats.

## 8. Conclusions

1. Phylogenomic and comparative structural analyses of ancient families of proteins that catalyze hydrolysis of the phosphoester bond have been performed. The following enzyme families have been analyzed: P-loop GTPases, RadA/RecA recombinases, chaperone GroEL, branched-chain  $\alpha$ -ketoacid dehydrogenase kinase, chaperone Hsc70 and actin, and membrane pyrophosphatases. In each family we observed (1) members that were potassium-dependent and/or contained  $K^+$  ions in the active site and (2) potassium-independent members with lysine or arginine residues as catalysts; these residues could be either permanently present or enter the active site during the oligomerization or the interaction with a functional partner protein. These arginine and lysine residues occupy the same positions as potassium ions in the homologous,  $K^+$ -dependent enzymes. Based on the results of our analyses, we suggest, for the inspected protein families, that the appearance of the  $K^+$ -binding sites could have preceded in evolution the recruitment of positively charged residues (lysine or arginine "fingers") in the same catalytic capacity. Specifically, for the ubiquitous translation factors, the putative  $K^+$ -binding sites were not known before and were identified, from the structure comparison, in the course of this work. The obtained results are in agreement with the proposed hypothesis of the origin of the first cells in the  $K^+$ -enriched habitats at terrestrial anoxic geothermal fields.

2. Phylogenomic analysis of the rotary membrane ATPases/ATP synthases revealed a separate subfamily that we named N-ATPases. These enzymes are separated from the known groups of F-, V- and A-type ATPases on the phylogenetic trees, have a specific operon organization with two additional subunits and contain a complete set of  $Na^+$ -binding ligands in their membrane *c*-subunits. Thus, these enzymes are apparently capable of ATP-coupled  $Na^+$  translocation across the membrane. The discovery of a separate family of  $Na^+$ -translocating rotary ATPases provides an additional support to the idea of evolutionary primacy of  $Na^+$ -dependent bioenergetics.

3. Phylogenomic analysis of the cytochrome *bc* complexes suggests that these enzyme complexes initially emerged within the bacterial lineage and were then transferred to archaea

via lateral gene transfer on several independent occasions. The emergence of the ancestral cytochrome *bc* complex in bacteria may have been driven by the emergence of the (bacterio)chlorophyll-based photosynthesis which is accompanied by generation of high-potential electron acceptors. Our analysis indicates that the ancestral form of the cytochrome *bc* complex was a *b<sub>6</sub>f*-type complex; the fusion of the cytochrome *b<sub>6</sub>* and the subunit IV into a "long" cytochrome *b* could have happened in different lineages independently.

4. Phylogenomic and comparative structural analyses of several respiratory enzymes allowed us to trace how these enzymes became involved in triggering of apoptosis in *Metazoa*. We could track the emergence of a specific cardiolipin-binding site within the cytochrome *bc* complex and the evolution of a patch of lysine residues that account for the binding of the cytochrome *c* to the Apaf-1 protein, which makes the cytochrome *c* a key component of the apoptosis signal chain in vertebrates. The involvement of energy converting enzymes in the apoptotic cascade could be considered as an adaptation to the formation of ROS in respiratory chain which could not be avoided under aerobic conditions.

## 9. Summary

In this thesis, phylogenomic and comparative structural analyses of several widespread energy converting enzymes were performed. The focus was on the major subfamilies of the enzymes that process nucleoside triphosphates (ATP and GTP) and on some key enzymes of the electron transfer chains. First, we analyzed the P-loop GTPases, RadA/RecA recombinases, chaperone GroEL, branched-chain  $\alpha$ -ketoacid dehydrogenase kinases, chaperone Hsc70, actins, and membrane pyrophosphatases. In the each inspected family we could identify (1) members which were potassium-dependent and/or contained  $K^+$  ions in the active site, and (2) potassium-independent enzymes with lysine or arginine residues as catalytic groups that occupy the positions of potassium ions in the homologous,  $K^+$ -dependent enzymes. Based on the results of our analyses, we suggest that the appearance of the  $K^+$ -binding sites could precede in evolution the recruitment of positively charged residues (lysine or arginine "fingers") with the latter providing more possibilities to control the enzyme reactions. Second, we have described the distinctive features of a phylogenetically separated subfamily of rotary membrane ATPases which we named N-ATPases. The N-ATPases have a specific operon organization with two additional subunits, absent in other rotary ATPases, and a complete set of  $Na^+$ -binding ligands in the membrane *c*-subunits. We made a prediction, which was later confirmed, that these enzymes are capable of  $Na^+$  translocation across the membrane and may confer salt tolerance on marine prokaryotes. Third, phylogenomic analysis of the cytochrome *bc* complexes suggests that these enzyme complexes initially emerged within the bacteria and were then transferred to archaea via lateral gene transfer on several independent occasions. Our analysis indicates that the ancestral form of the cytochrome *bc* complex was a *b<sub>6</sub>f*-type complex; the fusion of the cytochrome *b<sub>6</sub>* and the subunit IV to a "long" cytochrome *b* of the cytochrome *bc<sub>1</sub>* complexes could have happened in different lineages independently. Fourth, our phylogenomic and comparative structural analyses of the cytochrome *bc<sub>1</sub>* complex and of cytochrome *c* allowed us to trace how these enzymes became involved in triggering of apoptosis in *Metazoa*. We could trace the emergence of a specific cardiolipin-binding site within the cytochrome *bc* complex and the evolution of structural traits that account for the involvement of the cytochrome *c* as a trigger of apoptosis in vertebrates.

## 10. References

- Adachi, K., Oiwa, K., Yoshida, M., Nishizaka, T. and Kinoshita, K., Jr. (2012) Controlled rotation of the F(1)-ATPase reveals differential and continuous binding changes for ATP synthesis, *Nat Commun*, **3**, 1022.
- Al-Attar, S. and de Vries, S. (2012) Energy transduction by respiratory metallo-enzymes: From molecular mechanism to cell physiology, *Coord Chem Rev*, <http://dx.doi.org/10.1016/j.ccr.2012.05.022>.
- Al-Attar, S. and de Vries, S. (2013) Energy transduction by respiratory metallo-enzymes: From molecular mechanism to cell physiology, *Coordination Chemistry Reviews*, **257** 64– 80.
- Alberty, R.A. and Goldberg, R.N. (1992) Standard thermodynamic formation properties for the adenosine 5'-triphosphate series, *Biochemistry*, **31**, 10610-10615.
- Alnemri, E.S., Livingston, D.J., Nicholson, D.W., Salvesen, G., Thornberry, N.A., Wong, W.W. and Yuan, J. (1996) Human ICE/CED-3 protease nomenclature, *Cell*, **87**, 171.
- Althoff, T., Mills, D.J., Popot, J.L. and Kuhlbrandt, W. (2011) Arrangement of electron transport chain components in bovine mitochondrial supercomplex I<sub>1</sub>III<sub>2</sub>IV<sub>1</sub>, *EMBO J*, **30**, 4652-4664.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.
- Anand, B., Surana, P. and Prakash, B. (2010) Deciphering the catalytic machinery in 30S ribosome assembly GTPase YqeH, *PLoS One*, **5**, e9944.
- Anbar, A.D. and Holland, H.D. (1992) The photochemistry of manganese and the origin of Banded Iron Formations, *Geochim Cosmochim Acta*, **56**, 2595-2603.
- Andreyev, A.Y., Kushnareva, Y.E. and Starkov, A.A. (2005) Mitochondrial metabolism of reactive oxygen species, *Biochemistry (Mosc)*, **70**, 200-214.
- Anisimova, M. and Gascuel, O. (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative, *Syst Biol*, **55**, 539-552.
- Anisimova, M., Gil, M., Dufayard, J.F., Dessimoz, C. and Gascuel, O. (2011) Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes, *Syst Biol*, **60**, 685-699.
- Aoyama, H., Muramoto, K., Shinzawa-Itoh, K., Hirata, K., Yamashita, E., Tsukihara, T., Ogura, T. and Yoshikawa, S. (2009) A peroxide bridge between Fe and Cu ions in the O<sub>2</sub> reduction site of fully oxidized cytochrome c oxidase could suppress the proton pump, *Proc Natl Acad Sci U S A*, **106**, 2165-2169.
- Archibald, J.M., Logsdon, J.M., Jr. and Doolittle, W.F. (2000) Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in CCT genes, *Mol Biol Evol*, **17**, 1456-1466.
- Arias-Cartin, R., Grimaldi, S., Arnoux, P., Guigliarelli, B. and Magalon, A. (2012) Cardiolipin binding in bacterial respiratory complexes: Structural and functional implications, *Biochim Biophys Acta*, **1817**, 1937-1949.
- Ash, M.R., Guilfoyle, A., Clarke, R.J., Guss, J.M., Maher, M.J. and Jormakka, M. (2010) Potassium-activated GTPase reaction in the G Protein-coupled ferrous iron transporter B, *J Biol Chem*, **285**, 14594-14602.
- Atkinson, G.C., Baldauf, S.L. and Haurlyuk, V. (2008) Evolution of nonstop, no-go and nonsense-mediated mRNA decay and their termination factor-derived components, *Bmc Evol Biol*, **8**, 290.
- Bagshaw, C.R. and Trentham, D.R. (1973) The reversibility of adenosine triphosphate cleavage by myosin, *Biochem J*, **133**, 323-328.

- Baker, L.A., Watt, I.N., Runswick, M.J., Walker, J.E. and Rubinstein, J.L. (2012) Arrangement of subunits in intact mammalian mitochondrial ATP synthase determined by cryo-EM, *Proc Natl Acad Sci U S A*, **109**, 11675-11680.
- Balaban, R.S., Nemoto, S. and Finkel, T. (2005) Mitochondria, oxidants, and aging, *Cell*, **120**, 483-495.
- Ballhausen, B., Altendorf, K. and Deckers-Hebestreit, G. (2009) Constant c10 ring stoichiometry in the Escherichia coli ATP synthase analyzed by cross-linking, *J Bacteriol*, **191**, 2400-2404.
- Baniulis, D., Zhang, H., Zakharova, T., Hasan, S.S. and Cramer, W.A. (2011) Purification and crystallization of the cyanobacterial cytochrome b6/f complex, *Methods Mol Biol*, **684**, 65-77.
- Baradaran, R., Berrisford, J.M., Minhas, G.S. and Sazanov, L.A. (2013) Crystal structure of the entire respiratory complex I, *Nature*, **494**, 443-448.
- Barks, H.L., Buckley, R., Grieves, G.A., Di Mauro, E., Hud, N.V. and Orlando, T.M. (2010) Guanine, adenine, and hypoxanthine production in UV-irradiated formamide solutions: relaxation of the requirements for prebiotic purine nucleobase formation, *ChemBiochem*, **11**, 1240-1243.
- Baymann, F., Lebrun, E., Brugna, M., Schoepp-Cothenet, B., Giudici-Orticoni, M.T. and Nitschke, W. (2003) The redox protein construction kit: pre-last universal common ancestor evolution of energy-conserving enzymes, *Philos Trans R Soc Lond B Biol Sci*, **358**, 267-274.
- Baymann, F., Lebrun, E. and Nitschke, W. (2004) Mitochondrial cytochrome c1 is a collapsed di-heme cytochrome, *Proc Natl Acad Sci U S A*, **101**, 17737-17740.
- Baymann, F. and Nitschke, W. (2010) Heliobacterial Rieske/cytb complex, *Photosynth Res*, **104**, 177-187.
- Baymann, F., Schoepp-Cothenet, B., Lebrun, E., van Lis, R. and Nitschke, W. (2012) Phylogeny of Rieske/cytb Complexes with a Special Focus on the Haloarchaeal Enzymes, *Genome Biol Evol*, **4**, 720-729.
- Becher, B. and Muller, V. (1994) Delta mu Na<sup>+</sup> drives the synthesis of ATP via an delta mu Na<sup>(+)</sup>-translocating F1F0-ATP synthase in membrane vesicles of the archaeon Methanosarcina mazei Go1, *J Bacteriol*, **176**, 2543-2550.
- Belogurov, G.A. and Lahti, R. (2002) A lysine substitute for K<sup>+</sup>. A460K mutation eliminates K<sup>+</sup> dependence in H<sup>+</sup>-pyrophosphatase of Carboxydotherrmus hydrogenoformans, *J Biol Chem*, **277**, 49651-49654.
- Belousoff, M.J., Davidovich, C., Zimmerman, E., Caspi, Y., Wekselman, I., Rozenszajn, L., Shapira, T., Sade-Falk, O., Taha, L., Bashan, A., Weiss, M.S. and Yonath, A. (2010) Ancient machinery embedded in the contemporary ribosome, *Biochem Soc Trans*, **38**, 422-427.
- Belozersky, A.N. (1957) On species specificity of nucleic acids in bacteria. In Oparin, A.I. (ed), *Vozniknovenie zhizni na Zemle (The Origin of Life on Earth)*. Akad. Nauk SSSR, Moscow, pp. 198-205.
- Belozersky, A.N. (1959) On the species specificity of the nucleic acids of bacteria. In Oparin, A.I., et al. (eds), *The Origin of Life on the Earth*. Pergamon, London, pp. 322-331.
- Benlekbir, S., Bueler, S.A. and Rubinstein, J.L. (2012) Structure of the vacuolar-type ATPase from Saccharomyces cerevisiae at 11-A resolution, *Nat Struct Mol Biol*, **19**, 1356-1362.
- Benner, S.A., Ellington, A.D. and Tauer, A. (1989) Modern metabolism as a palimpsest of the RNA world, *Proc Natl Acad Sci U S A*, **86**, 7054-7058.
- Berg, D.E., Davies, J., Allet, B. and Rochaix, J.D. (1975) Transposition of R factor genes to bacteriophage lambda, *Proc Natl Acad Sci U S A*, **72**, 3628-3632.
- Bergerat, A., de Massy, B., Gabelle, D., Varoutas, P.C., Nicolas, A. and Forterre, P. (1997) An atypical topoisomerase II from Archaea with implications for meiotic recombination, *Nature*, **386**, 414-417.
- Bergsten, J. (2005) A review of long-branch attraction, *Cladistics*, **21** 163-193.

- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank, *Nucleic acids research*, **28**, 235-242.
- Berry, E.A., Guergova-Kuras, M., Huang, L.S. and Crofts, A.R. (2000) Structure and function of cytochrome bc complexes, *Annu Rev Biochem*, **69**, 1005-1075.
- Berry, E.A. and Huang, L.S. (2003) Observations concerning the quinol oxidation site of the cytochrome bc<sub>1</sub> complex, *FEBS Letters*, **555**, 13-20.
- Berry, S. (2002) The chemical basis of membrane bioenergetics, *J Mol Evol*, **54**, 595-613.
- Bertsova, Y.V. and Bogachev, A.V. (2004) The origin of the sodium-dependent NADH oxidation by the respiratory chain of *Klebsiella pneumoniae*, *FEBS Lett*, **563**, 207-212.
- Beyenbach, K.W. and Wiczorek, H. (2006) The V-type H<sup>+</sup> ATPase: molecular structure and function, physiological roles and regulation, *J Exp Biol*, **209**, 577-589.
- Biegel, E., Schmidt, S., Gonzalez, J.M. and Muller, V. (2011) Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes, *Cell Mol Life Sci*, **68**, 613-634.
- Bird, L.J., Coleman, M.L. and Newman, D.K. (2013) Iron and copper act synergistically to delay anaerobic growth in bacteria, *Appl Environ Microbiol*.
- Blankenship, R.E. (1992) Origin and early evolution of photosynthesis, *Photosynth Res*, **33**, 91-111.
- Boltzmann, L. (1979) *Populäre Schriften*. Friedr. Vieweg & Sohn, Braunschweig.
- Bortnikova, S.B., Gavrilenko, G.M., Bessonova, E.P. and Lapukhov, A.S. (2009) The hydrogeochemistry of thermal springs on Mutnovskii Volcano, southern Kamchatka, *J Volcanology and Seismology*, **3**, 388-404.
- Bourne, H.R. (1997) G proteins. The arginine finger strikes again, *Nature*, **389**, 673-674.
- Boyer, P.D. (1993) The binding change mechanism for ATP synthase--some probabilities and possibilities, *Biochim Biophys Acta*, **1140**, 215-250.
- Brandes, J.A., Boctor, N.Z., Cody, G.D., Cooper, B.A., Hazen, R.M. and Yoder, H.S., Jr. (1998) Abiotic nitrogen reduction on the early Earth, *Nature*, **395**, 365-367.
- Brandt, K., Muller, D.B., Hoffmann, J., Hubert, C., Brutschy, B., Deckers-Hebestreit, G. and Muller, V. (2012) Functional production of the Na<sup>(+)</sup> F<sub>1</sub>F<sub>0</sub> ATP synthase from *Acetobacterium woodii* in *Escherichia coli* requires the native AtpI, *J Bioenerg Biomembr*.
- Brandt, U. (2006) Energy converting NADH:quinone oxidoreductase (complex I), *Annu Rev Biochem*, **75**, 69-92.
- Brenner, C. and Grimm, S. (2006) The permeability transition pore complex in cancer cell death, *Oncogene*, **25**, 4744-4756.
- Bryant, D.E., Marriott, K.E., Macgregor, S.A., Kilner, C., Pasek, M.A. and Kee, T.P. (2010) On the prebiotic potential of reduced oxidation state phosphorus: the H-phosphinate-pyruvate system, *Chem Commun (Camb)*, **46**, 3726-3728.
- Brzezinski, P. and Gennis, R.B. (2008) Cytochrome c oxidase: exciting progress and remaining mysteries, *J Bioenerg Biomembr*, **40**, 521-531.
- Burgers, P.M. and Eckstein, F. (1979) A study of the mechanism of DNA polymerase I from *Escherichia coli* with diastereomeric phosphorothioate analogs of deoxyadenosine triphosphate, *J Biol Chem*, **254**, 6889-6893.
- Busch, K.B., Deckers-Hebestreit, G., Hanke, G.T. and Mulikidjanian, A.Y. (2012) Dynamics of bioenergetic microcompartments, *Biol Chem*.
- Canman, C.E., Tang, H.Y., Normolle, D.P., Lawrence, T.S. and Maybaum, J. (1992) Variations in patterns of DNA damage induced in human colorectal tumor cells by 5-fluorodeoxyuridine: implications for mechanisms of resistance and cytotoxicity, *Proc Natl Acad Sci U S A*, **89**, 10474-10478.



- Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, *Mol Biol Evol*, **17**, 540-552.
- Castresana, J., Lubben, M., Saraste, M. and Higgins, D.G. (1994) Evolution of cytochrome oxidase, an enzyme older than atmospheric oxygen, *EMBO J*, **13**, 2516-2525.
- Castresana, J. and Moreira, D. (1999) Respiratory chains in the last common ancestor of living organisms, *J Mol Evol*, **49**, 453-460.
- Cech, T.R. and Brehm, S.L. (1981) Replication of the extrachromosomal ribosomal RNA genes of *Tetrahymena thermophila*, *Nucleic Acids Res*, **9**, 3531-3543.
- Chang, A., Scheer, M., Grote, A., Schomburg, I. and Schomburg, D. (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009, *Nucleic acids research*, **37**, D588-592.
- Charlebois, R.L. and Doolittle, W.F. (2004) Computing prokaryotic gene ubiquity: rescuing the core from extinction, *Genome research*, **14**, 2469-2477.
- Chaudhry, C., Farr, G.W., Todd, M.J., Rye, H.S., Brunger, A.T., Adams, P.D., Horwich, A.L. and Sigler, P.B. (2003) Role of the gamma-phosphate of ATP in triggering protein folding by GroEL-GroES: function, structure and energetics, *EMBO J*, **22**, 4877-4887.
- Cheek, S., Zhang, H. and Grishin, N.V. (2002) Sequence and structure classification of kinases, *J Mol Biol*, **320**, 855-881.
- Chen, Z., Yang, H. and Pavletich, N.P. (2008) Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures, *Nature*, **453**, 489-484.
- Cherepanov, D.A., Mulkidjanian, A.Y. and Junge, W. (1999) Transient accumulation of elastic energy in proton translocating ATP synthase, *FEBS Lett*, **449**, 1-6.
- Chernyak, B.V., Dibrov, P.A., Glagolev, A.N., Sherman, M.Y. and Skulachev, V.P. (1983) A novel type of energetics in a marine alkali-tolerant bacterium:  $\Delta\mu\text{Na}$ -driven motility and sodium cycle, *FEBS Lett*, **164**, 38-42.
- Choma, C., Lear JD, Nelson, M., Dutton, P., Robertson, D. and DeGrado, W. (1994) Design of a heme-binding four-helix bundle, *J. Am. Chem. Soc.*, **116**, 856-865.
- Cleland, W.W. and Hengge, A.C. (2006) Enzymatic mechanisms of phosphate and sulfate transfer, *Chem Rev*, **106**, 3252-3278.
- Cohn, C.A., Borda, M.J. and Schoonen, M.A. (2004) RNA decomposition by pyrite-induced radicals and possible role of lipids during the emergence of life, *Earth Planet. Sci. Lett.*, **225**, 271-278.
- Cole, C., Barber, J.D. and Barton, G.J. (2008) The Jpred 3 secondary structure prediction server, *Nucleic Acids Res*, **36**, W197-201.
- Connolly, B. and West, S.C. (1990) Genetic recombination in *Escherichia coli*: Holliday junctions made by RecA protein are resolved by fractionated cell-free extracts, *Proc Natl Acad Sci U S A*, **87**, 8476-8480.
- Conway, T.W. and Lipmann, F. (1964) Characterization of a Ribosome-Linked Guanosine Triphosphatase in *Escherichia Coli* Extracts, *Proc Natl Acad Sci U S A*, **52**, 1462-1469.
- Costanzo, G., Pino, S., Botta, G., Saladino, R. and Di Mauro, E. (2011) May cyclic nucleotides be a source for abiotic RNA synthesis?, *Orig Life Evol Biosph*, **41**, 559-562.
- Costanzo, G., Saladino, R., Crestini, C., Ciciriello, F. and Di Mauro, E. (2007) Nucleoside phosphorylation by phosphate minerals, *J Biol Chem*, **282**, 16729-16735.
- Cramer, W.A., Hasan, S.S. and Yamashita, E. (2011) The Q cycle of cytochrome bc complexes: a structure perspective, *Biochim Biophys Acta*, **1807**, 788-802.
- Cramer, W.A. and Knaff, D.B. (1991) *Energy Transduction in Biological Membranes*. Springer-Verlag, New York.

- Cramer, W.A., Zhang, H., Yan, J., Kurisu, G. and Smith, J.L. (2006) Transmembrane traffic in the cytochrome b6f complex, *Annu Rev Biochem*, **75**, 769-790.
- Crofts, A.R. and Wraight, C.A. (1983) The electrochemical domain of photosynthesis, *Biochim Biophys Acta*, **726**, 149-185.
- Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator, *Genome Res*, **14**, 1188-1190.
- Crowley, P.J., Berry, E.A., Cromartie, T., Daldal, F., Godfrey, C.R., Lee, D.W., Phillips, J.E., Taylor, A. and Viner, R. (2008) The role of molecular modeling in the design of analogues of the fungicidal natural products crocacin A and D, *Bioorg Med Chem*, **16**, 10345-10355.
- Dashdorj, N., Zhang, H., Kim, H., Yan, J., Cramer, W.A. and Savikhin, S. (2005) The single chlorophyll a molecule in the cytochrome b6f complex: unusual optical properties protect the complex against singlet oxygen, *Biophysical Journal*, **88**, 4178-4187.
- David, L.A. and Alm, E.J. (2011) Rapid evolutionary innovation during an Archaeal genetic expansion, *Nature*, **469**, 93-96.
- Davidovich, C., Belousoff, M., Bashan, A. and Yonath, A. (2009) The evolving ribosome: from non-coded peptide bond formation to sophisticated translation machinery, *Res Microbiol*, **160**, 487-492.
- Davidson, T., Beck, E., Ganapathy, A., Montgomery, R., Zafar, N., Yang, Q., Madupu, R., Goetz, P., Galinsky, K., White, O. and Sutton, G. (2010) The comprehensive microbial resource, *Nucleic Acids Res*, **38**, D340-345.
- Davis, B.K. (2002) Molecular evolution before the origin of species, *Prog Biophys Mol Biol*, **79**, 77-133.
- Dawson, T.M. and Dawson, V.L. (2003) Molecular pathways of neurodegeneration in Parkinson's disease, *Science*, **302**, 819-822.
- Dayhoff, M., Schwartz, R. and Orcutt, B. (1978) A model of evolutionary change in proteins. In, *Atlas of Protein Sequence and Structure*. Washington, Natl. Biomed. Res. Found., pp. 345-352.
- de Boer, P.A., Crossley, R.E. and Rothfield, L.I. (1989) A division inhibitor and a topological specificity factor coded for by the minicell locus determine proper placement of the division septum in *E. coli*, *Cell*, **56**, 641-649.
- de Meis, L., Behrens, M.I., Petretski, J.H. and Politi, M.J. (1985) Contribution of water to free energy of hydrolysis of pyrophosphate, *Biochemistry*, **24**, 7783-7789.
- Deamer, D.W. (1997) The first living systems: a bioenergetic perspective, *Microbiol Mol Biol Rev*, **61**, 239-261.
- Deamer, D.W. (2008) Origins of life: How leaky were primitive cells?, *Nature*, **454**, 37-38.
- Deamer, D.W. and Dworkin, J.P. (2005) Chemistry and physics of primitive membranes, *Top Curr Chem*, **259**, 1-27.
- Deckers-Hebestreit, G. and Altendorf, K. (1996) The F0F1-type ATP synthases of bacteria: structure and function of the F0 complex, *Annu Rev Microbiol*, **50**, 791-824.
- Deppenmeier, U. (2002) Redox-driven proton translocation in methanogenic Archaea, *Cell Mol Life Sci*, **59**, 1513-1533.
- Dibrov, P.A. (1991) The role of sodium ion transport in *Escherichia coli* energetics, *Biochim Biophys Acta*, **1056**, 209-224.
- Dibrova, D.V., Cherepanov, D.A., Galperin, M., Skulachev, V.P. and Mulikidjanian, A. (2013) Evolution of cytochrome bc complexes: from membrane electron translocases to triggers of apoptosis, *Biochim. Biophys. Acta*, (**Invited Review, accepted**).
- Dibrova, D.V., Chudetsky, M.Y., Galperin, M., Koonin, E. and Mulikidjanian, A. (2012) Role of energy in the emergence of biology from chemistry, *Orig Life Evol Biosph*, **42**, 459-468.

- Dibrova, D.V., Galperin, M.Y. and Mulikidjanian, A.Y. (2010) Characterization of the N-ATPase, a distinct, laterally transferred Na<sup>+</sup>-translocating form of the bacterial F-type membrane ATPase, *Bioinformatics*, **26**, 1473-1476.
- Dimroth, P. (1997) Primary sodium ion translocating enzymes, *Biochim Biophys Acta*, **1318**, 11-51.
- Dong, H. and Fillingame, R.H. (2010) Chemical reactivities of cysteine substitutions in subunit a of ATP synthase define residues gating H<sup>+</sup> transport from each side of the membrane, *J Biol Chem*, **285**, 39811-39818.
- Doublet, S., Tabor, S., Long, A.M., Richardson, C.C. and Ellenberger, T. (1998) Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution, *Nature*, **391**, 251-258.
- Drory, O. and Nelson, N. (2006) The emerging structure of vacuolar ATPases, *Physiology (Bethesda)*, **21**, 317-325.
- Drose, S. and Brandt, U. (2008) The mechanism of mitochondrial superoxide production by the cytochrome bc<sub>1</sub> complex, *J Biol Chem*, **283**, 21649-21654.
- Dupont, C.L., Butcher, A., Valas, R.E., Bourne, P.E. and Caetano-Anolles, G. (2010) History of biological metal utilization inferred through phylogenomic analysis of protein structures, *Proc Natl Acad Sci U S A*, **107**, 10567-10572.
- Dutta, D., Bandyopadhyay, K., Datta, A.B., Sardesai, A.A. and Parrack, P. (2009) Properties of HflX, an enigmatic protein from Escherichia coli, *J Bacteriol*, **191**, 2307-2314.
- Dutta, R. and Inouye, M. (2000) GHKL, an emergent ATPase/kinase superfamily, *Trends Biochem Sci*, **25**, 24-28.
- Dzioba, J., Hase, C.C., Gosink, K., Galperin, M.Y. and Dibrov, P. (2003) Experimental verification of a sequence-based prediction: F(1)F(0)-type ATPase of Vibrio cholerae transports protons, not Na<sup>(+)</sup> ions, *J Bacteriol*, **185**, 674-678.
- Earnshaw, W.C., Martins, L.M. and Kaufmann, S.H. (1999) Mammalian caspases: structure, activation, substrates, and functions during apoptosis, *Annu Rev Biochem*, **68**, 383-424.
- Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity, *BMC Bioinformatics*, **5**, 113.
- Edgell, D.R. and Doolittle, W.F. (1997) Archaea and the origin(s) of DNA replication proteins, *Cell*, **89**, 995-998.
- Efremov, R.G., Baradaran, R. and Sazanov, L.A. (2010) The architecture of respiratory complex I, *Nature*, **465**, 441-445.
- Efremov, R.G. and Sazanov, L.A. (2011) Structure of the membrane domain of respiratory complex I, *Nature*, **476**, 414-420.
- Efremov, R.G. and Sazanov, L.A. (2012) The coupling mechanism of respiratory complex I - A structural and evolutionary perspective, *Biochim Biophys Acta*, **1817**, 1785-1795.
- Elston, T., Wang, H. and Oster, G. (1998) Energy transduction in ATP synthase, *Nature*, **391**, 510-513.
- Enari, M., Sakahira, H., Yokoyama, H., Okawa, K., Iwamatsu, A. and Nagata, S. (1998) A caspase-activated DNase that degrades DNA during apoptosis, and its inhibitor ICAD, *Nature*, **391**, 43-50.
- Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events, *Nature*, **402**, 86-90.
- Erzberger, J.P. and Berger, J.M. (2006) Evolutionary relationships and structural mechanisms of AAA+ proteins, *Annu Rev Biophys Biomol Struct*, **35**, 93-114.
- Esser, L., Gong, X., Yang, S., Yu, L., Yu, C.A. and Xia, D. (2006) Surface-modulated motion switch: capture and release of iron-sulfur protein in the cytochrome bc<sub>1</sub> complex, *Proc Natl Acad Sci U S A*, **103**, 13045-13050.

- Ettwig, K.F., Butler, M.K., Le Paslier, D., Pelletier, E., Mangenot, S., Kuypers, M.M., Schreiber, F., Dutilh, B.E., Zedelius, J., de Beer, D., Gloerich, J., Wessels, H.J., van Alen, T., Luesken, F., Wu, M.L., van de Pas-Schoonen, K.T., Op den Camp, H.J., Janssen-Megens, E.M., Francoijs, K.J., Stunnenberg, H., Weissenbach, J., Jetten, M.S. and Strous, M. (2010) Nitrite-driven anaerobic methane oxidation by oxygenic bacteria, *Nature*, **464**, 543-548.
- Fasano, O., De Vendittis, E. and Parmeggiani, A. (1982) Hydrolysis of GTP by elongation factor Tu can be induced by monovalent cations in the absence of other effectors, *J Biol Chem*, **257**, 3145-3150.
- Fearnley, I.M. and Walker, J.E. (1992) Conservation of sequences of subunits of mitochondrial complex I and their relationships with other proteins, *Biochim Biophys Acta*, **1140**, 105-134.
- Felsenstein, J. (1973) Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters, *Systematic Zoology*, **22**, 240-249.
- Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap, *Evolution*, **39**, 783-791.
- Feniouk, B.A., Kozlova, M.A., Knorre, D.A., Cherepanov, D.A., Mulkidjanian, A.Y. and Junge, W. (2004) The proton-driven rotor of ATP synthase: ohmic conductance (10 fS), and absence of voltage gating, *Biophys J*, **86**, 4094-4109.
- Fernandez-Medarde, A. and Santos, E. (2011) Ras in cancer and developmental diseases, *Genes Cancer*, **2**, 344-358.
- Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching, *Nucleic Acids Res*, **39**, W29-37.
- Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R. and Bateman, A. (2010) The Pfam protein families database, *Nucleic Acids Res*, **38**, D211-222.
- Fitch, W. (1971) Toward defining the course of evolution: minimum change for a specific tree topology, *Systematic Zoology*, **20**, 406-416.
- Foucher, A.E., Reiser, J.B., Ebel, C., Housset, D. and Jault, J.M. (2012) Potassium acts as a GTPase-activating element on each nucleotide-binding domain of the essential *Bacillus subtilis* EngA, *PLoS One*, **7**, e46795.
- Fox, G.E. (2010) Origin and evolution of the ribosome, *Cold Spring Harb Perspect Biol*, **2**, a003483.
- Friedrich, T., Weidner, U., Nehls, U., Fecke, W., Schneider, R. and Weiss, H. (1993) Attempts to define distinct parts of NADH:ubiquinone oxidoreductase (complex I), *J Bioenerg Biomembr*, **25**, 331-337.
- Friedrich, T. and Weiss, H. (1997) Modular evolution of the respiratory NADH:ubiquinone oxidoreductase and the origin of its modules, *J Theor Biol*, **187**, 529-540.
- Fritz, M., Klyszejko, A.L., Morgner, N., Vonck, J., Brutschy, B., Muller, D.J., Meier, T. and Muller, V. (2008) An intermediate step in the evolution of ATPases: a hybrid F(0)-V(0) rotor in a bacterial Na(+) F(1)F(0) ATP synthase, *FEBS J*, **275**, 1999-2007.
- Fukata, Y., Amano, M. and Kaibuchi, K. (2001) Rho-Rho-kinase pathway in smooth muscle contraction and cytoskeletal reorganization of non-muscle cells, *Trends Pharmacol Sci*, **22**, 32-39.
- Furbacher, P.N., Tae, G.S. and Cramer, W.A. (1996) Evolution and origins of the cytochrome bc1 and b6f complexes. In Baltscheffsky, H. (ed), *Origin and evolution of biological energy conversion*. Wiley-VCH, New York, pp. 221-253.
- Galluzzi, L., Vitale, I., Abrams, J.M., Alnemri, E.S., Baehrecke, E.H., Blagosklonny, M.V., Dawson, T.M., Dawson, V.L., El-Deiry, W.S., Fulda, S., Gottlieb, E., Green, D.R., Hengartner, M.O., Kepp, O., Knight, R.A., Kumar, S., Lipton, S.A., Lu, X., Madeo, F., Malorni, W., Mehlen, P., Nunez, G., Peter, M.E., Piacentini, M., Rubinsztein, D.C., Shi, Y.,

- Simon, H.U., Vandenabeele, P., White, E., Yuan, J., Zhivotovsky, B., Melino, G. and Kroemer, G. (2012) Molecular definitions of cell death subroutines: recommendations of the Nomenclature Committee on Cell Death 2012, *Cell Death Differ*, **19**, 107-120.
- Gao, X., Wen, X., Esser, L., Quinn, B., Yu, L., Yu, C.A. and Xia, D. (2003) Structural basis for the quinone reduction in the bc1 complex: a comparative analysis of crystal structures of mitochondrial cytochrome bc1 with bound substrate and inhibitors at the Qi site, *Biochemistry*, **42**, 9067-9080.
- Gemperli, A.C., Dimroth, P. and Steuber, J. (2002) The respiratory complex I (NDH I) from *Klebsiella pneumoniae*, a sodium pump, *J Biol Chem*, **277**, 33811-33817.
- George, P., Witonsky, R.J., Trachtman, M., Wu, C., Dorwart, W., Richman, L., Richman, W., Shurayh, F. and Lentz, B. (1970) "Squiggle-H<sub>2</sub>O". An enquiry into the importance of solvation effects in phosphate ester and anhydride reactions, *Biochim Biophys Acta*, **223**, 1-15.
- Gibbons, C., Montgomery, M.G., Leslie, A.G. and Walker, J.E. (2000) The structure of the central stalk in bovine F(1)-ATPase at 2.4 Å resolution, *Nat Struct Biol*, **7**, 1055-1061.
- Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E.A., Young, L., Qi, Z.Q., Segall-Shapiro, T.H., Calvey, C.H., Parmar, P.P., Hutchison, C.A., 3rd, Smith, H.O. and Venter, J.C. (2010) Creation of a bacterial cell controlled by a chemically synthesized genome, *Science*, **329**, 52-56.
- Gilbert, W. (1986) Origin of life: The RNA world, *Nature*, **319**, 618.
- Glagolev, A.N. and Skulachev, V.P. (1978) The proton pump is a molecular engine of motile bacteria, *Nature*, **272**, 280-282.
- Glansdorff, N., Xu, Y. and Labedan, B. (2008) The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner, *Biol Direct*, **3**, 29.
- Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooshef, S., Lewis, M.R., Maruf, M., Hutchison, C.A., 3rd, Smith, H.O. and Venter, J.C. (2006) Essential genes of a minimal bacterium, *Proc Natl Acad Sci U S A*, **103**, 425-430.
- Glockner, F.O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., Gade, D., Beck, A., Borzym, K., Heitmann, K., Rabus, R., Schlesner, H., Amann, R. and Reinhardt, R. (2003) Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1, *Proc Natl Acad Sci U S A*, **100**, 8298-8303.
- Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J., Bowman, B.J., Manolson, M.F., Poole, R.J., Date, T., Oshima, T. and et al. (1989) Evolution of the vacuolar H<sup>+</sup>-ATPase: implications for the origin of eukaryotes, *Proc Natl Acad Sci U S A*, **86**, 6661-6665.
- Gogarten, J.P. and Townsend, J.P. (2005) Horizontal gene transfer, genome innovation and evolution, *Nat Rev Microbiol*, **3**, 679-687.
- Gopta, O.A., Feniouk, B.A., Junge, W. and Mulikjanian, A.Y. (1998) The cytochrome bc1 complex of *Rhodobacter capsulatus*: ubiquinol oxidation in a dimeric Q-cycle?, *FEBS Lett*, **431**, 291-296.
- Gorbalenya, A.E. and Koonin, E. (1993) Helicases: amino acid sequence comparisons and structure-function relationships, *Current Opinion in Structural Biology*, **3**, 419-429.
- Gorbalenya, A.E. and Koonin, E.V. (1990) Superfamily of UvrA-related NTP-binding proteins. Implications for rational classification of recombination/repair systems, *J Mol Biol*, **213**, 583-591.
- Gotoh, M., Sugawara, A., Akiyoshi, K., Matsumoto, I., Ourisson, G. and Nakatani, Y. (2007) Possible molecular evolution of biomembranes: from single-chain to double-chain lipids, *Chem Biodivers*, **4**, 837-848.

- Gottschalk, G. and Thauer, R.K. (2001) The Na<sup>(+)</sup>-translocating methyltransferase complex from methanogenic archaea, *Biochim Biophys Acta*, **1505**, 28-36.
- Govindjee, Whitmarsh, J. and Govindjee (1982) Introduction to photosynthesis: Energy conversion by plants and bacteria. In, *Photosynthesis*. Academic Press, New York, pp. 1-16.
- Grabowski, P.J., Zaug, A.J. and Cech, T.R. (1981) The intervening sequence of the ribosomal RNA precursor is converted to a circular RNA in isolated nuclei of Tetrahymena, *Cell*, **23**, 467-476.
- Granick, S. (1957) Speculations on the origins and evolution of photosynthesis, *Annals of the New York Academy of Sciences*, **69**, 292-308.
- Gray, M.W. (1989) The evolutionary origins of organelles, *Trends Genet*, **5**, 294-299.
- Green, D.R. and Reed, J.C. (1998) Mitochondria and apoptosis, *Science*, **281**, 1309-1312.
- Gregory, S.T. and Dahlberg, A.E. (2004) Peptide bond formation is all about proximity, *Nat Struct Mol Biol*, **11**, 586-587.
- Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins, *Proc Natl Acad Sci U S A*, **84**, 4355-4358.
- Grivennikova, V.G. and Vinogradov, A.D. (2006) Generation of superoxide by the mitochondrial Complex I, *Biochim Biophys Acta*, **1757**, 553-561.
- Guarente, L.P., Isberg, R.R., Syvanen, M. and Silhavy, T.J. (1980) Conferral of transposable properties to a chromosomal gene in Escherichia coli, *J Mol Biol*, **141**, 235-248.
- Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, *Syst Biol*, **59**, 307-321.
- Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, *Syst Biol*, **52**, 696-704.
- Guy, C.P., Haldenby, S., Brindley, A., Walsh, D.A., Briggs, G.S., Warren, M.J., Allers, T. and Bolt, E.L. (2006) Interactions of RadB, a DNA repair protein in archaea, with DNA and ATP, *J Mol Biol*, **358**, 46-56.
- Guyann, R.W. and Veech, R.L. (1973) The equilibrium constants of the adenosine triphosphate hydrolysis and the adenosine triphosphate-citrate lyase reactions, *J Biol Chem*, **248**, 6966-6972.
- Guzman, M.I. and Martin, S.T. (2009) Prebiotic metabolism: production by mineral photoelectrochemistry of alpha-ketocarboxylic acids in the reductive tricarboxylic acid cycle, *Astrobiology*, **9**, 833-842.
- Hagerhall, C. (1997) Succinate: quinone oxidoreductases. Variations on a conserved theme, *Biochim Biophys Acta*, **1320**, 107-141.
- Hagerhall, C. and Hederstedt, L. (1996) A structural model for the membrane-integral domain of succinate: quinone oxidoreductases, *FEBS Lett*, **389**, 25-31.
- Haldane, J.B.S. (1929) The origin of life, *The Rationalist Annual*, **148**, 3-10.
- Haldenby, S., White, M.F. and Allers, T. (2009) RecA family proteins in archaea: RadA and its cousins, *Biochem Soc Trans*, **37**, 102-107.
- Hanke, T., Richhardt, J., Polen, T., Sahm, H., Bringer, S. and Bott, M. (2012) Influence of oxygen limitation, absence of the cytochrome bc(1) complex and low pH on global gene expression in Gluconobacter oxydans 621H using DNA microarray technology, *J Biotechnol*, **157**, 359-372.
- Hao, G.F., Wang, F., Li, H., Zhu, X.L., Yang, W.C., Huang, L.S., Wu, J.W., Berry, E.A. and Yang, G.F. (2012) Computational discovery of picomolar Q(o) site inhibitors of cytochrome bc1 complex, *J Am Chem Soc*, **134**, 11168-11176.

- Harris, R.A., Popov, K.M., Zhao, Y., Kedishvili, N.Y., Shimomura, Y. and Crabb, D.W. (1995) A new family of protein kinases--the mitochondrial protein kinases, *Adv Enzyme Regul*, **35**, 147-162.
- Hartl, F.U., Bracher, A. and Hayer-Hartl, M. (2011) Molecular chaperones in protein folding and proteostasis, *Nature*, **475**, 324-332.
- Hasan, S.S., Yamashita, E., Baniulis, D. and Cramer, W.A. (2013) Quinone-dependent proton transfer pathways in the photosynthetic cytochrome b6f complex, *Proc Natl Acad Sci U S A*, **110**, 4297-4302.
- Hasan, S.S., Yamashita, E., Ryan, C.M., Whitelegge, J.P. and Cramer, W.A. (2011) Conservation of lipid functions in cytochrome bc complexes, *J Mol Biol*, **414**, 145-162.
- Hase, C.C., Fedorova, N.D., Galperin, M.Y. and Distrov, P.A. (2001) Sodium ion cycle in bacterial pathogens: evidence from cross-genome comparisons, *Microbiol Mol Biol Rev*, **65**, 353-370, table of contents.
- Hawkins, A.R. and Lamb, H.K. (1995) The molecular biology of multidomain proteins. Selected examples, *Eur J Biochem*, **232**, 7-18.
- Hazen, R.M., Bekker, A., Bish, D.L., Bleeker, W., Downs, R.T., Farquhar, J., Ferry, J.M., Grew, E.S., Knoll, A.H., Papineau, D., Ralph, J.P., Sverjensky, D.A. and Valley, J.W. (2011) Needs and opportunities in mineral evolution research, *Am Mineral*, **96**, 953-963.
- Hedderich, R. and Forzi, L. (2005) Energy-converting [NiFe] hydrogenases: more than just H<sub>2</sub> activation, *J Mol Microbiol Biotechnol*, **10**, 92-104.
- Hemp, J. and Gennis, R.B. (2008) Diversity of the Heme-Copper Superfamily in Archaea: Insights from Genomics and Structural Modeling In Schäfer, G. and Penefsky, H.S. (eds), *Results and Problems in Cell Differentiation*. Springer-Verlag, Berlin Heidelberg.
- Hemp, J., Laura, A.P. and Gennis, R.B. (2012) The heme-copper oxidoreductase superfamily: Diversity, evolution and ecology, *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, **1817** S107-S108.
- Henglein, A. (1984) Catalysis of photochemical reactions by colloidal semiconductors, *Pure & Appl. Chem.*, **56**, 1215—1224.
- Henikoff, J.G. and Henikoff, S. (1996) Using substitution probabilities to improve position-specific scoring matrices, *Comput Appl Biosci*, **12**, 135-143.
- Henikoff, S. and Henikoff, J.G. (1991) Automated assembly of protein blocks for database searching, *Nucleic Acids Res*, **19**, 6565-6572.
- Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci U S A*, **89**, 10915-10919.
- Henikoff, S., Henikoff, J.G. and Pietrokovski, S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations, *Bioinformatics*, **15**, 471-479.
- Henikoff, S., Wallace, J.C. and Brown, J.P. (1990) Finding protein similarities with nucleotide sequence databases, *Methods Enzymol*, **183**, 111-132.
- Higgins, D.G. and Sharp, P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer, *Gene*, **73**, 237-244.
- Hilario, E. and Gogarten, J.P. (1993) Horizontal transfer of ATPase genes--the tree of life becomes a net of life, *Biosystems*, **31**, 111-119.
- Hohmann-Marriott, M.F. and Blankenship, R.E. (2011) Evolution of photosynthesis, *Annual review of plant biology*, **62**, 515-548.
- Horovitz, A., Fridmann, Y., Kafri, G. and Yifrach, O. (2001) Review: allostery in chaperonins, *J Struct Biol*, **135**, 104-114.
- Horowitz, N.H. (1945) On the Evolution of Biochemical Syntheses, *Proc Natl Acad Sci U S A*, **31**, 153-157.

- Horsefield, R., Yankovskaya, V., Sexton, G., Whittingham, W., Shiomi, K., Omura, S., Byrne, B., Cecchini, G. and Iwata, S. (2006) Structural and computational analysis of the quinone-binding site of complex II (succinate-ubiquinone oxidoreductase): a mechanism of electron transfer and proton conduction during ubiquinone reduction, *J Biol Chem*, **281**, 7309-7316.
- Hou, S., Makarova, K.S., Saw, J.H., Senin, P., Ly, B.V., Zhou, Z., Ren, Y., Wang, J., Galperin, M.Y., Omelchenko, M.V., Wolf, Y.I., Yutin, N., Koonin, E.V., Stott, M.B., Mountain, B.W., Crowe, M.A., Smirnova, A.V., Dunfield, P.F., Feng, L., Wang, L. and Alam, M. (2008) Complete genome sequence of the extremely acidophilic methanotroph isolate V4, *Methylacidiphilum infernorum*, a representative of the bacterial phylum Verrucomicrobia, *Biol Direct*, **3**, 26.
- Huang, L.S., Cobessi, D., Tung, E.Y. and Berry, E.A. (2005) Binding of the respiratory chain inhibitor antimycin to the mitochondrial bc1 complex: a new crystal structure reveals an altered intramolecular hydrogen-bonding pattern, *J Mol Biol*, **351**, 573-597.
- Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: Visual molecular dynamics, *Journal of Molecular Graphics*, **14**, 33-38.
- Hunte, C., Zickermann, V. and Brandt, U. (2010) Functional modules and structural basis of conformational coupling in mitochondrial complex I, *Science*, **329**, 448-451.
- Hutchison, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O. and Venter, J.C. (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome, *Science*, **286**, 2165-2169.
- Huttemann, M., Pecina, P., Rainbolt, M., Sanderson, T.H., Kagan, V.E., Samavati, L., Doan, J.W. and Lee, I. (2011) The multiple functions of cytochrome c and their regulation in life and death decisions of the mammalian cell: From respiration to apoptosis, *Mitochondrion*, **11**, 369-381.
- Im, C.H., Hwang, S.M., Son, Y.S., Heo, J.B., Bang, W.Y., Suwastika, I.N., Shiina, T. and Bahk, J.D. (2011) Nuclear/nucleolar GTPase 2 proteins as a subfamily of Y1qF/YawG GTPases function in pre-60S ribosomal subunit maturation of mono- and dicotyledonous plants, *J Biol Chem*, **286**, 8620-8632.
- Inagaki, Y. and Ford Doolittle, W. (2000) Evolution of the eukaryotic translation termination system: origins of release factors, *Mol Biol Evol*, **17**, 882-889.
- Israelachvili, J.N., Mitchell, D.J. and Ninham, B.W. (1977) Theory of self-assembly of lipid bilayers and vesicles, *Biochim Biophys Acta*, **470**, 185-201.
- Ivanov, V. and Mizuuchi, K. (2010) Multiple modes of interconverting dynamic pattern formation by bacterial cell division proteins, *Proc Natl Acad Sci U S A*, **107**, 8071-8078.
- Ivey, D.M., Sturr, M.G., Krulwich, T.A. and Hicks, D.B. (1994) The abundance of atp gene transcript and of the membrane F1F0-ATPase as a function of the growth pH of alkaliphilic *Bacillus firmus* OF4, *J Bacteriol*, **176**, 5167-5170.
- Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. and Miyata, T. (1989) Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes, *Proc Natl Acad Sci U S A*, **86**, 9355-9359.
- Iwai, M., Takizawa, K., Tokutsu, R., Okamuro, A., Takahashi, Y. and Minagawa, J. (2010) Isolation of the elusive supercomplex that drives cyclic electron flow in photosynthesis, *Nature*, **464**, 1210-1213.
- Iyer, L.M., Leipe, D.D., Koonin, E.V. and Aravind, L. (2004) Evolutionary history and higher order classification of AAA+ ATPases, *J Struct Biol*, **146**, 11-31.
- Jain, N., Dhimole, N., Khan, A.R., De, D., Tomar, S.K., Sajish, M., Dutta, D., Parrack, P. and Prakash, B. (2009) *E. coli* HflX interacts with 50S ribosomal subunits in presence of nucleotides, *Biochem Biophys Res Commun*, **379**, 201-205.
- Jekely, G. (2006) Did the last common ancestor have a biological membrane?, *Biol Direct*, **1**, 35.



- Ji, J., Kline, A.E., Amoscato, A., Samhan-Arias, A.K., Sparvero, L.J., Tyurin, V.A., Tyurina, Y.Y., Fink, B., Manole, M.D., Puccio, A.M., Okonkwo, D.O., Cheng, J.P., Alexander, H., Clark, R.S., Kochanek, P.M., Wipf, P., Kagan, V.E. and Bayir, H. (2012) Lipidomics identifies cardiolipin oxidation as a mitochondrial target for redox therapy of brain injury, *Nat Neurosci*, **15**, 1407-1413.
- Joliot, P., Joliot, A. and Vermeglio, A. (2005) Fast oxidation of the primary electron acceptor under anaerobic conditions requires the organization of the photosynthetic chain of *Rhodospira rubra* in supercomplexes, *Biochim Biophys Acta*, **1706**, 204-214.
- Jordan, I.K., Kondrashov, F.A., Adzhubei, I.A., Wolf, Y.I., Koonin, E.V., Kondrashov, A.S. and Sunyaev, S. (2005) A universal trend of amino acid gain and loss in protein evolution, *Nature*, **433**, 633-638.
- Jormakka, M., Byrne, B. and Iwata, S. (2003) Protonmotive force generation by a redox loop mechanism, *FEBS Lett*, **545**, 25-30.
- Jormakka, M., Tornroth, S., Byrne, B. and Iwata, S. (2002) Molecular basis of proton motive force generation: structure of formate dehydrogenase-N, *Science*, **295**, 1863-1868.
- Kabsch, W., Mannherz, H.G., Suck, D., Pai, E.F. and Holmes, K.C. (1990) Atomic structure of the actin:DNase I complex, *Nature*, **347**, 37-44.
- Kagan, V.E., Bayir, H.A., Belikova, N.A., Kapralov, O., Tyurina, Y.Y., Tyurin, V.A., Jiang, J., Stoyanovsky, D.A., Wipf, P., Kochanek, P.M., Greenberger, J.S., Pitt, B., Shvedova, A.A. and Borisenko, G. (2009) Cytochrome c/cardiolipin relations in mitochondria: a kiss of death, *Free Radic Biol Med*, **46**, 1439-1453.
- Kagan, V.E., Tyurin, V.A., Jiang, J., Tyurina, Y.Y., Ritov, V.B., Amoscato, A.A., Osipov, A.N., Belikova, N.A., Kapralov, A.A., Kini, V., Vlasova, I.I., Zhao, Q., Zou, M., Di, P., Svistunenko, D.A., Kurnikov, I.V. and Borisenko, G.G. (2005) Cytochrome c acts as a cardiolipin oxygenase required for release of proapoptotic factors, *Nat Chem Biol*, **1**, 223-232.
- Kall, L., Krogh, A. and Sonnhammer, E.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server, *Nucleic Acids Res*, **35**, W429-432.
- Kamerlin, S.C., Sharma, P.K., Prasad, R.B. and Warshel, A. (2013) Why nature really chose phosphate, *Q Rev Biophys*, **46**, 1-132.
- Kamm, K.E. and Stull, J.T. (1985) The function of myosin and myosin light chain kinase phosphorylation in smooth muscle, *Annu Rev Pharmacol Toxicol*, **25**, 593-620.
- Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. and Hirakawa, M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs, *Nucleic Acids Res*, **38**, D355-360.
- Kato, M., Chuang, J.L., Tso, S.C., Wynn, R.M. and Chuang, D.T. (2005) Crystal structure of pyruvate dehydrogenase kinase 3 bound to lipoyl domain 2 of human pyruvate dehydrogenase complex, *EMBO J*, **24**, 1763-1774.
- Kelley, D.S., Karson, J.A., Fruh-Green, G.L., Yoerger, D.R., Shank, T.M., Butterfield, D.A., Hayes, J.M., Schrenk, M.O., Olson, E.J., Proskurowski, G., Jakuba, M., Bradley, A., Larson, B., Ludwig, K., Glickson, D., Buckman, K., Bradley, A.S., Brazelton, W.J., Roe, K., Elend, M.J., Delacour, A., Bernasconi, S.M., Lilley, M.D., Baross, J.A., Summons, R.E. and Sylva, S.P. (2005) A serpentinite-hosted ecosystem: the Lost City hydrothermal field, *Science*, **307**, 1428-1434.
- Kellosalo, J., Kajander, T., Kogan, K., Pokharel, K. and Goldman, A. (2012) The structure and catalytic cycle of a sodium-pumping pyrophosphatase, *Science*, **337**, 473-476.
- Kerr, J.F. (1965) A histochemical study of hypertrophy and ischaemic injury of rat liver with special reference to changes in lysosomes, *J Pathol Bacteriol*, **90**, 419-435.

- Kerr, J.F., Wyllie, A.H. and Currie, A.R. (1972) Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics, *Br J Cancer*, **26**, 239-257.
- Kleckner, N. (1977) Translocatable elements in procaryotes, *Cell*, **11**, 11-23.
- Kleinschroth, T., Castellani, M., Trinh, C.H., Morgner, N., Brutschy, B., Ludwig, B. and Hunte, C. (2011) X-ray structure of the dimeric cytochrome bc(1) complex from the soil bacterium *Paracoccus denitrificans* at 2.7-Å resolution, *Biochim Biophys Acta*, **1807**, 1606-1615.
- Kluck, R.M., Ellerby, L.M., Ellerby, H.M., Naiem, S., Yaffe, M.P., Margoliash, E., Bredesen, D., Mauk, A.G., Sherman, F. and Newmeyer, D.D. (2000) Determinants of cytochrome c pro-apoptotic activity. The role of lysine 72 trimethylation, *J Biol Chem*, **275**, 16127-16133.
- Kolbe, M., Besir, H., Essen, L.O. and Oesterhelt, D. (2000) Structure of the light-driven chloride pump halorhodopsin at 1.8 Å resolution, *Science*, **288**, 1390-1396.
- Komori, K., Miyata, T., DiRuggiero, J., Holley-Shanks, R., Hayashi, I., Cann, I.K., Mayanagi, K., Shinagawa, H. and Ishino, Y. (2000) Both RadA and RadB are involved in homologous recombination in *Pyrococcus furiosus*, *J Biol Chem*, **275**, 33782-33790.
- Komoriya, Y., Ariga, T., Iino, R., Imamura, H., Okuno, D. and Noji, H. (2012) Principal role of the arginine finger in rotary catalysis of F1-ATPase, *J Biol Chem*, **287**, 15134-15142.
- Koonin, E.V. (2000) How many genes can make a cell: the minimal-gene-set concept, *Annu Rev Genomics Hum Genet*, **1**, 99-116.
- Koonin, E.V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor, *Nat Rev Microbiol*, **1**, 127-136.
- Koonin, E.V. (2010) The origin and early evolution of eukaryotes in the light of phylogenomics, *Genome Biol*, **11**, 209.
- Koonin, E.V., Makarova, K.S. and Aravind, L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification, *Annu Rev Microbiol*, **55**, 709-742.
- Koonin, E.V., Mushegian, A.R. and Bork, P. (1996) Non-orthologous gene displacement, *Trends Genet*, **12**, 334-336.
- Koonin, E.V., Senkevich, T.G. and Dolja, V.V. (2006) The ancient Virus World and evolution of cells, *Biology direct*, **1**, 29.
- Koonin, E.V., Wolf, Y.I. and Aravind, L. (2000) Protein fold recognition using sequence profiles and its application in structural genomics, *Adv Protein Chem*, **54**, 245-275.
- Korshunov, S.S., Skulachev, V.P. and Starkov, A.A. (1997) High protonic potential actuates a mechanism of production of reactive oxygen species in mitochondria, *FEBS Lett*, **416**, 15-18.
- Krawiec, S. and Riley, M. (1990) Organization of the bacterial chromosome, *Microbiol Rev*, **54**, 502-539.
- Krebs, E.G. and Beavo, J.A. (1979) Phosphorylation-dephosphorylation of enzymes, *Annu Rev Biochem*, **48**, 923-959.
- Krishtalik, L.I. (1996) Intramembrane electron transfer: Processes in the photosynthetic reaction center, *Biochim Biophys Acta*, **1273**, 139-149.
- Krissinel, E. and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallographica*, **D60**, 2256-2268.
- Kritsky, M.S., Kolesnikov, M.P. and Telegina, T.A. (2007) Modeling of abiogenic synthesis of ATP, *Dokl Biochem Biophys*, **417**, 313-315.
- Kroemer, G. (1997) Mitochondrial implication in apoptosis. Towards an endosymbiont hypothesis of apoptosis evolution, *Cell Death Differ*, **4**, 443-456.
- Kroemer, G., Galluzzi, L. and Brenner, C. (2007) Mitochondrial membrane permeabilization in cell death, *Physiol Rev*, **87**, 99-163.
- Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J Mol Biol*, **305**, 567-580.

- Kumagai, H., Fujiwara, T., Matsubara, H. and Saeki, K. (1997) Membrane localization, topology, and mutual stabilization of the rnfABC gene products in *Rhodobacter capsulatus* and implications for a new family of energy-coupling NADH oxidoreductases, *Biochemistry*, **36**, 5509-5521.
- Kurusu, G., Zhang, H., Smith, J.L. and Cramer, W.A. (2003) Structure of the cytochrome b6f complex of oxygenic photosynthesis: tuning the cavity, *Science*, **302**, 1009-1014.
- Kushnareva, Y., Andreyev, A.Y., Kuwana, T. and Newmeyer, D.D. (2012) Bax activation initiates the assembly of a multimeric catalyst that facilitates Bax pore formation in mitochondrial outer membranes, *PLoS Biol*, **10**, e1001394.
- Kushnareva, Y., Murphy, A.N. and Andreyev, A. (2002) Complex I-mediated reactive oxygen species generation: modulation by cytochrome c and NAD(P)<sup>+</sup> oxidation-reduction state, *Biochem J*, **368**, 545-553.
- Kwon, S.K., Kim, B.K., Song, J.Y., Kwak, M.J., Lee, C.H., Yoon, J.H., Oh, T.K. and Kim, J.F. (2013) Genomic Makeup of the Marine Flavobacterium *Nonlabens* (*Donghaeana*) *dokdonensis* and Identification of a Novel Class of Rhodopsins, *Genome Biol Evol*, **5**, 187-199.
- Kyrpides, N., Overbeek, R. and Ouzounis, C. (1999) Universal protein families and the functional content of the last universal common ancestor, *J Mol Evol*, **49**, 413-423.
- Lancaster, C.R., Gross, R. and Simon, J. (2001) A third crystal form of *Wolinella succinogenes* quinol:fumarate reductase reveals domain closure at the site of fumarate reduction, *Eur J Biochem*, **268**, 1820-1827.
- Lancaster, C.R. and Kroger, A. (2000) Succinate: quinone oxidoreductases: new insights from X-ray crystal structures, *Biochim Biophys Acta*, **1459**, 422-431.
- Lancaster, C.R., Kroger, A., Auer, M. and Michel, H. (1999) Structure of fumarate reductase from *Wolinella succinogenes* at 2.2 Å resolution, *Nature*, **402**, 377-385.
- Lane, N., Allen, J.F. and Martin, W. (2010) How did LUCA make a living? Chemiosmosis in the origin of life, *Bioessays*, **32**, 271-280.
- Lane, N. and Martin, W.F. (2012) The origin of membrane bioenergetics, *Cell*, **151**, 1406-1416.
- Laskey, R.A. and Madine, M.A. (2003) A rotary pumping model for helicase function of MCM proteins at a distance from replication forks, *EMBO Rep*, **4**, 26-30.
- Lazcano, A. and Forterre, P. (1999) The molecular search for the last common ancestor, *J Mol Evol*, **49**, 411-412.
- Lazcano, A. and Miller, S.L. (1999) On the origin of metabolic pathways, *J Mol Evol*, **49**, 424-431.
- Lazebnik, Y.A., Kaufmann, S.H., Desnoyers, S., Poirier, G.G. and Earnshaw, W.C. (1994) Cleavage of poly(ADP-ribose) polymerase by a proteinase with properties like ICE, *Nature*, **371**, 346-347.
- Le, N.P., Omote, H., Wada, Y., Al-Shawi, M.K., Nakamoto, R.K. and Futai, M. (2000) *Escherichia coli* ATP synthase alpha subunit Arg-376: the catalytic site arginine does not participate in the hydrolysis/synthesis reaction but is required for promotion to the steady state, *Biochemistry*, **39**, 2778-2783.
- Le, S.Q. and Gascuel, O. (2008) An improved general amino acid replacement matrix, *Mol Biol Evol*, **25**, 1307-1320.
- Lebrun, E., Santini, J.M., Brugna, M., Ducluzeau, A.L., Ouchane, S., Schoepp-Cothenet, B., Baymann, F. and Nitschke, W. (2006) The Rieske protein: a case study on the pitfalls of multiple sequence alignments and phylogenetic reconstruction, *Mol Biol Evol*, **23**, 1180-1191.
- Lee, J. and Yang, W. (2006) UvrD Helicase Unwinds DNA One Base Pair at a Time by a Two-Part Power Stroke, *Cell*, **127**, 1349-1360.
- Lehninger, A.L. (1950) Role of metal ions in enzyme systems, *Physiol Rev*, **30**, 393-429.

- Leipe, D.D., Aravind, L. and Koonin, E.V. (1999) Did DNA replication evolve twice independently?, *Nucleic Acids Res*, **27**, 3389-3401.
- Leipe, D.D., Koonin, E.V. and Aravind, L. (2003) Evolution and classification of P-loop kinases and related proteins, *J Mol Biol*, **333**, 781-815.
- Leipe, D.D., Wolf, Y.I., Koonin, E.V. and Aravind, L. (2002) Classification and evolution of P-loop GTPases and related ATPases, *J Mol Biol*, **317**, 41-72.
- Lemos, R.S., Fernandes, A.S., Pereira, M.M., Gomes, C.M. and Teixeira, M. (2002) Quinol:fumarate oxidoreductases and succinate:quinone oxidoreductases: phylogenetic relationships, metal centres and membrane attachment, *Biochim Biophys Acta*, **1553**, 158-170.
- Li, Y., He, Y. and Luo, Y. (2009) Conservation of a conformational switch in RadA recombinase from *Methanococcus maripaludis*, *Acta Crystallogr D Biol Crystallogr*, **65**, 602-610.
- Lin, S.M., Tsai, J.Y., Hsiao, C.D., Huang, Y.T., Chiu, C.L., Liu, M.H., Tung, J.Y., Liu, T.H., Pan, R.L. and Sun, Y.J. (2012) Crystal structure of a membrane-embedded H<sup>+</sup>-translocating pyrophosphatase, *Nature*, **484**, 399-403.
- Lin, Z., Kong, H., Nei, M. and Ma, H. (2006) Origins and evolution of the recA/RAD51 gene family: evidence for ancient gene duplication and endosymbiotic gene transfer, *Proc Natl Acad Sci U S A*, **103**, 10328-10333.
- Liu, X., Kim, C.N., Yang, J., Jemmerson, R. and Wang, X. (1996) Induction of apoptotic program in cell-free extracts: requirement for dATP and cytochrome c, *Cell*, **86**, 147-157.
- Lockshin, R.A. and Williams, C.M. (1965) Programmed Cell Death--I. Cytology of Degeneration in the Intersegmental Muscles of the Pernyi Silkworm, *J Insect Physiol*, **11**, 123-133.
- Lokhmatikov, A.V., Voskoboinikova, N.E., Cherepanov, D.A., Steinhoff, H.J., Skulachev, V.P. and Mulikjanian, A.Y. (2012) Oxidation of cardiolipin in liposomes: A new insight into the primary steps of mitochondria-triggered apoptosis, *Biochim Biophys Acta*, **1817**, S100.
- Lowenstein, J.M. (1960) The stimulation of transphosphorylation by alkali-metal ions, *Biochem J*, **75**, 269-274.
- Ludwig, M. and Bryant, D.A. (2011) Transcription Profiling of the Model Cyanobacterium *Synechococcus* sp. Strain PCC 7002 by Next-Gen (SOLiD) Sequencing of cDNA, *Front Microbiol*, **2**, 41.
- Luoto, H.H., Belogurov, G.A., Baykov, A.A., Lahti, R. and Malinen, A.M. (2011) Na<sup>+</sup>-translocating membrane pyrophosphatases are widespread in the microbial world and evolutionarily precede H<sup>+</sup>-translocating pyrophosphatases, *J Biol Chem*, **286**, 21633-21642.
- Macallum, A.B. (1926) The paleochemistry of the body fluids and tissues, *Physiol Rev*, **6**, 316-357.
- Machius, M., Chuang, J.L., Wynn, R.M., Tomchick, D.R. and Chuang, D.T. (2001) Structure of rat BCKD kinase: nucleotide-induced domain communication in a mitochondrial protein kinase, *Proc Natl Acad Sci U S A*, **98**, 11218-11223.
- Madej, M.G., Muller, F.G., Ploch, J. and Lancaster, C.R. (2009) Limited reversibility of transmembrane proton transfer assisting transmembrane electron transfer in a dihaem-containing succinate:quinone oxidoreductase, *Biochim Biophys Acta*, **1787**, 593-600.
- Madej, M.G., Nasiri, H.R., Hilgendorff, N.S., Schwalbe, H. and Lancaster, C.R. (2006) Evidence for transmembrane proton transfer in a dihaem-containing membrane protein complex, *EMBO J*, **25**, 4963-4970.
- Maeshima, M. (2000) Vacuolar H<sup>(+)</sup>-pyrophosphatase, *Biochim Biophys Acta*, **1465**, 37-51.
- Mandel, M., Moriyama, Y., Hulmes, J.D., Pan, Y.C., Nelson, H. and Nelson, N. (1988) cDNA sequence encoding the 16-kDa proteolipid of chromaffin granules implies gene duplication in the evolution of H<sup>+</sup>-ATPases, *Proc Natl Acad Sci U S A*, **85**, 5521-5524.
- Mansy, S.S., Schrum, J.P., Krishnamurthy, M., Tobe, S., Treco, D.A. and Szostak, J.W. (2008) Template-directed synthesis of a genetic polymer in a model protocell, *Nature*, **454**, 122-125.

- Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences, *Science*, **285**, 751-753.
- Martin, W., Baross, J., Kelley, D. and Russell, M.J. (2008) Hydrothermal vents and the origin of life, *Nat Rev Microbiol*, **6**, 805-814.
- Martin, W. and Russell, M.J. (2003) On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells, *Philos Trans R Soc Lond B Biol Sci*, **358**, 59-83; discussion 83-55.
- Martin, W. and Russell, M.J. (2007) On the origin of biochemistry at an alkaline hydrothermal vent, *Philos Trans R Soc Lond B Biol Sci*, **362**, 1887-1925.
- Martinez, S.E., Huang, D., Szczepaniak, A., Cramer, W.A. and Smith, J.L. (1994) Crystal structure of chloroplast cytochrome f reveals a novel cytochrome fold and unexpected heme ligation, *Structure*, **2**, 95-105.
- Mathiesen, C. and Hagerhall, C. (2002) Transmembrane topology of the NuoL, M and N subunits of NADH:quinone oxidoreductase and their homologues among membrane-bound hydrogenases and bona fide antiporters, *Biochim Biophys Acta*, **1556**, 121-132.
- Mathiesen, C. and Hagerhall, C. (2003) The 'antiporter module' of respiratory chain complex I includes the MrpC/NuoK subunit -- a revision of the modular evolution scheme, *FEBS Lett*, **549**, 7-13.
- McInerney, M.J., Rohlin, L., Mouttaki, H., Kim, U., Krupp, R.S., Rios-Hernandez, L., Sieber, J., Struchtemeyer, C.G., Bhattacharyya, A., Campbell, J.W. and Gunsalus, R.P. (2007) The genome of *Syntrophus aciditrophicus*: life at the thermodynamic limit of microbial growth, *Proc Natl Acad Sci U S A*, **104**, 7600-7605.
- Meier, T. and Dimroth, P. (2002) Intersubunit bridging by Na<sup>+</sup> ions as a rationale for the unusual stability of the c-rings of Na<sup>+</sup>-translocating F1F0 ATP synthases, *EMBO Rep*, **3**, 1094-1098.
- Meier, T., Krahl, A., Bond, P.J., Pogoryelov, D., Diederichs, K. and Faraldo-Gomez, J.D. (2009) Complete ion-coordination structure in the rotor ring of Na<sup>+</sup>-dependent F-ATP synthases, *J Mol Biol*, **391**, 498-507.
- Meier, T., Morgner, N., Matthies, D., Pogoryelov, D., Keis, S., Cook, G.M., Dimroth, P. and Brutschy, B. (2007) A tridecameric c ring of the adenosine triphosphate (ATP) synthase from the thermoalkaliphilic *Bacillus* sp. strain TA2.A1 facilitates ATP synthesis at low electrochemical proton potential, *Mol Microbiol*, **65**, 1181-1192.
- Meier, T., Polzer, P., Diederichs, K., Welte, W. and Dimroth, P. (2005) Structure of the rotor ring of F-Type Na<sup>+</sup>-ATPase from *Ilyobacter tartaricus*, *Science*, **308**, 659-662.
- Meier, T., Yu, J., Raschle, T., Henzen, F., Dimroth, P. and Muller, D.J. (2005) Structural evidence for a constant c11 ring stoichiometry in the sodium F-ATP synthase, *FEBS J*, **272**, 5474-5483.
- Melchior, N.C. (1954) Sodium and potassium complexes of adenosinetriphosphate: equilibrium studies, *J Biol Chem*, **208**, 615-627.
- Melchior, N.C. and Melchior, J.B. (1958) The role of complex metal ions in the yeast hexokinase reaction, *J Biol Chem*, **231**, 609-623.
- Menz, R.I., Walker, J.E. and Leslie, A.G. (2001) Structure of bovine mitochondrial F(1)-ATPase with nucleotide bound to all three catalytic sites: implications for the mechanism of rotary catalysis, *Cell*, **106**, 331-341.
- Mereschkowsky, K. (1910) Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis, einer neuen Lehre von der Entstehung der Organismen, *Biol Centralbl*, **30**, 353-367.
- Meyer Zu Tittingdorf, J.M., Rexroth, S., Schafer, E., Schlichting, R., Giersch, C., Dencher, N.A. and Seelert, H. (2004) The stoichiometry of the chloroplast ATP synthase oligomer III in

- Chlamydomonas reinhardtii* is not affected by the metabolic state, *Biochim Biophys Acta*, **1659**, 92-99.
- Mileykovskaya, E., Penczek, P.A., Fang, J., Mallampalli, V.K., Sparagna, G.C. and Dowhan, W. (2012) Arrangement of the respiratory chain complexes in *Saccharomyces cerevisiae* supercomplex III<sub>2</sub>IV<sub>2</sub> revealed by single particle cryo-electron microscopy, *J Biol Chem*, **287**, 23095-23103.
- Miller, D.L. and Westheimer, F.H. (1965) Hydrolysis of Gamma-Phenylpropyl Di- and Triphosphates, *Science*, **148**, 667.
- Miller, M.J., Oldenburg, M. and Fillingame, R.H. (1990) The essential carboxyl group in subunit c of the F1F0 ATP synthase can be moved and H(+)-translocating function retained, *Proc Natl Acad Sci U S A*, **87**, 4900-4904.
- Miller, S.L. (1953) A production of amino acids under possible primitive earth conditions, *Science*, **117**, 528-529.
- Miller, S.L. and Cleaves, H.J. (2006) Prebiotic chemistry on the primitive Earth. In Rigoutsos, I. and Stephanopoulos, G. (eds), *Systems Biology*. Oxford University Press, Oxford, pp. 4-56.
- Milner-White, E.J., Coggins, J.R. and Anton, I.A. (1991) Evidence for an ancestral core structure in nucleotide-binding proteins with the type A motif, *J Mol Biol*, **221**, 751-754.
- Mishra, R., Gara, S.K., Mishra, S. and Prakash, B. (2005) Analysis of GTPases carrying hydrophobic amino acid substitutions in lieu of the catalytic glutamine: implications for GTP hydrolysis, *Proteins*, **59**, 332-338.
- Mitchell, P. (1961) Coupling of phosphorylation to electron and hydrogen transfer by a chemiosmotic type of mechanism, *Nature*, **191**, 144-148.
- Mitchell, P. (1966) Chemiosmotic coupling in oxidative and photosynthetic phosphorylation, *Biol Rev Camb Philos Soc*, **41**, 445-502.
- Mitchell, P. (1975) Protonmotive redox mechanism of the cytochrome b-c<sub>1</sub> complex in the respiratory chain: protonmotive ubiquinone cycle, *FEBS Lett*, **56**, 1-6.
- Mitchell, P. (1976) Possible molecular mechanisms of the protonmotive function of cytochrome systems, *J Theor Biol*, **62**, 327-367.
- Mitchell, P. (1984) Bacterial flagellar motors and osmoelectric molecular rotation by an axially transmembrane well and turnstile mechanism, *FEBS Lett*, **176**, 287-294.
- Miyamoto, S., Nantes, I.L., Faria, P.A., Cunha, D., Ronsein, G.E., Medeiros, M.H. and Di Mascio, P. (2012) Cytochrome c-promoted cardiolipin oxidation generates singlet molecular oxygen, *Photochem Photobiol Sci*, **11**, 1536-1546.
- Moll, R. and Schafer, G. (1991) Purification and characterisation of an archaeobacterial succinate dehydrogenase complex from the plasma membrane of the thermoacidophile *Sulfolobus acidocaldarius*, *Eur J Biochem*, **201**, 593-600.
- Moparthy, V.K. and Hagerhall, C. (2011) The evolution of respiratory chain complex I from a smaller last common ancestor consisting of 11 protein subunits, *J Mol Evol*, **72**, 484-497.
- Moran, L.A., Horton, R.A., Scrimgeour, G. and Perry, M. (2011) *Principles of Biochemistry (5th Edition)*. Pearson.
- Morii, H., Nishihara, M. and Koga, Y. (2000) CTP:2,3-di-O-geranylgeranyl-sn-glycero-1-phosphate cytidyltransferase in the methanogenic archaeon *Methanothermobacter thermoautotrophicus*, *J Biol Chem*, **275**, 36568-36574.
- Muench, S.P., Huss, M., Song, C.F., Phillips, C., Wieczorek, H., Trinick, J. and Harrison, M.A. (2009) Cryo-electron microscopy of the vacuolar ATPase motor reveals its mechanical and regulatory complexity, *J Mol Biol*, **386**, 989-999.
- Muench, S.P., Trinick, J. and Harrison, M.A. (2011) Structural divergence of the rotary ATPases, *Q Rev Biophys*, **44**, 311-356.

- Mulkidjanian, A. and Galperin, M. (2010) Evolutionary origins of membrane proteins. In Frishman, D. (ed), *Structural Bioinformatics of Membrane Proteins*. Springer, Vienna, pp. 1-28.
- Mulkidjanian, A. and Junge, W. (1997) On the origin of photosynthesis as inferred from sequence analysis, *Photosynth Res*, **51**, 27-42.
- Mulkidjanian, A.Y. (2005) Ubiquinol oxidation in the cytochrome bc<sub>1</sub> complex: reaction mechanism and prevention of short-circuiting, *Biochim Biophys Acta*, **1709**, 5-34.
- Mulkidjanian, A.Y. (2006) Proton in the well and through the desolvation barrier, *Biochim Biophys Acta*, **1757**, 415-427.
- Mulkidjanian, A.Y. (2007) Proton translocation by the cytochrome bc<sub>1</sub> complexes of phototrophic bacteria: introducing the activated Q-cycle, *Photochem Photobiol Sci*, **6**, 19-34.
- Mulkidjanian, A.Y. (2010) Activated Q-cycle as a common mechanism for cytochrome bc<sub>1</sub> and cytochrome b<sub>6</sub>f complexes, *Biochim Biophys Acta*, **1797**, 1858-1868.
- Mulkidjanian, A.Y., Bychkov, A.Y., Dibrova, D.V., Galperin, M.Y. and Koonin, E.V. (2012) Origin of first cells at terrestrial, anoxic geothermal fields, *Proc Natl Acad Sci U S A*.
- Mulkidjanian, A.Y., Cherepanov, D.A. and Galperin, M.Y. (2003) Survival of the fittest before the beginning of life: selection of the first oligonucleotide-like polymers by UV light, *Bmc Evol Biol*, **3**.
- Mulkidjanian, A.Y., Dibrov, P. and Galperin, M.Y. (2008a) The past and present of sodium energetics: may the sodium-motive force be with you, *Biochim Biophys Acta*, **1777**, 985-992.
- Mulkidjanian, A.Y. and Galperin, M. (2013) A Time to Scatter Genes and a Time to Gather Them: Evolution of Photosynthesis Genes in Bacteria. In Beatty, J.T. (ed), *Advances in Botanical Research: Genome Evolution of Photosynthetic Bacteria*. Elsevier, San Diego.
- Mulkidjanian, A.Y. and Galperin, M.Y. (2007) Physico-chemical and evolutionary constraints for the formation and selection of first biopolymers: Towards the consensus paradigm of the abiogenic origin of life, *Chemistry & Biodiversity*, **4**, 2003-2015.
- Mulkidjanian, A.Y. and Galperin, M.Y. (2009) On the origin of life in the zinc world. 2. Validation of the hypothesis on the photosynthesizing zinc sulfide edifices as cradles of life on Earth, *Biol Direct*, **4**, 27.
- Mulkidjanian, A.Y. and Galperin, M.Y. (2010) On the abundance of zinc in the evolutionarily old protein domains, *P Natl Acad Sci USA*, **107**, E137-E137.
- Mulkidjanian, A.Y., Galperin, M.Y. and Koonin, E.V. (2009) Co-evolution of primordial membranes and membrane proteins, *Trends Biochem Sci*, **34**, 206-215.
- Mulkidjanian, A.Y., Galperin, M.Y., Makarova, K.S., Wolf, Y.I. and Koonin, E.V. (2008b) Evolutionary primacy of sodium bioenergetics, *Biol Direct*, **3**, 13.
- Mulkidjanian, A.Y., Koonin, E.V., Makarova, K.S., Mekhedov, S.L., Sorokin, A., Wolf, Y.I., Dufresne, A., Partensky, F., Burd, H., Kaznadzey, D., Haselkorn, R. and Galperin, M.Y. (2006) The cyanobacterial genome core and the origin of photosynthesis, *P Natl Acad Sci USA*, **103**, 13126-13131.
- Mulkidjanian, A.Y., Makarova, K.S., Galperin, M.Y. and Koonin, E.V. (2007) Inventing the dynamo machine: the evolution of the F-type and V-type ATPases, *Nat Rev Microbiol*, **5**, 892-899.
- Muller, V. and Gruber, G. (2003) ATP synthases: structure, function and evolution of unique energy converters, *Cell Mol Life Sci*, **60**, 474-494.
- Muramoto, K., Hirata, K., Shinzawa-Itoh, K., Yoko-o, S., Yamashita, E., Aoyama, H., Tsukihara, T. and Yoshikawa, S. (2007) A histidine residue acting as a controlling site for dioxygen reduction and proton pumping by cytochrome c oxidase, *Proc Natl Acad Sci U S A*, **104**, 7881-7886.
- Muramoto, K., Ohta, K., Shinzawa-Itoh, K., Kanda, K., Taniguchi, M., Nabekura, H., Yamashita, E., Tsukihara, T. and Yoshikawa, S. (2010) Bovine cytochrome c oxidase structures enable

- O<sub>2</sub> reduction with minimization of reactive oxygens and provide a proton-pumping gate, *Proc Natl Acad Sci U S A*, **107**, 7740-7745.
- Murata, T., Yamato, I., Kakinuma, Y., Leslie, A.G. and Walker, J.E. (2005) Structure of the rotor of the V-Type Na<sup>+</sup>-ATPase from *Enterococcus hirae*, *Science*, **308**, 654-659.
- Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol*, **247**, 536-540.
- Mushegian, A. (2005) Protein content of minimal and ancestral ribosome, *RNA*, **11**, 1400-1406.
- Mushegian, A.R. and Koonin, E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes, *Proc Natl Acad Sci U S A*, **93**, 10268-10273.
- Nakamura, Y. and Ito, K. (1998) How protein reads the stop codon and terminates translation, *Genes Cells*, **3**, 265-278.
- Nakanishi-Matsui, M., Sekiya, M., Nakamoto, R.K. and Futai, M. (2010) The mechanism of rotating proton pumping ATPases, *Biochim Biophys Acta*, **1797**, 1343-1352.
- Nealson, K.H.R., R. (2003) Evolution of Metabolism. In Schlesinger, W.H. (ed), *Treatise on Geochemistry*. Elsevier.
- Needleman, S. and Wunsch, C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.*, **48**, 443-453.
- Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J.O., Deppenmeier, U. and Martin, W.F. (2012) Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea, *Proc Natl Acad Sci U S A*, **109**, 20537-20542.
- Nelson, D. and Cox, M. (2005) *Lehninger Principles of Biochemistry*. W.H. Freeman and Co., New York.
- Nelson, D.J. and Carter, C.E. (1969) Purification and characterization of Thymidine 5-monophosphate kinase from *Escherichia coli* B, *J Biol Chem*, **244**, 5254-5262.
- Neumann, S., Fuchs, A., Mulkidjanian, A. and Frishman, D. (2010) Current status of membrane protein structure classification, *Proteins-Structure Function and Bioinformatics*, **78**, 1760-1773.
- Neuwald, A.F., Aravind, L., Spouge, J.L. and Koonin, E.V. (1999) AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes, *Genome Res*, **9**, 27-43.
- Nishi, T. and Forgac, M. (2002) The vacuolar (H<sup>+</sup>)-ATPases--nature's most versatile proton pumps, *Nat Rev Mol Cell Biol*, **3**, 94-103.
- Nitschke, W., Liebl, U., Matsuura, K. and Kramer, D.M. (1995) Membrane-bound c-type cytochromes in *Heliobacillus mobilis*. In vivo study of the hemes involved in electron donation to the photosynthetic reaction center, *Biochemistry*, **34**, 11831-11839.
- Nitschke, W., van Lis, R., Schoepp-Cothenet, B. and Baymann, F. (2010) The "green" phylogenetic clade of Rieske/cytb complexes, *Photosynth Res*, **104**, 347-355.
- Nobes, C. and Hall, A. (1994) Regulation and function of the Rho subfamily of small GTPases, *Curr Opin Genet Dev*, **4**, 77-81.
- Noji, H., Yasuda, R., Yoshida, M. and Kinosita, K., Jr. (1997) Direct observation of the rotation of F<sub>1</sub>-ATPase, *Nature*, **386**, 299-302.
- Nomura, S.M., Yoshikawa, Y., Yoshikawa, K., Dannenmuller, O., Chasserot-Golaz, S., Ourisson, G. and Nakatani, Y. (2001) Towards proto-cells: "primitive" lipid vesicles encapsulating giant DNA and its histone complex, *ChemBiochem*, **2**, 457-459.
- Nury, H., Dahout-Gonzalez, C., Trezeguet, V., Lauquin, G., Brandolin, G. and Pebay-Peyroula, E. (2005) Structural basis for lipid-mediated interactions between mitochondrial ADP/ATP carrier monomers, *FEBS Lett*, **579**, 6031-6036.



- O'Brien, M.C. and McKay, D.B. (1995) How potassium affects the activity of the molecular chaperone Hsc70. I. Potassium is required for optimal ATPase activity, *J Biol Chem*, **270**, 2247-2250.
- Ochiai, E.I. (1978) The evolution of the environment and its influence on the evolution of life, *Orig Life*, **9**, 81-91.
- Ochiai, E.I. (1983) Copper and the biological evolution, *Biosystems*, **16**, 81-86.
- Oesper, P. (1950) Sources of the high energy content in energy-rich phosphates, *Arch Biochem*, **27**, 255-270.
- Oesterhelt, D. and Stoeckenius, W. (1973) Functions of a new photoreceptor membrane, *Proc Natl Acad Sci U S A*, **70**, 2853-2857.
- Ohta, K., Muramoto, K., Shinzawa-Itoh, K., Yamashita, E., Yoshikawa, S. and Tsukihara, T. (2010) X-ray structure of the NO-bound Cu(B) in bovine cytochrome c oxidase, *Acta Crystallogr Sect F Struct Biol Cryst Commun*, **66**, 251-253.
- Olsen, G.J., Woese, C.R. and Overbeek, R. (1994) The winds of (evolutionary) change: breathing new life into microbiology, *J Bacteriol*, **176**, 1-6.
- Oparin, A.I. (1924) *The Origin of Life*. Moskowskiy rabochiy, Moscow.
- Oparin, A.I. (1957) *The origin of life on the Earth*. Academic Press, Inc, New York.
- Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH--a hierarchic classification of protein domain structures, *Structure*, **5**, 1093-1108.
- Orgel, L.E. (2008) The implausibility of metabolic cycles on the prebiotic Earth, *PLoS Biol*, **6**, e18.
- Ostermann, J., Horwich, A.L., Neupert, W. and Hartl, F.U. (1989) Protein folding in mitochondria requires complex formation with hsp60 and ATP hydrolysis, *Nature*, **341**, 125-130.
- Page, M.J. and Di Cera, E. (2006) Role of Na<sup>+</sup> and K<sup>+</sup> in enzyme function, *Physiol Rev*, **86**, 1049-1092.
- Pallen, M.J., Bailey, C.M. and Beatson, S.A. (2006) Evolutionary links between FliH/YscL-like proteins from bacterial type III secretion systems and second-stalk components of the FoF1 and vacuolar ATPases, *Protein Sci*, **15**, 935-941.
- Palsdottir, H. and Hunte, C. (2004) Lipids in membrane protein structures, *Biochim Biophys Acta*, **1666**, 2-18.
- Pandelia, M.E., Lubitz, W. and Nitschke, W. (2012) Evolution and diversification of Group 1 [NiFe] hydrogenases. Is there a phylogenetic marker for O<sub>2</sub>-tolerance?, *Biochim Biophys Acta*, **1817**, 1565-1575.
- Pascal, R. and Boiteau, L. (2011) Energy flows, metabolism and translation, *Philos Trans R Soc Lond B Biol Sci*, **366**, 2949-2958.
- Pasek, M.A., Kee, T.P., Bryant, D.E., Pavlov, A.A. and Lunine, J.I. (2008) Production of potentially prebiotic condensed phosphates by phosphorus redox chemistry, *Angew Chem Int Ed Engl*, **47**, 7918-7920.
- Patel, S.S. and Picha, K.M. (2000) Structure and function of hexameric helicases, *Annu Rev Biochem*, **69**, 651-697.
- Pebay-Peyroula, E., Dahout-Gonzalez, C., Kahn, R., Trezeguet, V., Lauquin, G.J. and Brandolin, G. (2003) Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside, *Nature*, **426**, 39-44.
- Pereira, M.M. and Teixeira, M. (2003) Is a Q-cycle-like mechanism operative in dihaemic succinate:quinone and quinol:fumarate oxidoreductases?, *FEBS Lett*, **543**, 1-4.
- Pereto, J., Lopez-Garcia, P. and Moreira, D. (2004) Ancestral lipid biosynthesis and early membrane evolution, *Trends Biochem Sci*, **29**, 469-477.

- Perier, C., Bove, J. and Vila, M. (2012) Mitochondria and programmed cell death in Parkinson's disease: apoptosis and beyond, *Antioxid Redox Signal*, **16**, 883-895.
- Petit, P.X., Susin, S.A., Zamzami, N., Mignotte, B. and Kroemer, G. (1996) Mitochondria and programmed cell death: back to the future, *FEBS Lett*, **396**, 7-13.
- Pettersson, E., Lundeberg, J. and Ahmadian, A. (2009) Generations of sequencing technologies, *Genomics*, **93**, 105-111.
- Pettigrew, G.W. and Moore, G.R. (1987) *Cytochrome c - Biological aspects*. Springer-Verlag, Berlin.
- Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. and Delsuc, F. (2005) Heterotachy and long-branch attraction in phylogenetics, *Bmc Evol Biol*, **5**, 50.
- Phipps, B.M., Typke, D., Hegerl, R., Volker, S., Hoffmann, A., Stetter, K.O. and Baumeister, W. (1993) Structure of a molecular chaperone from a thermophilic archaeobacterium, *Nature*, **361**, 475 - 477.
- Pierre, Y., Breyton, C., Lemoine, Y., Robert, B., Vernotte, C. and Popot, J.L. (1997) On the presence and role of a molecule of chlorophyll a in the cytochrome b6 f complex, *J.Biol Chem.*, **272**, 21901-21908.
- Pinti, D.L. (2005) The origin and evolution of the oceans. In Gargaud, M., *et al.* (eds), *Lectures in Astrobiology*. Springer-Verlag, Berlin, pp. 83-111.
- Pisa, K.Y., Weidner, C., Maischak, H., Kavermann, H. and Muller, V. (2007) The coupling ion in the methanoarchaeal ATP synthases: H(+) vs. Na(+) in the A(1)A(o) ATP synthase from the archaeon *Methanosarcina mazei* Go1, *FEMS Microbiol Lett*, **277**, 56-63.
- Pogoryelov, D., Krah, A., Langer, J.D., Yildiz, O., Faraldo-Gomez, J.D. and Meier, T. (2010) Microscopic rotary mechanism of ion translocation in the F(o) complex of ATP synthases, *Nat Chem Biol*, **6**, 891-899.
- Pogoryelov, D., Reichen, C., Klyszejko, A.L., Brunisholz, R., Muller, D.J., Dimroth, P. and Meier, T. (2007) The oligomeric state of c rings from cyanobacterial F-ATP synthases varies from 13 to 15, *J Bacteriol*, **189**, 5895-5902.
- Pogoryelov, D., Sudhir, P.R., Kovacs, L., Gombos, Z., Brown, I. and Garab, G. (2003) Sodium dependency of the photosynthetic electron transport in the alkaliphilic cyanobacterium *Arthrospira platensis*, *J Bioenerg Biomembr*, **35**, 427-437.
- Posfai, J., Bhagwat, A.S., Posfai, G. and Roberts, R.J. (1989) Predictive motifs derived from cytosine methyltransferases, *Nucleic Acids Res*, **17**, 2421-2435.
- Powner, M.W., Gerland, B. and Sutherland, J.D. (2009) Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions, *Nature*, **459**, 239-242.
- Pratt, D.A., Tallman, K.A. and Porter, N.A. (2011) Free radical oxidation of polyunsaturated lipids: New mechanistic insights and the development of peroxy radical clocks, *Acc Chem Res*, **44**, 458-467.
- Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res*, **35**, D61-65.
- Putnam, C.D., Clancy, S.B., Tsuruta, H., Gonzalez, S., Wetmur, J.G. and Tainer, J.A. (2001) Structure and mechanism of the RuvB Holliday junction branch migration motor, *J Mol Biol*, **311**, 297-310.
- Qian, X., He, Y., Wu, Y. and Luo, Y. (2006) Asp302 determines potassium dependence of a RadA recombinase from *Methanococcus voltae*, *J Mol Biol*, **360**, 537-547.
- Rahlf, S. and Muller, V. (1997) Sequence of subunit c of the Na(+)-translocating F1F0 ATPase of *Acetobacterium woodii*: proposal for determinants of Na+ specificity as revealed by sequence comparisons, *FEBS Lett*, **404**, 269-271.

- Ranson, N.A., White, H.E. and Saibil, H.R. (1998) Chaperonins, *Biochem J*, **333** ( Pt 2), 233-242.
- Rastogi, V.K. and Girvin, M.E. (1999) Structural changes linked to proton translocation by subunit c of the ATP synthase, *Nature*, **402**, 263-268.
- Raymond, J. and Blankenship, R.E. (2004) The evolutionary development of the protein complement of photosystem 2, *Biochim Biophys Acta*, **1655**, 133-139.
- Rees, D.M., Leslie, A.G. and Walker, J.E. (2009) The structure of the membrane extrinsic region of bovine ATP synthase, *Proc Natl Acad Sci U S A*, **106**, 21597-21601.
- Reubold, T.F., Wohlgenuth, S. and Eschenburg, S. (2011) Crystal structure of full-length Apaf-1: how the death signal is relayed in the mitochondrial pathway of apoptosis, *Structure*, **19**, 1074-1083.
- Roberts, P.G. and Hirst, J. (2012) The Deactive Form of Respiratory Complex I from Mammalian Mitochondria is a Na<sup>+</sup>/H<sup>+</sup> Antiporter, *J Biol Chem*.
- Rodriguez, A., Oliver, H., Zou, H., Chen, P., Wang, X. and Abrams, J.M. (1999) Dark is a Drosophila homologue of Apaf-1/CED-4 and functions in an evolutionarily conserved death pathway, *Nat Cell Biol*, **1**, 272-279.
- Rojas, N.R., Kamtekar, S., Simons, C.T., McLean, J.E., Vogel, K.M., Spiro, T.G., Farid, R.S. and Hecht, M.H. (1997) De novo heme proteins from designed combinatorial libraries, *Protein Sci*, **6**, 2512-2524.
- Rutherford, A.W., Osyczka, A. and Rappaport, F. (2012) Back-reactions, short-circuits, leaks and other energy wasteful reactions in biological electron transfer: redox tuning to survive life in O(2), *FEBS Lett*, **586**, 603-616.
- Sagan, L. (1967) On the origin of mitosing cells, *J Theor Biol*, **14**, 255-274.
- Saitou, N. and Imanishi, T. (1989) Relative Efficiencies of the Fitch-Margoliash, Maximum-Parsimony, Maximum-Likelihood, Minimum-Evolution, and Neighbor-joining Methods of Phylogenetic Tree Construction in Obtaining the Correct Tree, *Mol. Biol. Evol.*, **6**, 514-525.
- Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol Biol Evol*, **4**, 406-425.
- Saladino, R., Botta, G., Pino, S., Costanzo, G. and Di Mauro, E. (2012a) From the one-carbon amide formamide to RNA all the steps are prebiotically possible, *Biochimie*, **94**, 1451-1456.
- Saladino, R., Botta, G., Pino, S., Costanzo, G. and Di Mauro, E. (2012b) Genetics first or metabolism first? The formamide clue, *Chem Soc Rev*, **41**, 5526-5565.
- Saladino, R., Crestini, C., Ciciriello, F., Di Mauro, E. and Costanzo, G. (2006) Origin of informational polymers: differential stability of phosphoester bonds in ribomonomers and ribooligomers, *J Biol Chem*, **281**, 5790-5796.
- Sandler, S.J., Hugenholtz, P., Schleper, C., DeLong, E.F., Pace, N.R. and Clark, A.J. (1999) Diversity of radA genes from cultured and uncultured archaea: comparative analysis of putative RadA proteins and their use as a phylogenetic marker, *J Bacteriol*, **181**, 907-915.
- Saraste, M., Sibbald, P.R. and Wittinghofer, A. (1990) The P-loop--a common motif in ATP- and GTP-binding proteins, *Trends Biochem Sci*, **15**, 430-434.
- Sattley, W.M., Madigan, M.T., Swingley, W.D., Cheung, P.C., Clocksin, K.M., Conrad, A.L., Dejesa, L.C., Honchak, B.M., Jung, D.O., Karbach, L.E., Kurdoglu, A., Lahiri, S., Mastrian, S.D., Page, L.E., Taylor, H.L., Wang, Z.T., Raymond, J., Chen, M., Blankenship, R.E. and Touchman, J.W. (2008) The genome of *Heliobacterium modesticaldum*, a phototrophic representative of the Firmicutes containing the simplest photosynthetic apparatus, *J Bacteriol*, **190**, 4687-4696.
- Saum, R., Schlegel, K., Meyer, B. and Muller, V. (2009) The FIFO ATP synthase genes in *Methanosarcina acetivorans* are dispensable for growth and ATP synthesis, *FEMS Microbiol Lett*, **300**, 230-236.

- Sayle, R.A. and Milner-White, E.J. (1995) RASMOL: biomolecular graphics for all, *Trends Biochem Sci*, **20**, 374.
- Sazanov, L.A. and Hinchliffe, P. (2006) Structure of the hydrophilic domain of respiratory complex I from *Thermus thermophilus*, *Science*, **311**, 1430-1436.
- Schafer, G. (1996) Bioenergetics of the archaeobacterium *Sulfolobus*, *Biochim Biophys Acta*, **1277**, 163-200.
- Scheffzek, K., Ahmadian, M.R., Kabsch, W., Wiesmuller, L., Lautwein, A., Schmitz, F. and Wittinghofer, A. (1997) The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants, *Science*, **277**, 333-338.
- Schmehl, M., Jahn, A., Meyer zu Vilsendorf, A., Hennecke, S., Masepohl, B., Schuppler, M., Marxer, M., Oelze, J. and Klipp, W. (1993) Identification of a new class of nitrogen fixation genes in *Rhodobacter capsulatus*: a putative membrane complex involved in electron transport to nitrogenase, *Mol Gen Genet*, **241**, 602-615.
- Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences, *Nucleic Acids Res*, **18**, 6097-6100.
- Schoepp-Cothenet, B., Lieutaud, C., Baymann, F., Vermeglio, A., Friedrich, T., Kramer, D.M. and Nitschke, W. (2009) Menaquinone as pool quinone in a purple bacterium, *Proc Natl Acad Sci U S A*, **106**, 8549-8554.
- Schoepp-Cothenet, B., van Lis, R., Atteia, A., Baymann, F., Capowicz, L., Ducluzeau, A.L., Duval, S., Ten Brink, F., Russell, M.J. and Nitschke, W. (2013) On the universal core of bioenergetics, *Biochim Biophys Acta*, **1827**, 79-93.
- Schoffstall, A. (1976) Prebiotic phosphorylation of nucleosides in formamide, *Origins of Life*, **7**, 399-412.
- Schoonen, M., Smirnov, A. and Cohn, C. (2004) A perspective on the role of minerals in prebiotic synthesis, *Ambio*, **33**, 539-551.
- Schutz, M., Brugna, M., Lebrun, E., Baymann, F., Huber, R., Stetter, K.O., Hauska, G., Toci, R., Lemesle-Meunier, D., Tron, P., Schmidt, C. and Nitschke, W. (2000) Early evolution of cytochrome bc complexes, *J Mol Biol*, **300**, 663-675.
- Schütz, M., Zirngibl, S., Coutre, J., Büttner, M., Xie, D.-L., Nelson, N., Deutzmann, R. and Hauska, G. (1994) A transcription unit for the Rieske FeS-protein and cytochrome b in *Chlorobium limicola*, *Photosynth Res*, **39**, 163-174.
- Schweins, T. and Wittinghofer, A. (1994) GTP-binding proteins. Structures, interactions and relationships, *Curr Biol*, **4**, 547-550.
- Scrima, A. and Wittinghofer, A. (2006) Dimerisation-dependent GTPase reaction of MnmE: how potassium acts as GTPase-activating element, *EMBO J*, **25**, 2940-2951.
- Senior, A.E., Nadanaciva, S. and Weber, J. (2000) Rate acceleration of ATP hydrolysis by F(1)F(o)-ATP synthase, *J Exp Biol*, **203**, 35-40.
- Serrano-Andrés, L. and Merchán, M. (2009) Are the five natural DNA/RNA base monomers a good choice from natural selection?: A photochemical perspective, *Journal of Photochemistry and Photobiology C: Photochemistry Reviews*, **10**, 21-32.
- Setubal, J.C., dos Santos, P., Goldman, B.S., Ertesvag, H., Espin, G., Rubio, L.M., Valla, S., Almeida, N.F., Balasubramanian, D., Cromes, L., Curatti, L., Du, Z., Godsy, E., Goodner, B., Hellner-Burris, K., Hernandez, J.A., Houmiel, K., Imperial, J., Kennedy, C., Larson, T.J., Latreille, P., Ligon, L.S., Lu, J., Maerk, M., Miller, N.M., Norton, S., O'Carroll, I.P., Paulsen, I., Raulfs, E.C., Roemer, R., Rosser, J., Segura, D., Slater, S., Stricklin, S.L., Studholme, D.J., Sun, J., Viana, C.J., Wallin, E., Wang, B., Wheeler, C., Zhu, H., Dean, D.R., Dixon, R. and Wood, D. (2009) Genome sequence of *Azotobacter vinelandii*, an obligate aerobe specialized to support diverse anaerobic metabolic processes, *J Bacteriol*, **191**, 4534-4545.

- Sharonov, G.V., Feofanov, A.V., Bocharova, O.V., Astapova, M.V., Dedukhova, V.I., Chernyak, B.V., Dolgikh, D.A., Arseniev, A.S., Skulachev, V.P. and Kirpichnikov, M.P. (2005) Comparative analysis of proapoptotic activity of cytochrome c mutants in living cells, *Apoptosis*, **10**, 797-808.
- Shimizu, S., Narita, M. and Tsujimoto, Y. (1999) Bcl-2 family proteins regulate the release of apoptogenic cytochrome c by the mitochondrial channel VDAC, *Nature*, **399**, 483-487.
- Shinzawa-Itoh, K., Aoyama, H., Muramoto, K., Terada, H., Kurauchi, T., Tadehara, Y., Yamasaki, A., Sugimura, T., Kurono, S., Tsujimoto, K., Mizushima, T., Yamashita, E., Tsukihara, T. and Yoshikawa, S. (2007) Structures and physiological roles of 13 integral lipids of bovine heart cytochrome c oxidase, *EMBO J*, **26**, 1713-1725.
- Siegel, R.M., Frederiksen, J.K., Zacharias, D.A., Chan, F.K., Johnson, M., Lynch, D., Tsien, R.Y. and Lenardo, M.J. (2000) Fas preassociation required for apoptosis signaling and dominant inhibition by pathogenic mutations, *Science*, **288**, 2354-2357.
- Sies, H. (1993) Vitamin E in biologischen Systemen: Die Biochemie von Tocopheroxyl-Radikalen und ihre Rolle bei Krankheiten im Menschen. In Schmidt, K. and Wildmeiste, W. (eds), *Vitamin E in der modernen Medizin*. MKM Verlagsgesellschaft, Lenggies/Obb., pp. 11-20.
- Singleton, M.R., Dillingham, M.S. and Wigley, D.B. (2007) Structure and mechanism of helicases and nucleic acid translocases, *Annu Rev Biochem*, **76**, 23-50.
- Skordalakes, E. and Berger, J.M. (2003) Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading, *Cell*, **114**, 135-146.
- Skordalakes, E. and Berger, J.M. (2006) Structural insights into RNA-dependent ring closure and ATPase activation by the Rho termination factor, *Cell*, **127**, 553-564.
- Skulachev, V. (1988) *Membrane Bioenergetics*. Springer-Verlag, Berlin.
- Skulachev, V.P. (1972) Solution of the problem of energy coupling in terms of chemiosmotic theory, *J Bioenerg*, **3**, 25-38.
- Skulachev, V.P. (1996) Why are mitochondria involved in apoptosis? Permeability transition pores and apoptosis as selective mechanisms to eliminate superoxide-producing mitochondria and cell, *FEBS Lett*, **397**, 7-10.
- Skulachev, V.P. (1998) Cytochrome c in the apoptotic and antioxidant cascades, *FEBS Lett*, **423**, 275-280.
- Skulachev, V.P. (2006) Bioenergetic aspects of apoptosis, necrosis and mitoptosis, *Apoptosis*, **11**, 473-485.
- Skulachev, V.P. (2007) A biochemical approach to the problem of aging: "megaproject" on membrane-penetrating ions. The first results and prospects, *Biochemistry (Mosc)*, **72**, 1385-1396.
- Skulachev, V.P., Anisimov, V.N., Antonenko, Y.N., Bakeeva, L.E., Chernyak, B.V., Elichev, V.P., Filenko, O.F., Kalinina, N.I., Kapelko, V.I., Kolosova, N.G., Kopnin, B.P., Korshunova, G.A., Lichinitser, M.R., Obukhova, L.A., Pasyukova, E.G., Pisarenko, O.I., Roginsky, V.A., Ruuge, E.K., Senin, II, Severina, II, Skulachev, M.V., Spivak, I.M., Tashlitsky, V.N., Tkachuk, V.A., Vyssokikh, M.Y., Yaguzhinsky, L.S. and Zorov, D.B. (2009) An attempt to prevent senescence: a mitochondrial approach, *Biochim Biophys Acta*, **1787**, 437-461.
- Sleep, N.H., Meibom, A., Fridriksson, T., Coleman, R.G. and Bird, D.K. (2004) H<sub>2</sub>-rich fluids from serpentinization: geochemical and biotic implications, *Proc Natl Acad Sci U S A*, **101**, 12818-12823.
- Slesarev, A.I., Mezhevaya, K.V., Makarova, K.S., Polushin, N.N., Shcherbinina, O.V., Shakhova, V.V., Belova, G.I., Aravind, L., Natale, D.A., Rogozin, I.B., Tatusov, R.L., Wolf, Y.I., Stetter, K.O., Malykh, A.G., Koonin, E.V. and Kozyavkin, S.A. (2002) The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens, *Proc Natl Acad Sci U S A*, **99**, 4644-4649.

- Smith, J.M., Dowson, C.G. and Spratt, B.G. (1991) Localized sex in bacteria, *Nature*, **349**, 29-31.
- Smith, T. and Waterman, M. (1981) Comparison of biosequences, *Advances in Applied Mathematic*, **2**, 482-489.
- Sobolewski, A.L. and Domcke, W. (2006) The chemical physics of the photostability of life, *Europhysics News*, **37**, 20-23.
- Solmaz, S.R. and Hunte, C. (2008) Structure of complex III with bound cytochrome c in reduced state and definition of a minimal core interface for electron transfer, *J Biol Chem*, **283**, 17542-17549.
- Soontharapirakkul, K., Promden, W., Yamada, N., Kageyama, H., Incharoensakdi, A., Iwamoto-Kihara, A. and Takabe, T. (2011) Halotolerant cyanobacterium *Aphanothece halophytica* contains an Na<sup>+</sup>-dependent F1F0-ATP synthase with a potential role in salt-stress tolerance, *J Biol Chem*, **286**, 10169-10176.
- Sorokin, D.Y., Lucker, S., Vejmekova, D., Kostrikina, N.A., Kleerebezem, R., Rijpstra, W.I., Damste, J.S., Le Paslier, D., Muyzer, G., Wagner, M., van Loosdrecht, M.C. and Daims, H. (2012) Nitrification expanded: discovery, physiology and genomics of a nitrite-oxidizing bacterium from the phylum Chloroflexi, *ISME J*, **6**, 2245-2256.
- Sousa, F.L., Alves, R.J., Pereira-Leal, J.B., Teixeira, M. and Pereira, M.M. (2011) A bioinformatics classifier and database for heme-copper oxygen reductases, *PLoS One*, **6**, e19117.
- Spirin, A.S. (1960) On macromolecular structure of native high-polymer ribonucleic acid in solution, *J. Mol. Biol.*, **2**, 436-446.
- Spirin, A.S. (2005) The RNA World and Its Evolution, *Molecular Biology (Moscow)*, **39**, 466-472.
- Stanier, R.Y. and Van Niel, C.B. (1962) The concept of a bacterium, *Arch Mikrobiol*, **42**, 17-35.
- Stark, B.C., Kole, R., Bowman, E.J. and Altman, S. (1978) Ribonuclease P: an enzyme with an essential RNA component, *Proc Natl Acad Sci U S A*, **75**, 3717-3721.
- Steitz, T.A. and Moore, P.B. (2003) RNA, the first macromolecular catalyst: the ribosome is a ribozyme, *Trends Biochem Sci*, **28**, 411-418.
- Stock, D., Leslie, A.G. and Walker, J.E. (1999) Molecular architecture of the rotary motor in ATP synthase, *Science*, **286**, 1700-1705.
- Storbeck, S., Rolfes, S., Raux-Deery, E., Warren, M.J., Jahn, D. and Layer, G. (2010) A novel pathway for the biosynthesis of heme in Archaea: genome-based bioinformatic predictions and experimental evidence, *Archaea*, **2010**, 175050.
- Sträter, N., Lipscomb, W., Klabunde, T. and Krebs, B. (1996) Two-Metal Ion Catalysis in Enzymatic Acyl- and Phosphoryl-Transfer Reactions, *Angewandte Chemie International Edition in English*, **35**, 2024-2055.
- Stroebel, D., Choquet, Y., Popot, J.L. and Picot, D. (2003) An atypical haem in the cytochrome b(6)f complex, *Nature*, **426**, 413-418.
- Suga, M., Yano, N., Muramoto, K., Shinzawa-Itoh, K., Maeda, T., Yamashita, E., Tsukihara, T. and Yoshikawa, S. (2011) Distinguishing between Cl<sup>-</sup> and O<sub>2</sub>(<sup>2-</sup>) as the bridging element between Fe<sup>3+</sup> and Cu<sup>2+</sup> in resting-oxidized cytochrome c oxidase, *Acta Crystallogr D Biol Crystallogr*, **67**, 742-744.
- Sumi, M., Yohda, M., Koga, Y. and Yoshida, M. (1997) F0F1-ATPase genes from an archaeobacterium, *Methanosarcina barkeri*, *Biochem Biophys Res Commun*, **241**, 427-433.
- Sun, F., Huo, X., Zhai, Y., Wang, A., Xu, J., Su, D., Bartlam, M. and Rao, Z. (2005) Crystal structure of mitochondrial respiratory membrane protein complex II, *Cell*, **121**, 1043-1057.
- Suzuki, T., Ozaki, Y., Sone, N., Feniouk, B.A. and Yoshida, M. (2007) The product of uncI gene in F1Fo-ATP synthase operon plays a chaperone-like role to assist c-ring assembly, *Proc Natl Acad Sci U S A*, **104**, 20776-20781.

- Svensson-Ek, M., Abramson, J., Larsson, G., Tornroth, S., Brzezinski, P. and Iwata, S. (2002) The X-ray crystal structures of wild-type and EQ(I-286) mutant cytochrome c oxidases from *Rhodobacter sphaeroides*, *J Mol Biol*, **321**, 329-339.
- Swartz, T.H., Ikewada, S., Ishikawa, O., Ito, M. and Krulwich, T.A. (2005) The Mrp system: a giant among monovalent cation/proton antiporters?, *Extremophiles*, **9**, 345-354.
- Swerdlow, R.H. (2012) Mitochondria and cell bioenergetics: increasingly recognized components and a possible etiologic cause of Alzheimer's disease, *Antioxid Redox Signal*, **16**, 1434-1455.
- Swierczek, M., Cieluch, E., Sarewicz, M., Borek, A., Moser, C.C., Dutton, P.L. and Osyczka, A. (2010) An electronic bus bar lies in the core of cytochrome bc<sub>1</sub>, *Science*, **329**, 451-454.
- Swingle, W.D., Chen, M., Cheung, P.C., Conrad, A.L., Dejesa, L.C., Hao, J., Honchak, B.M., Karbach, L.E., Kurdoglu, A., Lahiri, S., Mastrian, S.D., Miyashita, H., Page, L., Ramakrishna, P., Satoh, S., Sattley, W.M., Shimada, Y., Taylor, H.L., Tomo, T., Tsuchiya, T., Wang, Z.T., Raymond, J., Mimuro, M., Blankenship, R.E. and Touchman, J.W. (2008) Niche adaptation and genome expansion in the chlorophyll d-producing cyanobacterium *Acaryochloris marina*, *Proc Natl Acad Sci U S A*, **105**, 2005-2010.
- Szathmary, E. (2007) Coevolution of metabolic networks and membranes: the scenario of progressive sequestration, *Philos Trans R Soc Lond B Biol Sci*, **362**, 1781-1787.
- Szostak, J.W., Bartel, D.P. and Luisi, P.L. (2001) Synthesizing life, *Nature*, **409**, 387-390.
- Szostak, J.W. and Ricardo, A. (2009) Origin of Life on Earth, *Scientific American*, **301**, 54 - 61.
- Szymanska, R., Dlużewska, J., Slesak, I. and Kruk, J. (2011) Ferredoxin:NADP<sup>+</sup> oxidoreductase bound to cytochrome b(6)f complex is active in plastoquinone reduction: implications for cyclic electron transport, *Physiol Plant*, **141**, 289-298.
- Tait, S.W. and Green, D.R. (2010) Mitochondria and cell death: outer membrane permeabilization and beyond, *Nat Rev Mol Cell Biol*, **11**, 621-632.
- Talavera, G. and Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments, *Syst Biol*, **56**, 564-577.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, *Mol Biol Evol*, **28**, 2731-2739.
- Taran, Y.A., Varley, N.R., Inguaggiato, S. and Cienfuegos, E. (2010) Geochemistry of H<sub>2</sub>- and CH<sub>4</sub>-enriched hydrothermal fluids of Socorro Island, Revillagigedo Archipelago, Mexico. Evidence for serpentinization and abiogenic methane, *Geofluids*, **10**, 542-555.
- Tatusov, R.L., Altschul, S.F. and Koonin, E.V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks, *Proc Natl Acad Sci U S A*, **91**, 12091-12095.
- Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A. (2003) The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **4**, 41.
- Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families, *Science*, **278**, 631-637.
- Tetas, M. and Lowenstein, J.M. (1963) The effect of bivalent metal ions on the hydrolysis of adenosine di- and triphosphate, *Biochemistry*, **2**, 350-357.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res*, **22**, 4673-4680.

- Tietze, M., Beuchle, A., Lamla, I., Orth, N., Dehler, M., Greiner, G. and Beifuss, U. (2003) Redox potentials of methanophenazine and CoB-S-S-CoM, factors involved in electron transport in Methanogenic archaea, *ChemBiochem*, **4**, 333-335.
- Tocilescu, M.A., Zickermann, V., Zwicker, K. and Brandt, U. (2010) Quinone binding and reduction by respiratory complex I, *Biochim Biophys Acta*, **1797**, 1883-1890.
- Toei, M., Gerle, C., Nakano, M., Tani, K., Gyobu, N., Tamakoshi, M., Sone, N., Yoshida, M., Fujiyoshi, Y., Mitsuoka, K. and Yokoyama, K. (2007) Dodecamer rotor ring defines H<sup>+</sup>/ATP ratio for ATP synthesis of prokaryotic V-ATPase from *Thermus thermophilus*, *Proc Natl Acad Sci U S A*, **104**, 20256-20261.
- Trent, J.D., Nimmegern, E., Wall, J.S., Hartl, F.U. and Horwich, A.L. (1991) A molecular chaperone from a thermophilic archaeobacterium is related to the eukaryotic protein t-complex polypeptide-1, *Nature*, **354**, 490-493.
- Trifonov, E.N. (2000) Consensus temporal order of amino acids and evolution of the triplet code, *Gene*, **261**, 139-151.
- Tsukatani, Y., Azai, C., Kondo, T., Itoh, S. and Oh-Oka, H. (2008) Parallel electron donation pathways to cytochrome c(z) in the type I homodimeric photosynthetic reaction center complex of *Chlorobium tepidum*, *Biochim Biophys Acta*, **1777**, 1211-1217.
- Tsukihara, T., Shimokata, K., Katayama, Y., Shimada, H., Muramoto, K., Aoyama, H., Mochizuki, M., Shinzawa-Itoh, K., Yamashita, E., Yao, M., Ishimura, Y. and Yoshikawa, S. (2003) The low-spin heme of cytochrome c oxidase as the driving element of the proton-pumping process, *Proc Natl Acad Sci U S A*, **100**, 15304-15309.
- UniProtConsortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acids Res*, **40**, D71-75.
- van Raaij, M.J., Abrahams, J.P., Leslie, A.G. and Walker, J.E. (1996) The structure of bovine F1-ATPase complexed with the antibiotic inhibitor aurovertin B, *Proc Natl Acad Sci U S A*, **93**, 6913-6917.
- Verlander, M.S., Lohrmann, R. and Orgel, L.E. (1973) Catalysts for the self-polymerization of adenosine cyclic 2', 3'-phosphate, *J Mol Evol*, **2**, 303-316.
- Vermeiglio, A. and Joliot, P. (1999) The photosynthetic apparatus of *Rhodobacter sphaeroides*, *Trends Microbiol*, **7**, 435-440.
- Vetter, I.R. and Wittinghofer, A. (1999) Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer, *Q Rev Biophys*, **32**, 1-56.
- Vignais, P.M. and Billoud, B. (2007) Occurrence, classification, and biological function of hydrogenases: an overview, *Chem Rev*, **107**, 4206-4272.
- Viitanen, P.V., Lubben, T.H., Reed, J., Goloubinoff, P., O'Keefe, D.P. and Lorimer, G.H. (1990) Chaperonin-facilitated refolding of ribulosebiphosphate carboxylase and ATP hydrolysis by chaperonin 60 (groEL) are K<sup>+</sup> dependent, *Biochemistry*, **29**, 5665-5671.
- Viitanen, P.V., Schmidt, M., Buchner, J., Suzuki, T., Vierling, E., Dickson, R., Lorimer, G.H., Gatenby, A. and Soll, J. (1995) Functional characterization of the higher plant chloroplast chaperonins, *J Biol Chem*, **270**, 18158-18164.
- Vik, S.B. and Antonio, B.J. (1994) A mechanism of proton translocation by F1F0 ATP synthases suggested by double mutants of the a subunit, *J Biol Chem*, **269**, 30364-30369.
- von Ballmoos, C., Cook, G.M. and Dimroth, P. (2008) Unique rotary ATP synthase and its biological diversity, *Annu Rev Biophys*, **37**, 43-64.
- von Ballmoos, C. and Dimroth, P. (2007) Two distinct proton binding sites in the ATP synthase family, *Biochemistry*, **46**, 11800-11809.
- Von Damm, K.L. (2001) Lost City found, *Nature*, **412**, 127-128.
- Voorhees, R.M., Schmeing, T.M., Kelley, A.C. and Ramakrishnan, V. (2010) The mechanism for activation of GTP hydrolysis on the ribosome, *Science*, **330**, 835-838.



- Wächtershäuser, G. (1988) Before enzymes and templates: theory of surface metabolism, *Microbiol Rev*, **52**, 452-484.
- Wächtershäuser, G. (1990) Evolution of the first metabolic cycles, *Proc Natl Acad Sci U S A*, **87**, 200-204.
- Wächtershäuser, G. (1992) Groundworks for an evolutionary biochemistry: the iron-sulphur world, *Prog Biophys Mol Biol*, **58**, 85-201.
- Wächtershäuser, G. (2007) On the chemistry and evolution of the pioneer organism, *Chem Biodivers*, **4**, 584-602.
- Wajant, H. (2002) The Fas signaling pathway: more than a paradigm, *Science*, **296**, 1635-1636.
- Wald, G. (1964) The Origins of Life, *Proc Natl Acad Sci U S A*, **52**, 595-611.
- Walker, J. (1998) ATP synthesis by rotary catalysis (Nobel lecture), *Angew Chem Int Ed Engl*, **37**, 2309-2319.
- Wallace, I.M., Blackshields, G. and Higgins, D.G. (2005) Multiple sequence alignments, *Curr Opin Struct Biol*, **15**, 261-266.
- Wallin, I.E. (1923) The Mitochondria Problem, *The American Naturalist*, **57**, 255-261.
- Walter, P., Ibrahimi, I. and Blobel, G. (1981) Translocation of proteins across the endoplasmic reticulum. I. Signal recognition protein (SRP) binds to in-vitro-assembled polysomes synthesizing secretory protein, *J Cell Biol*, **91**, 545-550.
- Wang, C. and Youle, R.J. (2009) The role of mitochondria in apoptosis\*, *Annu Rev Genet*, **43**, 95-118.
- Wang, J. and Boisvert, D.C. (2003) Structural basis for GroEL-assisted protein folding from the crystal structure of (GroEL-KMgATP)<sub>14</sub> at 2.0Å resolution, *J Mol Biol*, **327**, 843-855.
- Watanabe, H., Kamita, Y., Nakamura, T., Takimoto, A. and Yamanaka, T. (1979) The terminal oxidase of *Photobacterium phosphoreum*. A novel cytochrome, *Biochim Biophys Acta*, **547**, 70-78.
- Watt, I.N., Montgomery, M.G., Runswick, M.J., Leslie, A.G. and Walker, J.E. (2010) Bioenergetic cost of making an adenosine triphosphate molecule in animal mitochondria, *Proc Natl Acad Sci U S A*, **107**, 16823-16827.
- Weber, J. and Senior, A.E. (2003) ATP synthesis driven by proton transport in F<sub>1</sub>F<sub>0</sub>-ATP synthase, *FEBS Lett*, **545**, 61-70.
- Wei, M.C., Zong, W.X., Cheng, E.H., Lindsten, T., Panoutsakopoulou, V., Ross, A.J., Roth, K.A., MacGregor, G.R., Thompson, C.B. and Korsmeyer, S.J. (2001) Proapoptotic BAX and BAK: a requisite gateway to mitochondrial dysfunction and death, *Science*, **292**, 727-730.
- Welte, C., Kratzer, C. and Deppenmeier, U. (2010) Involvement of Ech hydrogenase in energy conservation of *Methanosarcina mazei*, *FEBS J*, **277**, 3396-3403.
- Weng, M., Makaroff, C.A. and Zalkin, H. (1986) Nucleotide sequence of *Escherichia coli* pyrG encoding CTP synthetase, *J Biol Chem*, **261**, 5568-5574.
- West, S.C. (1996) The RuvABC proteins and Holliday junction processing in *Escherichia coli*, *J Bacteriol*, **178**, 1237-1241.
- Westheimer, F.H. (1987) Why nature chose phosphates, *Science*, **235**, 1173-1178.
- Whelan, S. and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, *Mol Biol Evol*, **18**, 691-699.
- Widger, W.R., Cramer, W.A., Herrmann, R.G. and Trebst, A. (1984) Sequence homology and structural similarity between cytochrome b of mitochondrial complex III and the chloroplast b<sub>6</sub>-f complex: position of the cytochrome b hemes in the membrane, *Proc Natl Acad Sci U S A*, **81**, 674-678.
- Wilbanks, S.M. and McKay, D.B. (1995) How potassium affects the activity of the molecular chaperone Hsc70. II. Potassium binds specifically in the ATPase active site, *J Biol Chem*, **270**, 2251-2257.

- Willey, J.M., Waterbury, J.B. and Greenberg, E.P. (1987) Sodium-coupled motility in a swimming cyanobacterium, *J Bacteriol*, **169**, 3429-3434.
- Williams, R.J.P. and Frausto da Silva, J.J.R. (2006) *The Chemistry of Evolution: The Development of our Ecosystem*. Elsevier, Amsterdam.
- Winkler, J.R., Wittung-Stafshede, P., Leckner, J., Malmstrom, B.G. and Gray, H.B. (1997) Effects of folding on metalloprotein active sites, *Proc Natl Acad Sci U S A*, **94**, 4246-4249.
- Witt, H., Malatesta, F., Nicoletti, F., Brunori, M. and Ludwig, B. (1998) Cytochrome-c-binding site on cytochrome oxidase in *Paracoccus denitrificans*, *Eur J Biochem*, **251**, 367-373.
- Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms, *Proc Natl Acad Sci U S A*, **74**, 5088-5090.
- Woese, C.R., Kandler, O. and Wheelis, M.L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, *Proc Natl Acad Sci U S A*, **87**, 4576-4579.
- Wolf, Y.I. and Koonin, E.V. (2007) On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization, *Biol Direct*, **2**, 14.
- Wu, Y., Qian, X., He, Y., Moya, I.A. and Luo, Y. (2005) Crystal structure of an ATPase-active form of Rad51 homolog from *Methanococcus voltae*. Insights into potassium dependence, *J Biol Chem*, **280**, 722-728.
- Wyllie, A.H., Kerr, J.F. and Currie, A.R. (1980) Cell death: the significance of apoptosis, *Int Rev Cytol*, **68**, 251-306.
- Xiong, J., Inoue, K. and Bauer, C.E. (1998) Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from *Heliobacillus mobilis*, *Proc Natl Acad Sci U S A*, **95**, 14851-14856.
- Yamada, K., Kunishima, N., Mayanagi, K., Ohnishi, T., Nishino, T., Iwasaki, H., Shinagawa, H. and Morikawa, K. (2001) Crystal structure of the Holliday junction migration motor protein RuvB from *Thermus thermophilus* HB8, *Proc Natl Acad Sci U S A*, **98**, 1442-1447.
- Yamanaka, K., Hwang, J. and Inouye, M. (2000) Characterization of GTPase activity of TrmE, a member of a novel GTPase superfamily, from *Thermotoga maritima*, *J Bacteriol*, **182**, 7078-7082.
- Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G.J. and Woese, C.R. (1985) Mitochondrial origins, *Proc Natl Acad Sci U S A*, **82**, 4443-4447.
- Yankovskaya, V., Horsefield, R., Tornroth, S., Luna-Chavez, C., Miyoshi, H., Leger, C., Byrne, B., Cecchini, G. and Iwata, S. (2003) Architecture of succinate dehydrogenase and reactive oxygen species generation, *Science*, **299**, 700-704.
- Yanyushin, M.F. (2002) Fractionation of cytochromes of phototrophically grown *Chloroflexus aurantiacus*. Is there a cytochrome bc complex among them?, *FEBS Lett*, **512**, 125-128.
- Yanyushin, M.F., del Rosario, M.C., Brune, D.C. and Blankenship, R.E. (2005) New class of bacterial membrane oxidoreductases, *Biochemistry*, **44**, 10037-10045.
- Yasuda, R., Noji, H., Yoshida, M., Kinoshita, K., Jr. and Itoh, H. (2001) Resolution of distinct rotational substeps by submillisecond kinetic analysis of F1-ATPase, *Nature*, **410**, 898-904.
- Yin, Y., Yang, S., Yu, L. and Yu, C.A. (2010) Reaction mechanism of superoxide generation during ubiquinol oxidation by the cytochrome bc1 complex, *J Biol Chem*, **285**, 17038-17045.
- Youngman, E.M., Brunelle, J.L., Kochaniak, A.B. and Green, R. (2004) The active site of the ribosome is composed of two layers of conserved nucleotides with distinct roles in peptide bond formation and peptide release, *Cell*, **117**, 589-599.
- Yu, J., Hederstedt, L. and Piggot, P.J. (1995) The cytochrome bc complex (menaquinone:cytochrome c reductase) in *Bacillus subtilis* has a nontraditional subunit organization, *J Bacteriol*, **177**, 6751-6760.

- Yu, J. and Le Brun, N.E. (1998) Studies of the cytochrome subunits of menaquinone:cytochrome c reductase (bc complex) of *Bacillus subtilis*. Evidence for the covalent attachment of heme to the cytochrome b subunit, *J Biol Chem*, **273**, 8860-8866.
- Yu, T., Wang, X., Purring-Koch, C., Wei, Y. and McLendon, G.L. (2001) A mutational epitope for cytochrome C binding to the apoptosis protease activation factor-1, *J Biol Chem*, **276**, 13034-13038.
- Yue, H., Kang, Y., Zhang, H., Gao, X. and Blankenship, R.E. (2012) Expression and characterization of the diheme cytochrome c subunit of the cytochrome bc complex in *Heliobacterium modesticaldum*, *Arch Biochem Biophys*, **517**, 131-137.
- Yutin, N., Makarova, K.S., Mekhedov, S.L., Wolf, Y.I. and Koonin, E.V. (2008) The deep archaeal roots of eukaryotes, *Mol Biol Evol*, **25**, 1619-1630.
- Yutin, N., Puigbo, P., Koonin, E.V. and Wolf, Y.I. (2012) Phylogenomics of prokaryotic ribosomal proteins, *PLoS One*, **7**, e36972.
- Zamzami, N., Susin, S.A., Marchetti, P., Hirsch, T., Gomez-Monterrey, I., Castedo, M. and Kroemer, G. (1996) Mitochondrial control of nuclear apoptosis, *J Exp Med*, **183**, 1533-1544.
- Zhang, H., Whitelegge, J.P. and Cramer, W.A. (2001) Ferredoxin:NADP<sup>+</sup> oxidoreductase is a subunit of the chloroplast cytochrome b6f complex, *J Biol Chem*, **276**, 38159-38165.
- Zhang, Y. and Fillingame, R.H. (1995) Changing the ion binding specificity of the *Escherichia coli* H<sup>(+)</sup>-transporting ATP synthase by directed mutagenesis of subunit c, *J Biol Chem*, **270**, 87-93.
- Zuckerlandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history, *J Theor Biol*, **8**, 357-366.

## 11. Abbreviations

ATP (GTP, CTP)	Adenosine triphosphate (guanosine ..., cytosine ...)
ADP (GDP, CDP)	Adenosine dihosphate (guanosine ..., cytosine ...)
AMP	Adenosine monohosphate
LUCA	Last universal cellular ancestor
CoA	Coenzyme A
COG	Cluster of orthologous groups
LGT	Lateral gene transfer
CL	Cardiolipin
2D	Two-dimensional
3D	Three-dimensional
FMN	Flavin mononucleotide
FAD	Flavin adenine dinucleotide
Q	Quinone
NAD <sup>+</sup>	Nicotinamide adenine dinucleotide
NADP <sup>+</sup>	Nicotinamide adenine dinucleotide phosphate
PRC	Photosynthetic reaction center
$\Delta\tilde{\mu}_{H^+}$	Transmembrane difference in electrochemical potentials of hydrogen ions (proton potential)
$\Delta\tilde{\mu}_{Na^+}$	Transmembrane difference in electrochemical potentials of sodium ions (sodium potential)
$\Delta\tilde{\mu}_i$	Transmembrane ion potential
PMF	Proton-motive force
SMF	Sodium-motive force
OMM	Outer mitochondrial membrane
ROS	Reactive oxygen species
Apaf-1	Apoptotic protease activating factor

## 12. Supplementary material

List of supplementary materials:

**Table S1.** Products of ubiquitous genes and their association with essential inorganic cations and anions.

**Table S2.** List of 179 prokaryotic species with fully deciphered genomes that were used upon phylogenomic analysis.

**Table S3.** List of 35 eukaryotic species with fully deciphered genomes that were used upon phylogenomic analysis.

**Table S4.** List of 33 proteobacterial sequences with fully deciphered genomes that were used upon phylogenomic analysis.

This page is intentionally left blank

**Table S1. Products of ubiquitous genes and their association with essential inorganic cations and anions (taken from (Mulkidjanian *et al.*, 2012)).**

The lists of ubiquitous genes were extracted from refs. (1, 2). The data on the dependence of functional activity on particular metals were taken from the BRENDA database (3). According to the BRENDA database, the enzymatic activity of most  $Mg^{2+}$ -dependent enzymes could be routinely restored by  $Mn^{2+}$ . As concentration of  $Mg^{2+}$  ions in the cell is ca.  $10^{-2}$  M, whereas that of  $Mn^{2+}$  ions is ca.  $10^{-6}$  M, the data on the functional importance of  $Mn^{2+}$  were not included in the table for many enzymes. The presence of metals in protein structures was as listed in the Protein Data Bank (4) entries. The table includes all enzymes represented by orthologs in all cellular life forms as well as several cases when a function is ubiquitous (e.g., DNA polymerase, DNA primase) whereas the enzymes responsible for that function are represented by two or more non-orthologous forms (5). Upward arrows indicate the activation by the particular ion and downward arrows indicate the inhibition by this ion. If low ion concentrations activate the enzyme while high amounts of the same ion cause its inhibition then the  $\uparrow\downarrow$  sign is used.

Protein function	EC number (if available)	Functionally relevant inorganic anions	Functional dependence on monovalent cations	Monovalent cations in at least some structures	Functional dependence on divalent cations	Divalent cations in at least some structures
<b>Products of ubiquitous genes, according to (Koonin 2000 (1))</b>						
Ribosome as whole	-	-	$K^+$ (6, 7)	$K^+$ , $Na^+$ (in the 1JJ2 structure)*	$Mg^{2+}$ (8)	$Mg^{2+}$ (1JJ2, 3OH7, 1MMS), $\uparrow\uparrow Cd^{2+}$ (1MMS) $Zn^{2+}$ (1HR0)
Ribosomal protein L1	-	-	-	-	-	-
Ribosomal protein L10	-	-	-	-	-	-
Ribosomal protein L11	-	-	-	-	-	$Mg^{2+}$ , $Cd^{2+}$ (1MMS)
Ribosomal protein L13	-	-	-	$K^+$ , $Na^+$ (1JJ2)	-	-
Ribosomal protein L14	-	-	-	-	-	-
Ribosomal protein L15	-	-	-	$K^+$ , $Na^+$ (1JJ2)	-	$Mg^{2+}$ (3OH7)
Ribosomal protein L16/L10E	-	-	-	$K^+$ , $Na^+$ (1JJ2)	-	$Mg^{2+}$ (3OH7)
Ribosomal protein L18	-	-	-	-	-	-
Ribosomal protein L2	-	-	-	$K^+$ , $Na^+$ (1JJ2)	-	$Mg^{2+}$ (1JJ2, 3OH7)
Ribosomal protein L22	-	-	-	$K^+$ , $Na^+$ (1JJ2)	-	-
Ribosomal protein L24	-	-	-	$K^+$ , $Na^+$ (1JJ2)	-	$Mg^{2+}$ (1JJ2)
Ribosomal protein L29	-	-	-	-	-	-

\* Upon the crystallization of the ribosomes of *H. marismortui*, the media contained 1.7 M of  $Na^+$ , so that mostly  $Na^+$  ions are observed in the 1JJ2 structure. *H. marismortui* is a halophylic archaea and can accumulate up to 3M  $K^+$  under high salt conditions; thus the  $Na^+$ -containing sites in the 1JJ2 structure should be mostly occupied by  $K^+$  ions *in vivo*, as in other ribosomal structures.

$\uparrow\uparrow$  Here and below  $Cd^{2+}$  is not a native cofactor. High amounts of  $CdCl_2$  were added upon ribosome crystallization to improve the phasing. Other divalent cations occupy the place of  $Cd^{2+}$  ions in the native ribosome, see a detailed discussion in (9).

Ribosomal protein L3	-	-	-	K <sup>+</sup> , Na <sup>+</sup> (1JJ2)	-	Mg <sup>2+</sup> (1JJ2, 3OH7)
Ribosomal protein L4	-	-	-	K <sup>+</sup> , Na <sup>+</sup> (1JJ2)	-	Mg <sup>2+</sup> (3OH7)
Ribosomal protein L5	-	-	-	-	-	-
Ribosomal protein L6	-	-	-	-	-	-
Ribosomal protein S10	-	-	-	-	-	-
Ribosomal protein S11	-	-	-	-	-	-
Ribosomal protein S12	-	-	-	-	-	Mg <sup>2+</sup> (3OH7)
Ribosomal protein S13	-	-	-	-	-	-
Ribosomal protein S14	-	-	-	-	-	Zn <sup>2+</sup> (1HR0)
Ribosomal protein S15	-	-	-	-	-	-
Ribosomal protein S17	-	-	-	-	-	-
Ribosomal protein S19	-	-	-	-	-	-
Ribosomal protein S2	-	-	-	-	-	-
Ribosomal protein S3	3.1.25.1	-	-	-	-	-
Ribosomal protein S5	-	-	-	-	-	-
Ribosomal protein S7	-	-	-	-	-	-
Ribosomal protein S8	-	-	-	-	-	-
Ribosomal protein S9	-	-	-	-	-	-
Translation elongation factor G (EF-2)	3.6.5.3	phosphate	↑ NH <sub>4</sub> <sup>+</sup> , Na <sup>+</sup> (10)	-	Mg <sup>2+</sup>	Mg <sup>2+</sup> (1WRI)
Translation elongation factor Tu (EF-1)	3.6.5.3	phosphate	↑ NH <sub>4</sub> <sup>+</sup> , K <sup>+</sup> , ↓ Na <sup>+</sup> (10, 11)	-	Mg <sup>2+</sup> /Mn <sup>2+</sup>	Mg <sup>2+</sup> (2XQD)
Translation initiation factor 2	3.6.5.3	phosphate	↑ NH <sub>4</sub> <sup>+</sup> , ↓ Na <sup>+</sup> (10)	-	Mg <sup>2+</sup> , Zn <sup>2+</sup> (p subunit) (12)	Mg <sup>2+</sup> (1G7T), Zn <sup>2+</sup> (2QMU, 2D74)
Translation initiation factor IF-1	3.6.5.3	phosphate	↑ NH <sub>4</sub> <sup>+</sup> , ↓ Na <sup>+</sup> (10)	-	Mg <sup>2+</sup>	Mg <sup>2+</sup> (1HR0)
Translation elongation factor P/ translation initiation factor eIF5-a	-	-	-	-	-	No metals seen
Seryl-tRNA synthetase	6.1.1.11	pyrophosphate	↓ K <sup>+</sup> (13)	Na <sup>+</sup>	Mg <sup>2+</sup> (14), Zn <sup>2+</sup> (15)	Mg <sup>2+</sup> , Zn <sup>2+</sup>
Methionyl-tRNA synthetase	6.1.1.10	pyrophosphate	-	-	Mg <sup>2+</sup> (16), Zn <sup>2+</sup> (17)	Zn <sup>2+</sup>
Histidyl-tRNA synthetase	6.1.1.21	pyrophosphate	-	-	Mg <sup>2+</sup> (18)	-
Tryptophanyl-tRNA synthetase	6.1.1.2	pyrophosphate	↓ K <sup>+</sup> (13)	Cs <sup>+</sup>	Mg <sup>2+</sup> , Zn <sup>2+</sup> (19)	Mg <sup>2+</sup> , Ca <sup>2+</sup> , ‡Cd <sup>2+</sup>
Tyrosyl-tRNA synthetase	6.1.1.1	pyrophosphate	↑ K <sup>+</sup> , ↓ Na <sup>+</sup> (20), ↑ K <sup>+</sup> (21)	K <sup>+</sup>	Mg <sup>2+</sup> (20, 22, 23)	Mg <sup>2+</sup>
Phenylalanyl-tRNA synthetase	6.1.1.20	pyrophosphate	-	-	Mg <sup>2+</sup> (18), Zn <sup>2+</sup> (24, 25)	Mg <sup>2+</sup> /Mn <sup>2+</sup> , Zn <sup>2+</sup>
Aspartyl-tRNA synthetase	6.1.1.12	pyrophosphate	↑ ↓ K <sup>+</sup> , ↓ Na <sup>+</sup> (26)	-	Mg <sup>2+</sup> (27)	Mg <sup>2+</sup> /Mn <sup>2+</sup>
Valyl-tRNA synthetase	6.1.1.9	pyrophosphate	↑ ↓ K <sup>+</sup> (13)	-	Mg <sup>2+</sup> (28)	Zn <sup>2+</sup>
Isoleucyl-tRNA synthetase	6.1.1.5	pyrophosphate	-	K <sup>+</sup>	Mg <sup>2+</sup> (18), Zn <sup>2+</sup> (29)	Mg <sup>2+</sup> , Zn <sup>2+</sup>
Leucyl-tRNA synthetase	6.1.1.4	pyrophosphate	-	-	Mg <sup>2+</sup> (30)	Zn <sup>2+</sup> , Hg <sup>2+</sup>
Threonyl-tRNA synthetase	6.1.1.3	pyrophosphate	-	-	Mg <sup>2+</sup> (31), Zn <sup>2+</sup> (32)	Zn <sup>2+</sup>
Arginyl-tRNA synthetase	6.1.1.19	pyrophosphate	↓ K <sup>+</sup> (33)	-	Mg <sup>2+</sup> (18)	Mg <sup>2+</sup>
Prolyl-tRNA synthetase	6.1.1.15	pyrophosphate	-	-	Mg <sup>2+</sup> (34), Zn <sup>2+</sup> (35)	Mg <sup>2+</sup> /Mn <sup>2+</sup> , Zn <sup>2+</sup>



Alanyl-tRNA synthetase	6.1.1.7	pyrophosphate	-	-	Mg <sup>2+</sup> , Zn <sup>2+</sup> (36)	Mg <sup>2+</sup> , Zn <sup>2+</sup>
Pseudouridylate synthase	5.4.99.12	-	K <sup>+</sup> (37)	K <sup>+</sup>	Zn <sup>2+</sup> (38)	Zn <sup>2+</sup>
Methionine aminopeptidase	3.4.11.18	-	-	K <sup>+</sup> > Na <sup>+</sup>	Zn <sup>2+</sup> , Mn <sup>2+</sup> , Co <sup>2+</sup> , Ni <sup>2+</sup> , Fe <sup>2+</sup> (35-39, 40)	Mn <sup>2+</sup> /Zn <sup>2+</sup> / Co <sup>2+</sup> /Ni <sup>2+</sup> /Fe <sup>2+</sup>
Transcription antiterminator NusG	-	-	-	-	-	No metals seen
DNA-directed RNA polymerase, subunits $\alpha$ , $\beta$ , $\beta'$	2.7.7.6	pyrophosphate	-	Na <sup>+</sup> (3K4G)	Mg <sup>2+</sup> (41), Zn <sup>2+</sup> (42)	Mg <sup>2+</sup> /Mn <sup>2+</sup> , Zn <sup>2+</sup>
DNA polymerase III, subunit $\beta$ (sliding clamp)	2.7.7.7	-	-	-	-	-
Clamp loader ATPase (DNA polymerase III, subunit $\gamma$ and $\gamma'$ )	2.7.7.7	phosphate	-	-	Mg <sup>2+</sup>	Mg <sup>2+</sup> , Zn <sup>2+</sup>
Topoisomerase IA	5.99.1.2	-	K <sup>+</sup> > Na <sup>+</sup> (43)	-	Mg <sup>2+</sup> (44)	Zn <sup>2+</sup> (1CCY)
5'-3' exonuclease (including N-terminal domain of	3.1.11. <sup>a</sup>	phosphate	-	-	Mg <sup>2+</sup> (45)	Mn <sup>2+</sup> (1UT5), Zn <sup>2+</sup> (1TAQ)
	a = 3-6 (5->3) a = 1-2 (3->5)					
RecA/RadA (Rad51) recombinase	-	phosphate	K <sup>+</sup> (46)	K <sup>+</sup> (1XU4)	Mg <sup>2+</sup> (47)	Mg <sup>2+</sup> (1T4G)
Chaperonin GroEL	3.6.4.9	phosphate	K <sup>+</sup> (48)	K <sup>+</sup>	Mg <sup>2+</sup> (49)	Mg <sup>2+</sup>
O-sialoglycoprotease (MG2+046, PF00814) (50)/ apurinic endonuclease (51)	3.4.24.57	phosphate	-	-	Zn <sup>2+</sup> (52), Fe <sup>3+</sup> (51)	Mg <sup>2+</sup> (1IVN, 3ENO)
EMAP domain (OB-fold RNA-binding domain, MG2+449, PF01588)	-	-	-	-	-	-
Thymidylate kinase	2.7.4.9	phosphate	-	Na <sup>+</sup> (2WWF)	Mg <sup>2+</sup> (53)	Mg <sup>2+</sup>
Thioredoxin reductase	1.8.1.9	-	-	-	-	Mg <sup>2+</sup> (2A87)
Thioredoxin	-	-	-	-	-	Cd <sup>2+</sup> (3KD0) <sup>§</sup> , Zn <sup>2+</sup> (3P2A, 2XC2, 3HXS)
CDP-diglyceride-synthase	2.7.7.41	pyrophosphate	$\uparrow$ K <sup>+</sup> , NH <sub>4</sub> <sup>+</sup> , Rb <sup>+</sup> $\downarrow$ Li <sup>+</sup> , Na <sup>+</sup> Cs <sup>+</sup>	No entries	Mg <sup>2+</sup> (54)	No entries
Phosphomannomutase	5.4.2.8	-	-	-	Mg <sup>2+</sup> (55)	Mg <sup>2+</sup> /Zn <sup>2+</sup>
Catalytic subunit of the membrane ATP synthase	3.6.3.14	phosphate	-	-	Mg <sup>2+</sup> (56, 57)	Mg <sup>2+</sup>
Proteolipid subunits of the membrane ATP synthase	3.6.3.14	-	-	-	-	No metals seen
Triosephosphate isomerase	5.3.1.1	-	-	-	-	No metals seen
Glycine hydroxymethyltransferase	2.1.2.1	-	$\downarrow$ K <sup>+</sup> , NH <sub>4</sub> <sup>+</sup> , Na <sup>+</sup> (58, 59)	-	$\downarrow$ Mg <sup>2+</sup> , Mn <sup>2+</sup> , Ca <sup>2+</sup> (58, 59)	No metals seen
Preprotein translocase subunit SecY	-	-	-	-	-	Zn <sup>2+</sup> (2ZJS)
Signal recognition particle GTPase FtsY	3.6.5.4	phosphate	-	K <sup>+</sup> (2J7P)	Mg <sup>2+</sup> (60)	Mg <sup>2+</sup>
Predicted GTPase (YchF, PF06071, 1JAL, 2OHF, 2DBY, 2DWQ, 1N13)	-	phosphate	K <sup>+</sup> (61)	-	Mg <sup>2+</sup> (62)	No metals seen
<b>Additional ubiquitous gene products from ref. (Charlebois and Doolittle 2004 (2))</b>						
DNA primase (dnaG)	2.7.7.-	pyrophosphate	-	-	Zn <sup>2+</sup> (63)	Zn <sup>2+</sup> (2AU3, 1DOQ)
S-adenosylmethionine-6-N <sup>7</sup> ,N <sup>7</sup> - adenosyl (rRNA) dimethyltransferase	2.1.1.48	-	-	-	Mg <sup>2+</sup> (64)	No metals seen
Transcription pausing, L factor (NusA)	-	-	-	-	-	No metals seen

## References for Table S1

1. Koonin EV (2000) How many genes can make a cell: the minimal-gene-set concept. *Annu. Rev. Genom. Human Genet* 1:99-116.
2. Charlebois RL & Doolittle WF (2004) Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res* 14:2469-2477.
3. Chang A, Scheer M, Grote A, Schomburg I, & Schomburg D (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *NAR* 37:D588-592.
4. Berman HM, *et al.* (2000) The Protein Data Bank. *Nucl Acids Res* 28:235-242.
5. Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* 1:127-136.
6. Michelinaki M, Spanos A, Coutsogeorgopoulos C, & Kalpaxis DL (1997) New aspects on the kinetics of activation of ribosomal peptidyltransferase-catalyzed peptide bond formation by monovalent ions and spermine. *Biochim Biophys Acta* 1342:182-190.
7. Ioannou M & Coutsogeorgopoulos C (1997) Kinetic studies on the activation of eukaryotic peptidyltransferase by potassium. *Arch Biochem Biophys* 345:325-331.
8. Hsiao C & Williams LD (2009) A recurrent magnesium-binding motif provides a framework for the ribosomal peptidyl transferase center. *Nucleic Acids Res* 37:3134-3142.
9. Mulkidjanian AY & Galperin MY (2009) On the origin of life in the zinc world. 2. Validation of the hypothesis on the photosynthesizing zinc sulfide edifices as cradles of life on Earth. *Biol Direct* 4:27.
10. Conway TW (1964) On the role of ammonium or potassium ion in amino acid polymerization. *Proc. Natl. Acad. Sci. USA* 51:1216-1220.
11. Fasano O, De Vendittis E, & Parmeggiani A (1982) Hydrolysis of GTP by elongation factor Tu can be induced by monovalent cations in the absence of other effectors. *J Biol Chem* 257:3145-3150.
12. Donahue TF, Cigan AM, Pabich EK, & Valavicius BC (1988) Mutations at a Zn(II) finger motif in the yeast eIF-2 beta gene alter ribosomal start-site selection during the scanning process. *Cell* 54:621-632.
13. Jukubowski H & Pawelkiewicz J (1975) The plant aminoacyl-tRNA synthetases. Purification and characterization of valyl-tRNA, tryptophanyl-tRNA and seryl-tRNA synthetases from yellow-lupin seeds. *Eur J Biochem* 52:301-310.
14. Pachmann U & Zachau HG (1978) Yeast seryl tRNA synthetase: two sets of substrate sites involved in aminoacylation. *Nucleic Acids Res* 5:961-973.
15. Bilokapic S, *et al.* (2006) Structure of the unusual seryl-tRNA synthetase reveals a distinct zinc-dependent mode of substrate recognition. *EMBO J* 25:2498-2509.
16. Deobagkar DN & Gopinathan KP (1976) Two forms of methionyl-transfer RNA synthetase from *Mycobacterium smegmatis*. *Biochem Biophys Res Commun* 71:939-951.
17. Fourmy D, Mechulam Y, & Blanquet S (1995) Crucial role of an idiosyncratic insertion in the Rossman fold of class 1 aminoacyl-tRNA synthetases: the case of methionyl-tRNA synthetase. *Biochemistry* 34:15681-15688.
18. Airas RK (1996) Differences in the magnesium dependences of the class I and class II aminoacyl-tRNA synthetases from *Escherichia coli*. *Eur J Biochem* 240:223-231.
19. Kisselev LL (1993) Mammalian tryptophanyl-tRNA synthetases. *Biochimie* 75:1027-1039.
20. Austin J & First EA (2002) Catalysis of tyrosyl-adenylate formation by the human tyrosyl-tRNA synthetase. *J Biol Chem* 277:14812-14820.
21. Warner CK & Jacobson KB (1976) Mechanisms of suppression in *Drosophila*. IV. Specificity and properties of tyrosyl-tRNA synthetase. *Can J Biochem* 54:650- 656.
22. Brown P, *et al.* (1999) Molecular recognition of tyrosinyl adenylate analogues by prokaryotic tyrosyl tRNA synthetases. *Bioorg Med Chem* 7:2473-2485.
23. Hamano-Takaku F, *et al.* (2000) A mutant *Escherichia coli* tyrosyl-tRNA synthetase utilizes the unnatural amino acid azatyrosine more efficiently than

- tyrosine. *J Biol Chem* 275:40324-40328.
24. Brevet A, Plateau P, Cirakoglu B, Pailliez JP, & Blanquet S (1982) Zinc-dependent synthesis of 5',5'-diadenosine tetraphosphate by sheep liver lysyl- and phenylalanyl-tRNA synthetases. *J Biol Chem* 257:14613-14615.
  25. Plateau P, Mayaux JF, & Blanquet S (1981) Zinc(II)-dependent synthesis of diadenosine 5', 5''' -P(1),P(4) -tetraphosphate by *Escherichia coli* and yeast phenylalanyl transfer ribonucleic acid synthetases. *Biochemistry* 20:4654-4662.
  26. Vellekamp GJ & Kull FJ (1981) Allotropism in aspartyl-tRNA synthetase from porcine thyroid. *Eur J Biochem* 118:261-269.
  27. Norton SJ, Ravel JM, Lee C, & Shive W (1963) Purification and properties of the aspartyl ribonucleic acid synthetase of *Lactobacillus arabinosus*. *J Biol Chem* 238:269-274.
  28. Godar DE & Yang DC (1988) Mammalian high molecular weight and monomeric forms of valyl-tRNA synthetase. *Biochemistry* 27:2181-2186.
  29. Glasfeld E, Landro JA, & Schimmel P (1996) C-terminal zinc-containing peptide required for RNA recognition by a class I tRNA synthetase. *Biochemistry* 35:4139-4145.
  30. Chirikjian JG, Kanagalingam K, Lau E, & Fresco JR (1973) Purification and properties of leucyl-tRNA synthetase from Bakers' yeast. *J Biol Chem* 248:1074-1079.
  31. Freist W & Gauss DH (1995) Threonyl-tRNA synthetase. *Biol Chem Hoppe Seyler* 376:213-224.
  32. Sankaranarayanan R, *et al.* (1999) The structure of threonyl-tRNA synthetase-tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site. *Cell* 97:371-381.
  33. Airas RK (2006) Analysis of the kinetic mechanism of arginyl-tRNA synthetase. *Biochim Biophys Acta* 1764:307-319.
  34. Peterson PJ & Fowden L (1965) Purification, properties and comparative specificities of the enzyme prolyl-transfer ribonucleic acid synthetase from *Phaseolus aureus* and *Polygonatum multiflorum*. *Biochem J* 97:112-124.
  35. Kamtekar S, *et al.* (2003) The structural basis of cysteine aminoacylation of tRNA<sup>Pro</sup> by prolyl-tRNA synthetases. *Proc Natl Acad Sci USA* 100:1673-1678.
  36. Sood SM, Wu MX, Hill KA, & Slattery CW (1999) Characterization of zinc-depleted alanyl-tRNA synthetase from *Escherichia coli*: role of zinc. *Arch Biochem Biophys* 368:380-384.
  37. Green CJ, Kammen HO, & Penhoet EE (1982) Purification and properties of a mammalian tRNA pseudouridine synthase. *J Biol Chem* 257:3045-3052.
  38. Arluison V, Hountondji C, Robert B, & Grosjean H (1998) Transfer RNA-pseudouridine synthetase Pus1 of *Saccharomyces cerevisiae* contains one atom of zinc essential for its native conformation and tRNA recognition. *Biochemistry* 37:7268-7276.
  39. Ben-Bassat A, *et al.* (1987) Processing of the initiation methionine from proteins: properties of the *Escherichia coli* methionine aminopeptidase and its gene structure. *J Bacteriol* 169:751-757.
  40. Wang WL, *et al.* (2008) Discovery of inhibitors of *Escherichia coli* methionine aminopeptidase with the Fe(II)-form selectivity and antibacterial activity. *J Med Chem* 51:6110-6120.
  41. Sosunov V, *et al.* (2003) Unified two-metal mechanism of RNA synthesis and degradation by RNA polymerase. *EMBO J* 22:2234-2244.
  42. King RA, Markov D, Sen R, Severinov K, & Weisberg RA (2004) A conserved zinc binding domain in the largest subunit of DNA-dependent RNA polymerase modulates intrinsic transcription termination and antitermination but does not stabilize the elongation complex. *J Mol Biol* 342:1143-1154.
  43. Ivanov VA, Melnikov AA, & Terpilovska ON (1986) DNA topoisomerase I from rat brain neurons. *Biochim Biophys Acta* 866:154-160.
  44. Sutcliffe JA, Gootz TD, & Barrett JF (1989) Biochemical characteristics and physiological significance of major DNA topoisomerases. *Antimicrob Agents Chemother* 33:2027-2033.
  45. Dolberg M, Baur CP, & Knippers R (1991) Purification and characterization of a novel 5' exodeoxyribonuclease from the yeast *Saccharomyces cerevisiae*. *Eur J Biochem* 198:783-787.

46. Wu Y, Qian X, He Y, Moya IA, & Luo Y (2005) Crystal structure of an ATPase-active form of Rad51 homolog from *Methanococcus voltae*. Insights into potassium dependence. *J Biol Chem* 280:722-728.
47. Muller B, Burdett I, & West SC (1992) Unusual stability of recombination intermediates made by Escherichia coli RecA protein. *EMBO J* 11:2685-2693.
48. Viitanen PV, *et al.* (1990) Chaperonin-facilitated refolding of ribulosebisphosphate carboxylase and ATP hydrolysis by chaperonin 60 (groEL) are K<sup>+</sup> dependent. *Biochemistry* 29:5665-5671.
49. Horovitz A, Fridmann Y, Kafri G, & Yifrach O (2001) Review: allostery in chaperonins. *J Struct Biol* 135:104-114.
50. Katz C, Cohen-Or I, Gophna U, & Ron EZ (2010) The ubiquitous conserved glycopeptidase gcp prevents accumulation of toxic glycated proteins. *MBio* 1:3e00195-10.
51. Hecker A, *et al.* (2007) An archaeal orthologue of the universal protein KaeI is an iron metalloprotein which exhibits atypical DNA-binding properties and apurinic-endonuclease activity in vitro. *Nucleic Acids Res* 35:6042-6051.
52. Abdullah KM, Lo RY, & Mellors A (1991) Cloning, nucleotide sequence, and expression of the *Pasteurella haemolytica* A1 glycoprotease gene. *J Bacteriol* 173:5597-5603.
53. Nelson DJ & Carter CE (1969) Purification and characterization of Thymidine 5-monophosphate kinase from *Escherichia coli* B. *J Biol Chem* 244:5254-5262.
54. McCaman RE & Finnerty WR (1968) Biosynthesis of cytidine diphosphate-diglyceride by a particulate fraction from *Micrococcus cerificans*. *J Biol Chem* 243:5074-5080.
55. Guha SK & Rose ZB (1985) The synthesis of mannose 1-phosphate in brain. *Arch Biochem Biophys* 243:168-173.
56. Satoh S, Moritani C, Ohhashi T, Konishi K, & Ikeda M (1994) Chloroplast ATPase in *Acetabularia acetabulum*: purification and characterization of chloroplast F1-ATPase. *Biosci Biotechnol Biochem* 58:521-525.
57. Ferguson SA, Keis S, & Cook GM (2006) Biochemical and molecular characterization of a Na<sup>+</sup>-translocating F1Fo-ATPase from the thermoalkaliphilic bacterium *Clostridium paradoxum*. *J Bacteriol* 188:5045-5054.
58. Fujioka M (1969) Purification and properties of serine hydroxymethylase from soluble and mitochondrial fractions of rabbit liver. *Biochim Biophys Acta* 185:338-349.
59. Nakamura KD, Trewyn RW, & Parks LW (1973) Purification and characterization of serine transhydroxy-methylase from *Saccharomyces cerevisiae*. *Biochim Biophys Acta* 327:328-335.
60. Bange G, Wild K, & Sinning I (2007) Protein translocation: checkpoint role for SRP GTPase activation. *Curr Biol* 17:R980-982.
61. Anand B, Surana P, & Prakash B (2010) Deciphering the catalytic machinery in 30S ribosome assembly GTPase YqeH. *PLoS One* 5:e9944.
62. Teplyakov A, *et al.* (2003) Crystal structure of the YchF protein reveals binding sites for GTP and nucleic acid. *J Bacteriol* 185:4031-4037.
63. Pan H & Wigley DB (2000) Structure of the zinc-binding domain of *Bacillus stearothermophilus* DNA primase. *Structure* 8:231-239.
64. Andresson OS & Davies JE (1980) Some properties of the ribosomal RNA methyltransferase encoded by ksgA and the polarity of ksgA transcription. *Mol Gen Genet* 179:217-222.

**Table S2. List of 179 prokaryotic species with fully deciphered genomes that were used upon phylogenomic analysis.**Taxonomy was extracted from the NCBI Taxonomy web service (<http://www.ncbi.nlm.nih.gov/taxonomy>).

#	Domain	Phylum	Class	Species
1	Archaea	Crenarchaeota	Thermoprotei	<i>Desulfurococcus kamchatkensis</i> 1221n
2	Archaea	Crenarchaeota	Thermoprotei	<i>Hyperthermus butylicus</i>
3	Archaea	Crenarchaeota	Thermoprotei	<i>Sulfolobus solfataricus</i>
4	Archaea	Crenarchaeota	Thermoprotei	<i>Thermofilum pendens</i> Hrk 5
5	Archaea	Crenarchaeota	Thermoprotei	<i>Thermoproteus neutrophilus</i> V24Sta
6	Archaea	Euryarchaeota	Archaeoglobi	<i>Archaeoglobus fulgidus</i>
7	Archaea	Euryarchaeota	environmental	uncultured methanogenic archaeon RC-I
8	Archaea	Euryarchaeota	Halobacteria	<i>Halobacterium salinarum</i> R1
9	Archaea	Euryarchaeota	Methanobacteria	<i>Methanothermobacter thermautotrophicus</i>
10	Archaea	Euryarchaeota	Methanococci	<i>Methanocaldococcus jannaschii</i> DSM 2661
11	Archaea	Euryarchaeota	Methanococci	<i>Methanococcus aeolicus</i> Nankai-3
12	Archaea	Euryarchaeota	Methanomicrobia	<i>Methanocella paludicola</i> SANAE
13	Archaea	Euryarchaeota	Methanomicrobia	<i>Methanocorpusculum labreanum</i> Z
14	Archaea	Euryarchaeota	Methanomicrobia	<i>Methanosaeta thermophila</i> PT
15	Archaea	Euryarchaeota	Methanomicrobia	<i>Methanosarcina acetivorans</i>
16	Archaea	Euryarchaeota	Methanopyri	<i>Methanopyrus kandleri</i>
17	Archaea	Euryarchaeota	Thermococci	<i>Pyrococcus furiosus</i>
18	Archaea	Euryarchaeota	Thermoplasmata	<i>Picrophilus torridus</i> DSM 9790
19	Archaea	Euryarchaeota	Thermoplasmata	<i>Thermoplasma acidophilum</i>
20	Archaea	Euryarchaeota	unclassified	<i>Aciduliprofundum boonei</i> T469
21	Archaea	Korarchaeota	Korarchaeum	Candidatus <i>Korarchaeum cryptofilum</i> OPF8
22	Archaea	Nanoarchaeota	Nanoarchaeum	<i>Nanoarchaeum equitans</i>
23	Archaea	Thaumarchaeota	incertae sedis	<i>Nitrosopumilus maritimus</i> SCM1
24	Bacteria	Actinobacteria	Actinobacteria	<i>Corynebacterium efficiens</i> YS-314
25	Bacteria	Actinobacteria	Actinobacteria	<i>Corynebacterium kroppenstedtii</i> DSM 44385
26	Bacteria	Actinobacteria	Actinobacteria	<i>Micrococcus luteus</i> NCTC 2665
27	Bacteria	Actinobacteria	Actinobacteria	<i>Mycobacterium leprae</i>
28	Bacteria	Actinobacteria	Actinobacteria	<i>Mycobacterium marinum</i> M
29	Bacteria	Actinobacteria	Actinobacteria	<i>Mycobacterium tuberculosis</i> CDC1551
30	Bacteria	Actinobacteria	Actinobacteria	<i>Mycobacterium ulcerans</i> Agy99
31	Bacteria	Actinobacteria	Actinobacteria	<i>Nocardia farcinica</i> IFM10152
32	Bacteria	Actinobacteria	Actinobacteria	<i>Bifidobacterium longum</i>
33	Bacteria	Actinobacteria	Actinobacteria	<i>Gardnerella vaginalis</i> 409 05
34	Bacteria	Actinobacteria	Actinobacteria	<i>Atopobium parvulum</i> DSM 20469
35	Bacteria	Actinobacteria	Actinobacteria	<i>Rubrobacter xylanophilus</i> DSM 9941
36	Bacteria	Aquificae	Aquificae	<i>Aquifex aeolicus</i>
37	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	<i>Bacteroides fragilis</i> NCTC 9434
38	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	<i>Porphyromonas gingivalis</i> W83

## Supplementary material

39	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	Blattabacterium <i>Blattella germanica</i> Bge
40	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	Blattabacterium <i>Periplaneta americana</i> BPLAN
41	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	Candidatus <i>Sulcia muelleri</i> GWSS
42	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	<i>Capnocytophaga ochracea</i> DSM 7271
43	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	Flavobacteriaceae bacterium 3519 10
44	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	<i>Flavobacterium johnsoniae</i> UW101
45	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	<i>Flavobacterium psychrophilum</i> JIP02 86
46	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	<i>Gramella forsetii</i> KT0803
47	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	<i>Robiginitalea biformata</i> HTCC2501
48	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	<i>Chitinophaga pinensis</i> DSM 2588
49	Bacteria	Bacteroidetes/Chlorobi	Bacteroidetes	Candidatus <i>Amoebophilus asiaticus</i> 5a2
50	Bacteria	Bacteroidetes/Chlorobi	Chlorobi	<i>Chlorobaculum parvum</i> NCIB 8327
51	Bacteria	Chlamydiae/ Verrucomicrobia	Chlamydiae	<i>Chlamydia muridarum</i>
52	Bacteria	Chlamydiae/ Verrucomicrobia	Verrucomicrobia	<i>Opitutus terrae</i> PB90 1
53	Bacteria	Chloroflexi	Chloroflexi	<i>Chloroflexus aggregans</i> DSM 9485
54	Bacteria	Chloroflexi	Chloroflexi	<i>Chloroflexus aurantiacus</i> J 10 fl
55	Bacteria	Chloroflexi	Chloroflexi	<i>Chloroflexus</i> Y 400 fl
56	Bacteria	Chloroflexi	Chloroflexi	<i>Roseiflexus castenholzii</i> DSM 13941
57	Bacteria	Chloroflexi	Chloroflexi	<i>Roseiflexus</i> RS-1
58	Bacteria	Chloroflexi	Chloroflexi	<i>Herpetosiphon aurantiacus</i> ATCC 23779
59	Bacteria	Chloroflexi	Dehalococcoidetes	<i>Dehalococcoides ethenogenes</i> 195
60	Bacteria	Chloroflexi	Thermomicrobia	<i>Sphaerobacter thermophilus</i> DSM 20745
61	Bacteria	Chloroflexi	Thermomicrobia	<i>Thermomicrobium roseum</i> DSM 5159
62	Bacteria	Cyanobacteria	Chroococcales	<i>Cyanotheca</i> PCC 7425
63	Bacteria	Cyanobacteria	Nostocales	<i>Anabaena variabilis</i> ATCC 29413
64	Bacteria	Cyanobacteria	Nostocales	<i>Nostoc</i> sp
65	Bacteria	Cyanobacteria	unclassified	<i>Acaryochloris marina</i> MBIC11017
66	Bacteria	Deinococcus-Thermus	Deinococci	<i>Deinococcus radiodurans</i>
67	Bacteria	Deinococcus-Thermus	Deinococci	<i>Thermus thermophilus</i> HB8
68	Bacteria	Dictyoglomi	Dictyoglomia	<i>Dictyoglomus thermophilum</i> H 6 12
69	Bacteria	Elusimicrobia	Elusimicrobia	<i>Elusimicrobium minutum</i> Pei191
70	Bacteria	Fibrobacteres/ Acidobacteria	Acidobacteria	<i>Acidobacterium capsulatum</i> ATCC 51196
71	Bacteria	Fibrobacteres/ Acidobacteria	Fibrobacteres	<i>Fibrobacter succinogenes</i> S85
72	Bacteria	Firmicutes	Bacilli	<i>Bacillus subtilis</i>
73	Bacteria	Firmicutes	Bacilli	<i>Oceanobacillus iheyensis</i>
74	Bacteria	Firmicutes	Bacilli	<i>Listeria innocua</i>
75	Bacteria	Firmicutes	Bacilli	<i>Listeria monocytogenes</i>
76	Bacteria	Firmicutes	Bacilli	<i>Listeria seeligeri</i> serovar 1 2b SLCC3954
77	Bacteria	Firmicutes	Bacilli	<i>Listeria welshimeri</i> serovar 6b SLCC5334
78	Bacteria	Firmicutes	Bacilli	<i>Brevibacillus brevis</i> NBRC 100599

79	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Bacilli</i>	<i>Paenibacillus</i> JDR 2
80	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Bacilli</i>	<i>Staphylococcus aureus aureus</i> MRSA252
81	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Bacilli</i>	<i>Enterococcus faecalis</i> V583
82	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Bacilli</i>	<i>Lactobacillus casei</i>
83	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Bacilli</i>	<i>Streptococcus pneumoniae</i> G54
84	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Clostridia</i>	<i>Clostridium botulinum</i> A
85	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Clostridia</i>	<i>Clostridium difficile</i> 630
86	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Clostridia</i>	<i>Eubacterium eligens</i> ATCC 27750
87	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Clostridia</i>	<i>Heliobacterium modesticaldum</i> Ice1
88	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Clostridia</i>	<i>Desulfotobacterium hafniense</i> DCB 2
89	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Clostridia</i>	<i>Syntrophomonas wolfei</i> Goettingen
90	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Clostridia</i>	<i>Veillonella parvula</i> DSM 2008
91	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Clostridia</i>	<i>Haloferoxylum volcanium</i> H 168
92	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Clostridia</i>	<i>Natranaerobius thermophilus</i> JW NM WN LF
93	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Clostridia</i>	<i>Thermoanaerobacter tengcongensis</i>
94	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Clostridia</i>	<i>Coprothermobacter proteolyticus</i> DSM 5265
95	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Mollicutes</i>	<i>Acholeplasma laidlawii</i> PG 8A
96	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Mollicutes</i>	<i>Mycoplasma gallisepticum</i>
97	<i>Bacteria</i>	<i>Firmicutes</i>	<i>Mollicutes</i>	<i>Mycoplasma genitalium</i>
98	<i>Bacteria</i>	<i>Fusobacteria</i>	<i>Fusobacteria</i>	<i>Fusobacterium nucleatum</i>
99	<i>Bacteria</i>	<i>Fusobacteria</i>	<i>Fusobacteria</i>	<i>Leptotrichia buccalis</i> DSM 1135
100	<i>Bacteria</i>	<i>Fusobacteria</i>	<i>Fusobacteria</i>	<i>Sealdella termitidis</i> ATCC 33386
101	<i>Bacteria</i>	<i>Fusobacteria</i>	<i>Fusobacteria</i>	<i>Streptobacillus moniliformis</i> DSM 12112
102	<i>Bacteria</i>	<i>Gemmatimonadetes</i>	<i>Gemmatimonadetes</i>	<i>Gemmatimonas aurantiaca</i> T 27
103	<i>Bacteria</i>	<i>Nitrospirae</i>	<i>Nitrospira</i>	<i>Thermodesulfobivrio yellowstonii</i> DSM 11347
104	<i>Bacteria</i>	<i>Planctomycetes</i>	<i>Planctomycetacia</i>	<i>Rhodopirellula baltica</i>
105	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	<i>Phenylobacterium zucineum</i> HLK1
106	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	<i>Bartonella tribocorum</i> CIP 105476
107	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	<i>Agrobacterium radiobacter</i> K84
108	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	Candidatus <i>Liberibacter asiaticus</i> psy62
109	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	<i>Maricaulis maris</i> MCS10
110	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	<i>Dinoroseobacter shibae</i> DFL 12
111	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	<i>Acetobacter pasteurianus</i> IFO 3283 01
112	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	<i>Acidiphilium cryptum</i> JF-5
113	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	<i>Neorickettsia sennetsu</i> Miyayama
114	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	<i>Rickettsia typhi</i> wilmington
115	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	<i>Wolbachia</i> wRi
116	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	Candidatus <i>Pelagibacter ubique</i> HTCC1062
117	<i>Bacteria</i>	<i>Proteobacteria</i>	$\alpha$ -proteobacteria	<i>Zymomonas mobilis</i> ZM4
118	<i>Bacteria</i>	<i>Proteobacteria</i>	$\beta$ -proteobacteria	<i>Burkholderia cenocepacia</i> J2315
119	<i>Bacteria</i>	<i>Proteobacteria</i>	$\beta$ -proteobacteria	<i>Thiobacillus denitrificans</i> ATCC 25259

120	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\beta</math>-proteobacteria</i>	<i>Methylobacillus flagellatus</i> KT
121	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\beta</math>-proteobacteria</i>	<i>Chromobacterium violaceum</i>
122	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\beta</math>-proteobacteria</i>	<i>Neisseria gonorrhoeae</i> FA 1090
123	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\beta</math>-proteobacteria</i>	<i>Nitrosomonas europaea</i>
124	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\beta</math>-proteobacteria</i>	<i>Thauera</i> MZ1T
125	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Bdellovibrio bacteriovorus</i>
126	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Desulfobacterium autotrophicum</i> HRM2
127	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Desulfotalea psychrophila</i> LSv54
128	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Desulfohalobium retbaense</i> DSM 5692
129	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Lawsonia intracellularis</i> PHE MN1-00
130	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Geobacter uraniumreducens</i> Rf4
131	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Pelobacter carbinolicus</i>
132	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Anaeromyxobacter dehalogenans</i> 2CP-C
133	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Haliangium ochraceum</i> DSM 14365
134	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Myxococcus xanthus</i> DK 1622
135	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Sorangium cellulosum</i> So ce 56
136	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Syntrophus aciditrophicus</i> SB
137	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Campylobacter concisus</i> 13826
138	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\delta/\epsilon</math>-proteobacteria</i>	<i>Nautilia profundicola</i> AmH
139	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Acidithiobacillus ferrooxidans</i> ATCC 23270
140	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Aeromonas hydrophila</i> ATCC 7966
141	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Tolomonas auensis</i> DSM 9187
142	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Teredinibacter turnerae</i> T7901
143	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Colwellia psychrerythraea</i> 34H
144	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Pseudoalteromonas atlantica</i> T6c
145	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Psychromonas ingrahamii</i> 37
146	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Shewanella frigidimarina</i> NCIMB 400
147	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Shewanella halifaxensis</i> HAW EB4
148	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Dichelobacter nodosus</i> VCS1703A
149	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Allochromatium vinosum</i> DSM 180
150	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Nitrosococcus oceani</i> ATCC 19707
151	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Buchnera aphidicola</i>
152	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Buchnera aphidicola</i> 5A
153	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	Candidatus <i>Blochmannia floridanus</i>
154	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	Candidatus <i>Blochmannia pennsylvanicus</i>
155	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Escherichia coli</i> K 12 substr MG1655
156	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Photorhabdus asymbiotica</i>
157	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Sodalis glossinidius morsitans</i>
158	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Wigglesworthia glossinidia</i>
159	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Yersinia pestis</i> Pestoides F
160	<i>Bacteria</i>	<i>Proteobacteria</i>	<i><math>\gamma</math>-proteobacteria</i>	<i>Coxiella burnetii</i>



## Supplementary material

161	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Legionella longbeachae</i> NSW150
162	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Legionella pneumophila</i> Corby
163	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Methylococcus capsulatus</i> Bath
164	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Marinomonas</i> MWYL1
165	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Haemophilus influenzae</i>
166	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Acinetobacter sp</i> ADP1
167	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Azotobacter vinelandii</i> DJ
168	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Francisella philomiragia</i> ATCC 25017
169	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Thiomicrospira crunogena</i> XCL-2
170	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Vibrio cholerae</i>
171	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Xanthomonas albilineans</i>
172	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>γ-proteobacteria</i>	<i>Xylella fastidiosa</i>
173	<i>Bacteria</i>	<i>Proteobacteria</i>	<i>unclassified</i>	<i>Magnetococcus</i> MC-1
174	<i>Bacteria</i>	<i>Spirochaetes</i>	<i>Spirochaetes</i>	<i>Brachyspira hyodysenteriae</i> WA1
175	<i>Bacteria</i>	<i>Spirochaetes</i>	<i>Spirochaetes</i>	<i>Borrelia afzelii</i> PKo
176	<i>Bacteria</i>	<i>Spirochaetes</i>	<i>Spirochaetes</i>	<i>Treponema pallidum</i>
177	<i>Bacteria</i>	<i>Thermotogae</i>	<i>Thermotogae</i>	<i>Petrotoga mobilis</i> SJ95
178	<i>Bacteria</i>	<i>Thermotogae</i>	<i>Thermotogae</i>	<i>Thermotoga maritima</i>
179	<i>Bacteria</i>	<i>unclassified Bacteria</i>	<i>Thermobaculum</i>	<i>Thermobaculum terrenum</i> ATCC BAA 798

**Table S3. List of 35 eukaryotic species with fully deciphered genomes that were used upon phylogenomic analysis.**Taxonomy was extracted from the NCBI Taxonomy web service (<http://www.ncbi.nlm.nih.gov/taxonomy>).

#	Super phylum	Phylum	Class	Order	Species
1	Alveolata	Apicomplexa	Aconoidasida	Haemosporida	<i>Plasmodium vivax</i> SaI-1
2		Apicomplexa	Coccidia	Eucoccidiorida	<i>Cryptosporidium muris</i> RN66
3		Apicomplexa	Coccidia	Eucoccidiorida	<i>Toxoplasma gondii</i> ME49
4		Ciliophora	Intramacronucleata	Oligohymenophorea	<i>Paramecium tetraurelia</i> strain d4-2
5	Amoebozoa	Archamoebae	Entamoebidae	Entamoeba	<i>Entamoeba histolytica</i> HM-1:IMSS
6	Amoebozoa	Mycetozoa	Dictyosteliida	Dictyostelium	<i>Dictyostelium discoideum</i> AX4
7	Euglenozoa	Kinetoplastida	Trypanosomatidae	Trypanosoma	<i>Trypanosoma brucei</i> TREU927
8	Opisthokonta	Choanoflagellida	Codonosigidae	Monosiga	<i>Monosiga brevicollis</i> MX1
9		Fungi	Dikarya	Ascomycota	<i>Penicillium marneffei</i> ATCC 18224
10		Fungi	Dikarya	Ascomycota	<i>Saccharomyces cerevisiae</i> S288c
11		Fungi	Dikarya	Basidiomycota	<i>Puccinia graminis</i>
12		Fungi	Dikarya	Basidiomycota	<i>Ustilago maydis</i> 521
13		Fungi	Microsporidia	Apansporoblastina	<i>Encephalitozoon cuniculi</i> GB-M1
14		Metazoa	Eumetazoa	Bilateria	<i>Gallus gallus</i>
15		Metazoa	Eumetazoa	Bilateria	<i>Homo sapiens</i>
16		Metazoa	Eumetazoa	Bilateria	<i>Xenopus laevis</i>
17		Metazoa	Eumetazoa	Bilateria	<i>Danio rerio</i>
18		Metazoa	Eumetazoa	Bilateria	<i>Ciona intestinalis</i>
19		Metazoa	Eumetazoa	Bilateria	<i>Strongylocentrotus purpuratus</i>
20		Metazoa	Eumetazoa	Bilateria	<i>Saccoglossus kowalevskii</i>
21		Metazoa	Eumetazoa	Bilateria	<i>Apis mellifera</i>
22		Metazoa	Eumetazoa	Bilateria	<i>Ixodes scapularis</i>
23		Metazoa	Eumetazoa	Bilateria	<i>Drosophila melanogaster</i>
24		Metazoa	Eumetazoa	Bilateria	<i>Caenorhabditis elegans</i>
25		Metazoa	Eumetazoa	Cnidaria	<i>Nematostella vectensis</i>
26		Metazoa	Eumetazoa	Cnidaria	<i>Hydra magnipapillata</i>
27		Metazoa	Placozoa	Trichoplax	<i>Trichoplax adhaerens</i>
28	Parabasalia	Trichomonadida	Trichomonadidae	Trichomonas	<i>Trichomonas vaginalis</i> G3
29	Stramenopiles	Bacillariophyta	Bacillariophyceae	Bacillariophycidae	<i>Phaeodactylum tricorutum</i>
30	Stramenopiles	Bacillariophyta	Coscinodiscophyceae	Thalassiosirophycidae	<i>Thalassiosira pseudonana</i>
31	Viridiplantae	Chlorophyta	Chlorophyceae	Chlamydomonadales	<i>Volvox carteri f. nagariensis</i>
32		Chlorophyta	Mamiellophyceae	Mamiellales	<i>Micromonas sp.</i> RCC299
33		Streptophyta	Streptophytina	Embryophyta	<i>Physcomitrella patens</i>
34		Streptophyta	Streptophytina	Embryophyta	<i>Arabidopsis thaliana</i>
35		Streptophyta	Streptophytina	Embryophyta	<i>Selaginella moellendorffii</i>

**Table S4. List of 33 proteobacterial sequences with fully deciphered genomes that were used upon phylogenomic analysis.**Taxonomy was extracted from the NCBI Taxonomy web service (<http://www.ncbi.nlm.nih.gov/taxonomy>).

#	Phylum	Class	Order	Family	Species	
1	Proteobacteria	$\alpha$	Caulobacterales	Caulobacteraceae	<i>Phenylobacterium zucineum</i> HLK1	
2			Rhizobiales	Bradyrhizobiaceae	<i>Bradyrhizobium japonicum</i> USDA 110	
3			Rhizobiales	Rhizobiaceae	<i>Agrobacterium radiobacter</i> K84	
4			Rhodobacterales	Rhodobacteraceae	<i>Rhodobacter capsulatus</i> SB 1003	
5			Rhodospirillales	Rhodospirillaceae	<i>Magnetospirillum magneticum</i> AMB 1	
6			Rickettsiales	Anaplasmataceae	<i>Neorickettsia risticii</i> Illinois	
7			Rickettsiales	Rickettsiaceae	<i>Rickettsia typhi</i> Wilmington	
8			Sphingomonadales	Sphingomonadaceae	<i>Zymomonas mobilis</i> ZM4	
9		$\beta$	Burkholderiales	Alcaligenaceae	<i>Bordetella avium</i> 197N	
10			Burkholderiales	Oxalobacteraceae	<i>Collimonas fungivorans</i> Ter331	
11			Hydrogenophilales	Hydrogenophilaceae	<i>Thiobacillus denitrificans</i> ATCC 25259	
12			Nitrosomonadales	Nitrosomonadaceae	<i>Nitrosomonas europaea</i> ATCC 19718	
13		$\delta/\epsilon$	Deltaproteobacteria	Bdellovibrionales	<i>Bacteriovorax marinus</i> SJ	
14			Deltaproteobacteria	Desulfurellales	<i>Hippea maritima</i> DSM 10411	
15			Deltaproteobacteria	Desulfuromonadales	<i>Geobacter uraniireducens</i> Rf4	
16			Deltaproteobacteria	Myxococcales	<i>Haliangium ochraceum</i> DSM 14365	
17			Epsilonproteobacteria	Campylobacterales	<i>Helicobacter pylori</i> 26695	
18			Epsilonproteobacteria	Campylobacterales	<i>Sulfuricurvum kujiense</i> DSM 16994	
19			Epsilonproteobacteria	Nautiliales	<i>Nautilia profundicola</i> AmH	
20		$\gamma$	Aeromonadales	Aeromonadaceae	<i>Tolumonas auensis</i> DSM 9187	
21			Alteromonadales	Idiomarinaceae	<i>Idiomarina loihiensis</i> L2TR	
22			Alteromonadales	Shewanellaceae	<i>Shewanella frigidimarina</i> NCIMB 400	
23			Chromatiales	Chromatiaceae	<i>Nitrosococcus watsonii</i> C 113	
24			Chromatiales	Halothiobacillaceae	<i>Halothiobacillus neapolitanus</i> c2	
25			Enterobacteriales	Enterobacteriaceae	<i>Escherichia coli</i> K 12	
26			Enterobacteriales	Enterobacteriaceae	<i>Yersinia pestis</i> biovar Microtus	
27			Pasteurellales	Pasteurellaceae	<i>Haemophilus influenzae</i> 86 028NP	
28			Pseudomonadales	Pseudomonadaceae	<i>Azotobacter vinelandii</i> DJ	
29			Vibrionales	Vibrionaceae	<i>Vibrio cholerae</i> M66 2	
30		unclassified	Magnetococcus	<i>Magnetococcus</i> sp. MC-1	<i>Magnetococcus</i> MC 1	
31		Aquificae	Aquificae	Aquificales	Aquificaceae	<i>Aquifex aeolicus</i> VF5
32		Aquificae	Aquificae	Aquificales	Desulfurobacteriaceae	<i>Thermovibrio ammonificans</i> HB 1
33		Aquificae	Aquificae	Aquificales	Hydrogenothermaceae	<i>Persephonella marina</i> EX H1

### 13. Acknowledgements

First of all, I would like to thank PD Dr. Armen Y. Mulkidjanian for the opportunity to work on this exciting project, for all effort he put into me, for the ongoing help in understanding scientific method of thinking and writing, for the knowledge and spirit in bioenergetics which could only be given upon face-to-face communication.

I would like to thank Prof. Dr. Heinz-Jürgen Steinhoff for his continuous interest in my work and the possibility to use the infrastructure of his department.

Also I express my deep gratitude to Dr. Eugene V. Koonin from National Center of Biotechnology Information, NLM, NIH, who has kindly invited me to visit his laboratory in United States and thus offered me an opportunity to learn from the great specialists in bioinformatics and wonderful people. Except himself, my teachers in this field were Dr. Kira S. Makarova, Dr. Michael Y. Galperin and Dr. Yuri I. Wolf, to whom I address my thanks.

I would like to thank all people from the groups of PD Dr. Armen Y. Mulkidjanian and Prof. Dr. Heinz-Jürgen Steinhoff for their critical comments and ongoing discussions of my work during these 3 years. My special thanks are addressed to Alexei Lokhmatikov, who critically red the manuscript of this thesis.

I would like to remember here my first scientific supervisors, Dr. Vladimir K. Nikolaev and Dr. Vladimir V. Galatenko who have done their best in teaching me programming. Their friendly attitude and a sense of humor have shaped my intentions towards making a carrier in science.

I want to thank Dr. Stanislav Y. Rykov for his ongoing belief in me, for helping me in proofreading the manuscript of this thesis, and for much more.

Finally, I would like to deeply thank my parents, Vladimir V. Dibrov and Dr. Tat'yana V. Zhavoronkova, for everything they have done for me in my life. I can not mention all my family members here, but I express my gratitude to them all. I specifically thank my elder brother Sergey V. Zhavoronkov for the first lessons of critical thinking and argumentation, as well as for his ongoing support. My aunt Irina V. Zhavoronkova have taught me a lot of important daily life skills, and always treated me with kindness and love. My grandmother Nina M. Zhavoronkova was for me always an example of the person with steel will, but combined with a hot heart. My grandfather Victor D. Zhavoronkov is the most honest and diligent man I ever met. I am endlessly proud of them both.

## 14. Publications

### Publications in peer-reviewed journals:

1. **D.V. Dibrova**, M.Y. Galperin, A.Y. Mulkidjanian (2010) **Characterization of the N-ATPase, a distinct, laterally transferred Na<sup>+</sup>-translocating form of the bacterial F-type membrane ATPase.** *Bioinformatics*, 26, 1473-1476.
2. A.Y. Mulkidjanian, A.Y. Bychkov, **D.V. Dibrova**, M.Y. Galperin, E.V. Koonin (2012) **Origin of first cells at terrestrial, anoxic geothermal fields.** *Proc Natl Acad Sci USA*, 109, E821-830.
3. **D.V. Dibrova**, M.Y. Chudetsky, M.Y. Galperin, E.V. Koonin, A.Y. Mulkidjanian (2012) **Role of energy in the emergence of biology from chemistry.** *Orig Life Evol Biosph*, 42(5), 459-68.
4. A.Y. Mulkidjanian, A.Yu. Bychkov, **D.V. Dibrova**, M.Y. Galperin and E.V. Koonin. (2012) **Open questions on the origin of life at anoxic geothermal fields.** *Orig Life Evol Biosph*, 42(5), 507-16.
5. **D.V. Dibrova**, D.A. Cherepanov, M.Y. Galperin, V.P. Skulachev and A.Y. Mulkidjanian (2013) **Evolution of cytochrome *bc* complexes: from membrane electron translocases to triggers of apoptosis.** *Biochim. Biophys. Acta* (Invited Review, accepted).

**Conference talks and posters:**

1. **D.V. Dibrova**, K.S. Makarova, M.Y. Galperin, E.V. Koonin, and A.Y. Mulkidjanian (2011) **Comparative analysis of lipid biosynthesis in Archaea and Bacteria: What was the structure of first membrane lipids?** *Proceedings of the International Moscow Conference on Computational Molecular Biology*, Moscow, Russia, pp. 92-93.
2. **D.V. Dibrova**, M.Y. Galperin, and A.Y. Mulkidjanian (2011) **Evolution of membrane bioenergetics.** *Proceedings of the International Moscow Conference on Computational Molecular Biology*, Moscow, Russia, pp. 241-242.
3. **D.V. Dibrova**, K.S. Makarova, M.Y. Galperin, E.V. Koonin, and A.Y. Mulkidjanian (2012) **Comparative analysis of lipid biosynthesis in archaea, bacteria and eukaryotes: What was the structure of the first membrane lipids?** *Proceedings of the 3<sup>rd</sup> Moscow International Conference "Molecular Phylogenetics"* (Moscow, Russia, July 31-August 4, 2012), Moscow, Russia, pp. 56-57.
4. **D.V. Dibrova**, M.Y. Galperin, and A.Y. Mulkidjanian (2012) **The cytochrome *bc*<sub>1</sub> complex and the evolution of membrane bioenergetics.** *Biochim. Biophys. Acta*, 1817 (Abstracts of the 17th European Bioenergetics Conference in Freiburg, Germany), S91.
5. **D.V. Dibrova**, K.S. Makarova, M.Y. Galperin, E.V. Koonin, and A.Y. Mulkidjanian (2012) **Comparative analysis of lipid biosynthesis in archaea, bacteria and eukaryotes: What was the structure of the first membrane lipids?** *Biochim. Biophys. Acta*, 1817 (Abstracts of the 17th European Bioenergetics Conference in Freiburg, Germany), p. S154.
6. **D.V. Dibrova**, K.S. Makarova, M.Y. Galperin, E.V. Koonin, and A.Y. Mulkidjanian (2013) **Comparative analysis of lipid biosynthesis in bacteria and archaea: an update on co-evolution of membranes and membrane**

- bioenergetics. Proceedings of the 3<sup>rd</sup> D/UK Conference on Bioenergetics (DUKBEC) 2013**, Rauschholzhausen, Germany, P009.
7. **D.V. Dibrova**, M.Y. Galperin, E.V. Koonin, and A.Y. Mulkidjanian (2013) **Lysine and arginine “fingers” could replace K<sup>+</sup> as cofactors of phosphoester bond cleavage in the course of evolution. Proceedings of the 3<sup>rd</sup> D/UK Conference on Bioenergetics (DUKBEC) 2013**, Rauschholzhausen, Germany, P035.

This page is intentionally left blank



## 15. Lebenslauf

### Daria Dibrova

Martinsburg 29, App. B109

49078 Osnabrück

Tel.: 0541 9692680

Email: ddibrova@uos.de

geb. 9. September 1987

in Pjatigorsk, Russland

#### **Familienstand**

ledig, keine Kinder

**1994-2004**

#### **Schule**

Gymnasium Nr. 4, Pjatigorsk

(Auszeichnung: "Goldene Medaille für Prüfungsleistungen")

**09/2004 - 07/2009**

#### **Studium**

Fakultät für Bioengineering und Bioinformatik

der Lomonossow-Staatliche Universität Moskau

*Hauptfach:* Bioinformatik

*Nebenfach:* Bioengineering

Diplom mit Auszeichnung

**05/2009**

#### **Thema der Diplomarbeit:**

*"Bioinformatische Analyse von Zwei-Komponenten-Systemen"*

06/2007

**Studienbegleitende Praktika**

Bioinformatik- Praktikum (Betreuer: Dr. J.J. Goeman),  
Abteilung für medizinische Statistik, medizinisches Zentrum  
der Universität Leiden, Leiden, die Niederlande

ab 10/2009

**Promotionsstudium**

Promotionstudentin bei PD Dr. A. Mulkidjanian, Fachbereich  
Physik, Universität Osnabrück

03/2010 - 04/2010

**Promotionsbegleitende Weiterqualifizierung**

Gastaufenthalt in der AG von Dr. E. Koonin, *National Center  
for Biotechnology Information at National Library of  
Medicine at National Institutes of Health*, Bethesda,  
Vereinigte Staaten

**Sprachkenntnisse**

Russisch (Muttersprache)

Englisch (fließend)

Deutsch (Grundkenntnisse)

Französisch (Grundkenntnisse)

**Programmiersprachenkenntnisse**

C++ (gute Kenntnisse)

Perl (gute Kenntnisse)

Java (Grundkenntnisse)

R (Grundkenntnisse)

Osnabrück, den

---

## 16. Erklärung über die Eigenständigkeit

Ich habe die Dissertation selbständig angefertigt und mich außer der angegebenen keiner weiteren Hilfsmittel bedient. Alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen wurden, sind als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen.

Ich erkläre, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

.....  
(Ort, Datum)

.....  
(Unterschrift)