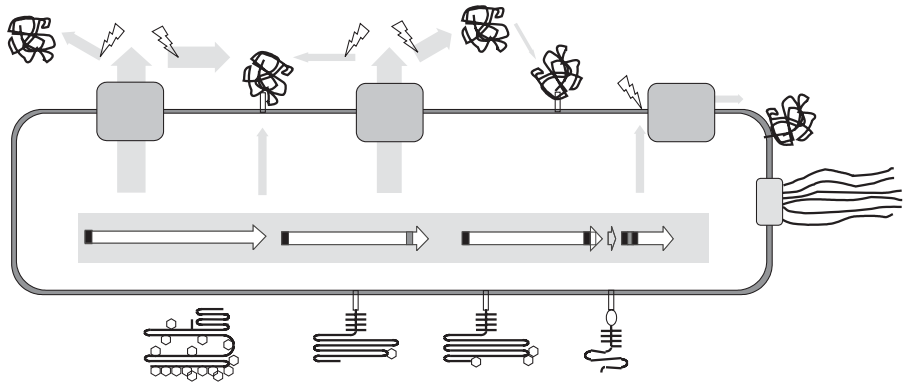
Computational Genome and Pathway Analysis of Halophilic Archaea



Dissertation

Michaela Falb

aus

Heiligenstadt

(Eichsfeld)

2005

Dissertation

zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Computational Genome and Pathway Analysis of Halophilic Archaea

Michaela Falb

aus
Heiligenstadt
(Eichsfeld)

2005

Erklärung

Diese Dissertation wurde im Sinne von §13 Abs. 3 bzw. 4 der Promotionsordnung vom 29. Januar 1998 von Prof. Dr. Dieter Oesterhelt betreut.

Ehrenwörtliche Versicherung

Diese Dissertation wurde selbständig, ohne unerlaubte Hilfe erarbeitet.

München, den 31. August 2005

Michaela Falb

Dissertation eingereicht am: 01.09.2005

1. Gutachter: Prof. Dr. Dieter Oesterhelt
2. Gutachter: Prof. Dr. Erich Bornberg-Bauer

Mündliche Prüfung am: 22.12.2005

CONTENTS

SUMMARY

Halophilic archaea inhabit hypersaline environments and share common physiological features such as acidic protein machineries in order to adapt to high internal salt concentrations as well as electron transport chains for oxidative respiration. Surprisingly, nutritional demands were found to differ considerably amongst haloarchaeal species, though, and in this project several complete genomes of halophilic archaea were analysed to predict their metabolic capabilities. Comparative analysis of gene equipments showed that haloarchaea adopted several strategies to utilize abundant cell material available in brines such as the acquisition of catabolic enzymes, secretion of hydrolytic enzymes, and elimination of biosynthesis gene clusters. For example, metabolic genes of the well-studied *Halobacterium salinarum* were found to be consistent with the known degradation of glycerol and amino acids. Further, the complex requirement of *H. salinarum* for various amino acids and vitamins in comparison with other halophiles was explained by the lack of several genes and gene clusters, e.g. for the biosynthesis of methionine, lysine, and thiamine. Nitrogen metabolism varied also among halophilic archaea, and the haloalkaliphile *Natronomonas pharaonis* was predicted to apply several modes of N-assimilation to cope with severe ammonium deficiencies in its highly alkaline habitat. This species was experimentally shown to possess a functional respiratory chain, but comparative analysis with several archaea suggests a yet unknown complex III analogue in *N. pharaonis*. Respiratory chains of halophilic and other respiratory archaea were found to share similar genes for pre-quinone electron transfer steps but show great diversity in post-quinone electron transfer steps indicating adaptation to changing environmental conditions in extreme habitats. Finally, secretomes of halophilic and non-halophilic archaea were predicted proposing that haloarchaea secretion proteins are predominantly exported via the twin-arginine pathway and commonly exhibit a lipobox motif for N-terminal lipid anchoring. In *N. pharaonis*, lipobox-containing proteins were most frequent suggesting that lipid anchoring might prevent protein extraction under alkaline conditions. By contrast, non-halophilic archaea seem to prefer the general secretion pathway for protein translocation and to retain only few secretion proteins by N-terminal lipid anchors. Membrane attachment was preferentially observed for interacting components of ABC transporters and respiratory chains and might further occur via postulated C-terminal anchors in archaea.

Within this project, the complete genome of the newly sequenced *N. pharaonis* was analysed with focus on curation of automatically generated data in order to retrieve reliable gene prediction and protein function assignment results as a basis for additional studies. Through the development of a post-processing routine and expert validation as well as by integration of proteomics data, a highly reliable gene set was created for *N. pharaonis* which was subsequently used to assess various microbial gene finders. This showed that all automatic gene tools predicted a rather correct gene set for the GC-rich *N. pharaonis* genome but produced insufficient results in respect to their start codon assignments. Available proteomics results for *N. pharaonis* and *H. salinarum* were further analysed for post-translational modifications, and N-terminal peptides of haloarchaeal proteins were found to be commonly processed by N-terminal methionine cleavage and to some extent further modified by N-acetylation. For general function assignment of predicted *N. pharaonis* proteins and for enzyme assignment in *H. salinarum*, similarity-based searches, gene-context methods such as neighbourhood analysis but also manual curation were applied in order to reduce the number of hypothetical proteins and to avoid cross-species transfer of misassigned functions. This permitted to reliably reconstruct the metabolism of *H. salinarum* and *N. pharaonis*. Generated metabolic data were stored in a newly developed metabolic database that also integrates experimental data retrieved from the literature. The pathway data can be assessed as coloured KEGG maps and were combined with data resulting from transcriptomics and proteomics techniques. In future, expert-curated reaction entries of the created metabolic database will be a valuable source for the design of metabolic experiments and will deliver a reliable input for metabolic models of halophilic archaea.

CHAPTER 1

Introduction to Halophilic Archaea

1.1 Hypersaline environments

Extremely halophilic archaea are a diverse group of euryarchaeota that inhabit highly saline (hypersaline) environments such as salt lakes, salt ponds and marine salterns (Figure 1.1), which are found in hot, dry areas of the world. However, haloarchaea have also been isolated from the surfaces of proteinaceous products like fish that is heavily salted with solar salt. Salt lakes can vary considerably in ionic composition, and the predominant ions depend on the surrounding topography, geology, and general climatic conditions (Table 1.1) (Madigan et al. 2000). While the Great Salt Lake in Utah (USA) is essentially 10-fold concentrated sea water with sodium and chloride as main cation and anion and significant sulfate levels, the Dead Sea (Israel) is relatively low in sodium and sulfate. Instead, high levels of magnesium and calcium are found due to the minerals in the surrounding rocks. The water chemistry of soda lakes like Lake Magadi in the African Rift Valley (Kenya) (Figure 1.1) resembles that of the Great Salt Lake, but because of high levels of carbonate in the surrounding rocks, the pH of soda lakes is high and the level of dissolved magnesium and calcium is very low. Marine salterns where seawater evaporates to produce solar salt are also habitats for haloarchaea. From samples of such a salt production plant in Spain, a new

Figure 1.1: **Hypersaline habitats. Left:** Aerial view of the Torrevieja's area (Costa Blanca, Spain), of a series of salterns, where seawater is evaporated to produce solar salt. The red-purple colour is mainly due to bacterioruberins of halophilic archaea such as *Halobacterium*. **Right:** A soda lake, Lake Magadi, in the African Rift Valley (Kenya). The pink colour at the shore of the lake is from haloalkaliphilic archaea, e.g. *Natronomonas*, which thrive under hypersaline and highly alkaline conditions.

Table 1.1: **Ionic composition of selected hypersaline environments** (from Madigan et al. 2000). The concentrations are given in g/l.

Ion	Great Salt Lake	Dead Sea	Typical soda lake	Seawater (for comparison)
Na^+	105	40	142	11
K^+	7	8	2	0.4
Mg^{2+}	11	44	< 0.1	1.3
Ca^{2+}	0.3	17	< 0.1	0.4
Cl^-	181	225	155	19
SO_4^{2-}	27	0.5	23	3
HCO_3^-/CO_3^{2-}	0.7	0.2	67	0.1
pH	7.7	6.1	11	8.1

square-shaped haloarchaeal species that thrives at the limits of water activity has recently be isolated and subsequently cultured (Figure 1.2) (Bolhuis et al. 2004).

Haloarchaea often occur in such large numbers in salt ponds, so that the waters of brines turn a red colour, which comes from their carotenoid pigments. Most of these haloarchaeal carotenoids are bacterioruberins (C50 carotenoids) (Oren 2002, pp. 179-183) that protect the cell against the effects of visible and ultraviolet light and probably reinforce the cell membrane. Massive growth of extreme halophiles (bloom) may spoil the produced salt and without further processing, a gram of solar salt may contain millions of viable cells surviving for several years under practical storage conditions (Tansill 1984). Halophiles were even isolated from a brine inclusion within a 250-million-year-old salt crystal (Vreeland et al. 2000). Rock salt sample, e.g. from Alpine salt sediments, yield also amplifiable DNA with 16S rRNA genes from haloarchaeal species suggesting that these already populated an ancient hypersaline ocean (Radax et al. 2001).

Despite rather harsh conditions in salt and soda lakes, these can be highly productive ecosystems where not only haloarchaea but also halophilic eukaryotes and bacteria are found. *Dunaliella,* an eukaryotic alga with the ability to produce glycerol, is the major primary

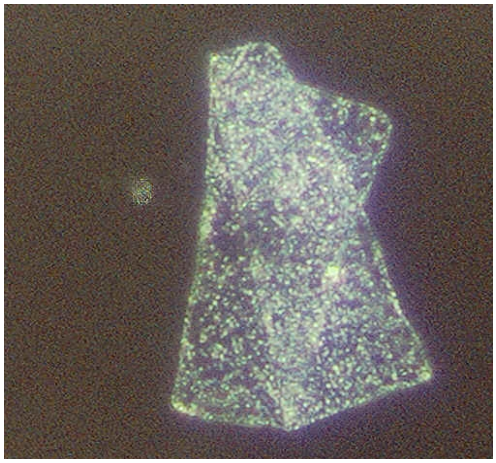


Figure 1. 2: **Phase-contrast microscopic image of *Haloquadratum walsbyi*** (from Bolhuis et al. 2004). This species were first described by Walsby (1980) who observed square cells of normal sizes (5x5 µm) (small cell on the left). In culture, however, very large cells and cell aggregates are found (large folded cell on the right).

oxygenic phototroph in most salt lakes. At higher salinity, cyanobacteria such as *Aphanothece* emerge. In contrast to salt lakes, anoxygenic phototrophic purple bacteria of the genus *Halorhodospira* predominate in soda lakes where *Dunaliella* is absent (Madigan et al. 2000). Organic matter originating from primary production by oxygenic or anoxygenic phototrophs enable the massive growth of chemoorganotrophic haloarchaea at extreme salinities. In addition, extremely halophilic methanogenes, e.g. *Methanohalophilus halophilus* and *Methanohalobium evestigatum*, the latter exhibiting an NaCl optimum of 4.5 M, have been found in hypersaline environments. The sediment of salt lakes is further inhabited by halophilic anaerobic bacteria such as *Haloanaerobacter* and *Halobacteroides* (DasSarma and Arora 2001).

1.2 Taxononomy of halophilic archaea

Organisms that grow at high salt concentrations are found in all three domains of life. These halophiles are often closely related to non-halophilic species though suggesting that adaptation to life in hypersaline environments has arisen many times during the evolution (Oren 2002, pp. 23-24). However, there are few families that consist entirely of halophiles; the families *Haloanaerobiaceae* and *Halobacteroidaceae* in the domain of bacteria and *Halobacteriaceae* in the domain of archaea. Although several halophilic methanogenes are known, only members of the *Halobacteriaceae* are usually considered as halophilic archaea or haloarchaea. Up to now 22 genera of *Halobacteriaceae* have been defined by rRNA sequencing and other criteria (NCBI taxonomy website). The family is often collectively referred to as “halobacteria” since the genus *Halobacterium* was the first halophilic archaeon to be described and remains the best-studied representative of the group.

In 1919, Kleebahn was the first to describe a salt-requiring, red-coloured, pleomorphic, rod-shaped bacterium isolated from salted fish, which he named *Bacillus halobius ruber* (Tansill 1984) and which was later renamed to *Halobacterium halobium*. Several other closely related *Halobacterium* strains were isolated in 1922 (*H. salinarium*) and 1934 (*H. cutirubrum*) causing much disagreement whether these should be upheld as separate species. In 1980, it was finally reported that the three strains exhibit identical 16S RNAs, therefore all of them were listed under the species description of *H. salinarium* (Tansill 1984). After that, the species name was later once again changed to *H. salinarum*, and recently the completely sequenced *Halobacterium species* strain NRC-1 was also re-classified as a strain of *H. salinarum* (Gruber et al. 2004).

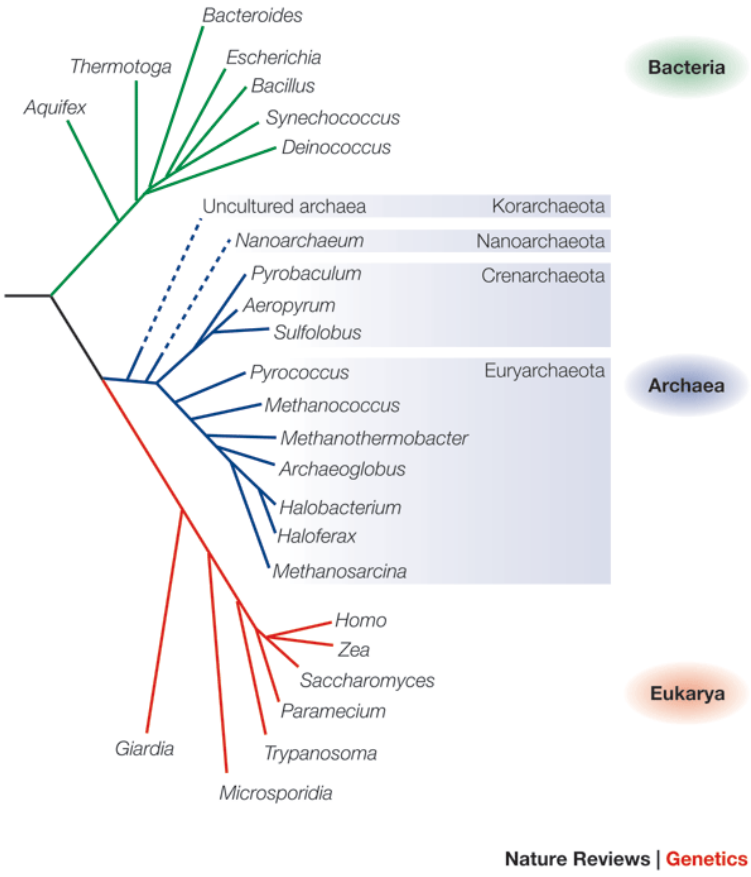


Figure 1.2: The rRNA tree of life (from Allers and Mevarech 2005). Molecular phylogeny clearly distinguishes archaea (blue) as a distinct domain of life independent from bacteria (green) and eukarya (red). Archaea comprise several phyla of which euryarchaeota (blue) are the most diverse group, including all methanogenes and halophiles but also thermophilic as well as psychrophilic species. Members of crenarchaeota are virtually all hyper-thermophiles. Of the remaining phyla, korarchaeota and nanoarchaeota, no or only one species have been cultured so far, thus, their position in the tree is only indicated by dashed branches.

Apart from *Halobacterium* whose retinal proteins and signal transduction have been studied in detail, *Haloferax* and *Haloarcula* species have been investigated in respect to their gas vesicle formation *(Haloferax mediterranei)*, chaperonins *(Haloferax volcanii)* as well as their nitrate *(H. mediterranei, Haloarcula marismortui)* and sugar metabolism *(H. mediterranei, Haloarcula vallismortis)*. Seven genera of *Halobacteriaceae* such as *Natronomonas*, thrive also under alkaliphilic conditions to befit their soda lake habitats reaching pH values of up to 11.

Halobacteriacaea belong to the phylum of euryarchaeota, and can be grouped most closely with *Methanomicrobia (Methanosarcina, Methanohalophilus)*. Euryarchaeota unite a diverse group of archaea characterized by their ability to generate methane (methanogenes) or by their growth in extreme habitats with high salt concentrations (halophiles) or at high (hyperthermophilic) and low (psychrophilic) temperatures (Figure 1.2). In contrast, crenarchaeota comprise almost exclusively hyperthermophilic species. Recently, two more archaeal phyla, korarchaeota and nanoarchaeota, have been suggested, but so far only *Nanoarchaeum equitans* with a genome size of only 0.49 Mb has been cultured and sequenced. Although archaea are often predominant under extreme environmental conditions as they might have existed on the early (archaean) earth, recent environmental studies have shown that not all archaea are extremophiles but are much more widespread than previously thought (Allers and Mevarech 2005).

Archaea have long been considered as bacteria due to their prokaryotic morphology, circular genomes, and gene organization in operons, but in 1977 Woese could clearly distinguish archaea as a third domain of life by applying rRNA phylogeny (Woese and Fox 1977). Their status as a separate domain is further supported by their unique features such as a distinctive cell membrane containing prenyl side chains that are ether-linked to *sn*-glycerol 1-phosphate.

Genome-sequencing projects gave further insights into the nature of archaea and it became clear that this is not completely represented by the rRNA tree, since archaea might have acquired many of their genes by horizontal gene transfer (Koonin et al. 2001). For example, *Methanosarcina mazei* exhibits a large genome that contains 30% genes with bacterial origin (Deppenmeier et al. 2002). Especially metabolically diverse methanogenic and halophilic archaea that cohabit with bacteria may have gained many genes for novel metabolic functions (Deppenmeier et al. 2002; Rosenshine et al. 1989). For example, respiratory-chain genes that enable aerobic life of halophilic archaea are thought to have been acquired from bacteria. A systematic analysis that estimates the extent of horizontal gene transfer among haloarchaea has to be yet conducted in future though.

The deciphering of archaeal genomes also showed that archaea unite bacterial and eukaryotic features. While their information-processing machinery resembles eukaryotic systems (for details see next subchapter), their core metabolic functions are more similar to bacteria (Allers and Mevarech 2005). However, there are also informational as well as operational genes that are unique to archaea such as the nucleoic protein Alba and methanogenesis enzymes, respectively. Archaea also frequently use variant, mostly non-orthologous enzymes, which can be explained by independent, convergent evolution. Considering that as much as 50% of the archaeal genes have no clear functions and many genomes lack enzyme genes for important central pathways such as the pentose-phosphate pathway, many novel archaeal-specific features await discovery (Allers and Mevarech 2005).

1.3 Information processing in archaea

Zillig and colleagues first showed that archaea use DNA-dependent RNA polymerases that are similar to those found in eukaryotes (Huet et al. 1983). As in eukaryota, the archaeal enzyme requires further basal factors for efficient promoter recognition including TATA-box binding protein (TBP) and transcription factor B (TFB) (Figure 1.3 in Chapter 1.4.1) (Bell and Jackson 2001). Many archaea contain several homologs of these factors that might have distinct roles in transcription (Table 1.2 in Chapter 1.5). For example, one of the *H. volcanii*

transcription factor B genes was upregulated as a response to heat shock (Thompson et al. 1999). Interestingly, archaea possess many homologs of bacterial transcription regulators and studies in *Archaeoglobus fulgidus* and *Methanococcus maripaludis* were shown to apply a bacterial mode of transcriptional regulation where a repressor binds at operator sites located close to the promoter (Allers and Mevarech 2005). For the gas-vesicle gene cluster of *H. salinarum*, transcriptional activation has been shown involving binding of a transcriptional activator to a TFB-recognition element (BRE), which might enable direct contact to the basal transcriptional machinery (Hofacker et al. 2004).

Archaeal translation has not been studied intensively but it is clear that core components resemble the ones of eukaryotes. Notably, *H. marismortui* and *H. walsbyi* contain several rRNA operons (Table 1.2) but it was argued that these do not indicate a chimerical nature of these species (Baliga et al. 2004). Translation initiation in archaea requires as much as ten initiation factors and no N-formylmethionine while bacteria require only three initiation factors and formylated methionine. Though, both, archaea and bacteria often group co-regulated genes in transcription units, transcriptional (promoter) and translational elements (Shine-Dalgarno sequence) differ between the two domains. Archaeal promoters seem to be less conserved in sequence and position often exhibiting a BRE motif adjacent to a TATA box and an AT peak around position -10 (Torarinsson et al. 2005). Archaeal Shine-Dalgarno sequences also differ from the bacterial consensus and are found mainly for genes within transcription units. For single genes and first genes of transcription units, translation initiation often operates on leaderless mRNA without Shine-Dalgarno sequences, a mechanism rarely found in bacteria.

Archaea and bacteria share a common genomic structure of a single circular chromosome and optional plasmids, but differ in the machinery used to carry out DNA replication. As with other informational processes such as transcription and translation, the archaeal proteins are more similar to eukaryotic homologs than the bacterial ones. Again, only a subset of eukaryotic proteins are found though, so that the simpler archaeal systems can serve as streamlined model to understand information processes in eukaryotes. Amongst Pyrococci, the predicted origin of replication is highly conserved and located adjacent to a conserved gene that resembles both eukaryotic genes, *cdc6* and *orc1* (Allers and Mevarech 2005). The latter codes for a subunit of the eukaryotic origin recognition complex suggesting that Cdc6/Orc1 functions as the initiator protein for archaeal replication. However, *Sulfolobus* species possess three and *Halobacterium* even 10 *cdc6*/*orc1* homologues within their genomes. The origin of the *Halobacterium* chromosome was recently experimentally identified though and is located next to one of the *cdc6*/*orc1* genes around the maximum of the cumulative GC skew plot though (Berquist and DasSarma 2003). This origin of replication was found to be conserved among other haloarchaeal genomes.

1.4 Physiology and metabolism of halophilic archaea

1.4.1 Osmotic adaptation

Members of the *Halobacteriaceae* require 2-4 M (12-23%) NaCl for optimal growth (Table 1.2 in Chapter 1.5) and most halophilic archaea thrive up to the limit of saturation for sodium chloride (around 5.5 M (32%) NaCl), although growth of some species is rather slow at this salinity. Haloarchaea are unable to grow below concentrations of 1.5 M (9%) NaCl, and *H. salinarum* has been shown to require large amounts of sodium. Sodium ions are, for example, needed, e.g. for cell wall integrity and many transport processes in *Halobacterium*, and cannot be replaced by other ions such as potassium (Madigan et al. 2000).

In contrast to most halophiles that accumulate or synthesize intracellular organic compounds (compatible solutes) to withstand the osmotic pressure that accompanies life in hypersaline environments, *Halobacterium* produces no compatible solutes. Instead, it pumps large amounts of potassium into the cytoplasm (salt-in strategy) so that the intracellular K^+ concentration is considerable higher than the extracellular Na^+ concentration. Thus, the inorganic potassium is employed as a compatible solute that keeps the cell in positive water balance and counteracts the tendency of the cell to become dehydrated at high osmotic pressure or ionic strength. However, under high ionic strength proteins tend to aggregate and often lose their activity, so that the complete intracellular machinery of haloarchaea requires adaptation to such an environment. It has been shown, for example, that the ribosomes of *Halobacterium* require high potassium levels for stability and halophilic enzymes exhibit highly polar surfaces in order to remain in solution (Figure 1.3) (Madigan et al. 2000; Kennedy et al. 2001). Therefore cytoplasmic proteins of halophiles reveal high ratios of acidic amino acids, mainly aspartate residues. Known proteins of haloalkaliphilic species have this typical amino acid distribution pattern, too, but haloalkaliphiles also produce a compatible

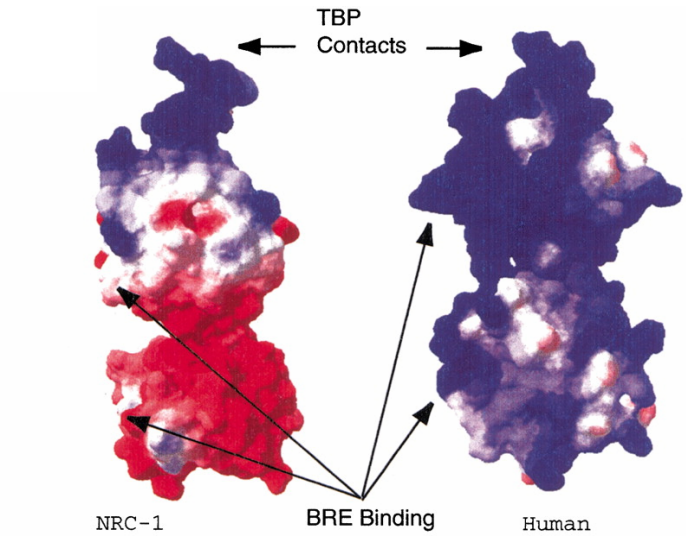


Figure 1.3: **Surface charge comparison** for a halophilic (NRC-1 TFBe) and a non-halophilic protein (Human transcription initiation factor TFIIB) (from Kennedy et al. 2001). Acidic character is indicated by red, basic character is indicated by blue, and neutral areas are indicated by white. The sites for BRE and TBP contacts are indicated within the structural models.

solute, 2-sulfotrehalose, whose concentration increases with the salinity of the medium (Desmarais et al. 1997).

The cell wall of *H. salinarum*, *H. volcanii*, and *Haloarcula japonica* is composed of a glycoprotein (Csg) with exceptionally high contents of acidic amino acids (Sumper 1993; Nakamura et al. 1995). In the extreme halophilic *H. salinarum*, further negative surface charges are introduced by sulfate groups of N-linked saccharide units and a large N-terminally linked, sulphated sugar unit (Figure 4.2 in Chapter 4.1) lacking in the more moderate halophile, *H. volcanii* (Sumper 1993). The negative charges of the acidic amino acids and sugar moieties are shielded by sodium ions, and are absolutely essential for maintaining cellular integrity (Madigan et al. 2000).

1.4.2 Nutritional demands, nutrient transport and sensing

Members of the *Halobacteriaceae* differ greatly in their nutritional demands. While simple growth requirements were first described for *H. mediterranei* and later for other species of the genera *Haloferax* and *Haloarcula*, *H. salinarum* exhibits very complex nutritional demands. Although designed synthetic media for this widely studied species contain 10 to 21 amino acids, vitamin supplements (folate, biotin, thiamine), and sometimes also glycerol (Oesterhelt and Krippahl 1973; Grey and Fitt 1976), growth curves often do not show a typical exponential growth phase (Oren 2002, pp. 125-126). There are also indications that even rich media based on yeast extract lacks some compounds to grow certain haloarchaeal strains, and it was reported that growth often improved when the medium is supplemented with a lysate of *H. salinarum* cells. In contrast, the synthetic medium for *H. volcanii* contains apart from simple carbon sources only the stimulatory vitamins thiamine and folate, inorganic salts as well as ammonium as nitrogen source (Kauri et al. 1990).

Most halophilic archaea preferentially use amino acids as carbon and energy source, but utilize also other compounds of hypersaline habitats such as glycerol and tricarboxylic acid (TCA) cycle intermediates that are excreted by *Dunaliella* and the cyanobacterium *Microcoleus chtonoplastes*, respectively (Oren 2002, pp. 125-126). For example, the synthetic medium for *H. volcanii* contains succinate and glycerol as carbon sources (Kauri et al. 1990). Sugars such as glucose, fructose, and sucrose are catabolized only by some haloarchaea, such as *H. mediterranei* and *Halorubrum saccharovorum* (Altekar and Rangaswamy 1992; Rawal et al. 1988). *Halobacterium* does not grow on sugars but growth of is stimulated by the addition of carbohydrates to the medium (Oren 2002, pp. 128-138) and glucose can be transformed to gluconate (Sonawat et al. 1990). Oxidation of carbohydrates is often incomplete, and *H. saccharovorum* was found to excrete acetate and pyruvate when grown on various sugars (Oren 2002, pp. 128-138). Acetate can also be metabolized but it was found that acetate is used very poorly by haloarchaea. Degradation of

fatty acids has not been reported yet but is likely since all genes for the fatty acid β-oxidation pathway are present in the halobacterial genomes. Many species of *Halobacteriaceae* produce exoenzymes for the degradation of polymeric substances, e.g. the alkaline serine protease halolysin of *Halobacterium* (Kamekura et al. 1992) and an α-amylase of *Natronococcus* (Kobayashi et al. 1994). Finally, several haloarchaeal isolates have been described to degrade aliphatic hydrocarbons and aromatic compounds (Oren 2002, pp. 128-138).

Ammonia and nitrate can be assimilated by some haloarchaea, e.g. by *Haloferax* species (Kauri et al. 1990; Martinez-Espinosa et al. 2001), but these ions are scarce in hypersaline environments and especially in soda lakes due to the lack of nitrifying organisms and high pH levels, respectively. Thus, amino acids are generally the preferred nitrogen source of most haloarchaea. In membrane vesicle studies it was shown, that *H. salinarum* facilitates the uptake of several amino acids such as leucine, glutamate, and tyrosine, mostly dependent on sodium (Oren 2002, p. 127). *Halobacterium* further correlates intracellular pools of amino acids where glutamate and aspartate are most prominent with their rate of transport. Membrane transport systems for acetate and propionate have been studied in the haloalkaliphile *Natronococcus occultus* involving amongst others a sodium-dependent high-affinity transporter. Glucose and fructose transport of *H. volcanii* is also sodium-driven (Oren 2002, p. 127).

H. salinarum is able to sense branched amino acids, methionine, cysteine, arginine, and several peptides and to move toward attractant signals. Arginine chemotaxis is enabled by the cytoplasmic transducer Car, while membrane-bound BasT is involved in sensing of the remaining amino acids (Kokoeva et al. 2002). *H. salinarum* contains several further transducers, the functions of some have been elucidated; HemAT/HtrVIII for oxygen-sensing (Hou et al. 2001), HtrI/HtrII for orange/blue light phototaxis (Hoff et al. 1997), and MpcT as proton motive force sensor (Koch and Oesterhelt 2005). *Halobacterium* tranducers trigger a signalling pathway to the flagellar motor which resembles bacterial signalling cascades and involves a typical bacterial-type two-component regulatory system (CheA/CheY). The *Halobacterium* motor is well understood on the functional level and, thus, a dynamic model could be established (Nutsch et al. 2003). However, the protein components of the archaeal motor remain to be identified on the genomic level.

1.4.3 Energy metabolism

Haloarchaea are aerobic chemoorganotrophs that degrade carbon sources such as amino acids, glycerol, and organic acids via the TCA cycle (Ghosh and Sonawat 1998) and a respiratory electron transport chain (Schafer et al. 1996). Due to the low solubility of oxygen in salt-saturated brines, molecular oxygen easily becomes a limiting factor for oxidative

respiration though. Some halophiles are able to cope by the production of gas vesicles that enable floating of the cell towards the water surface. Furthermore, aerotaxis has been observed for *Halobacterium*, which is triggered by the oxygen sensor HemAT. However, many halophiles can also grow anaerobically by using alternative electron acceptors such as dimethylsulfoxide, triethylamine N-oxide, fumarate, or nitrate (Oren 2002, pp. 128-138). While triethylamine N-oxide is often present in fish tissues as an osmotic solute, the ecological relevance of dimethylsulfoxide is unclear. Nitrate dissimilation might also be limited in hypersaline brines since it is unlikely to be regenerated by nitrification.

Halobacterium employs two further modes of energy conservation under anaerobic conditions. First, it is able of photophosphorylation by using the light-driven proton pump bacteriorhodopsin building up proton motive force for ATP generation. The retinal protein bacteriorhodopsin is one of the best studied proteins, and structurally and functionally resembles the rhodopsin of the eye (Figure 1.4). Since biosynthesis of the retinal moiety is oxygen-dependent, trace concentrations of oxygen are required for light-mediated ATP synthesis in *Halobacterium* though (Oesterhelt and Krippahl 1983). As a second possibility to cover energy requirements when grown anaerobically, *Halobacterium* is able to ferment arginine via the arginine deiminase pathway (Hartmann et al. 1980; Ruepp and Soppa 1996). In this pathway, arginine is converted to ornithine and carbamoylphosphate, which is further split into carbon dioxide and ammonia with concomitant ATP production. While plasmid-encoded enzymes for arginine fermentation are uncommon amongst haloarchaea (Oren 2002, pp. 128-138), genes for bacteriorhodopsin and other retinal proteins (halorhodopsin, sensory rhodopsin) were found in several other *Halobacteriaceaea* (Table 1.2).

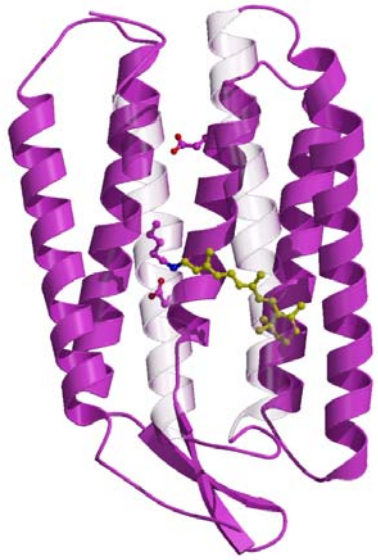


Figure 1.4: **The three-dimensional structure of bacteriorhodopsin** from *H. salinarum* (cross section of the structural model PDB:1BRR) (from Essen et al. 1998). Bacteriorhodopsin functions as a light-mediated proton pump building up a proton motif force that can be used for ATP generation. It contains a retinal molecule (green) that is responsible for absorption of green light with a maximum at 568 nm.

1.5 Genomes of halophilic archaea

Up to date, genomes of six haloarchaeal strains have been completely sequenced (Table 1.2). *Halobacterium salinarum* str. NRC-1 was published in 2000 as *Halobacterium sp.* NRC-1 (Ng et al. 2000), and recently the genome of *Haloarcula marismortui* became available (Baliga et al. 2004). Within our lab complete sequences of three haloarchaea have been finished, *Halobacterium salinarum* str. R1, *Natronomonas pharaonis*, and *Haloquadratum walsbyi*, which were predominantly analysed within this project. Furthermore, *Haloferax volcanii* has been completely sequenced but the genome sequence has also not been published yet. Among ongoing genome projects, there are two further *Halobacteriaceae*, *Halobaculum gomorrense* and the psychrotolerant halophile *Halorubrum lacusprofundi* (Allers and Mevarech 2005). Thus, comparative genomics of halophilic archaea will become even more important in the near future and will give further insights into life in hypersaline environments, especially into the versatile metabolic equipments of halophiles that underlie highly differing nutritional demands that have been described for haloarchaea.

Table 1.2: **Some features of completed haloarchaeal genomes.** Two strains of *H. salinarum* have been sequenced, strains NRC-1 and R1, which mainly differ in their ISH distribution and plasmid arrangements. *H. marismortui* exhibits three ribosomal RNA operons, two on its 3.13-Mb chromosome I, and one rRNA operon on the chromosomal-like megaplasmid designated as chromosome II (CHRII, 0.29 Mb). All halophilic genomes contain retinal proteins with different functions; bacteriorhodopsin (Bop), halorhodopsin (Hop), and sensorhodopsin (Sop).

	H. salinarum strains NRC-1/R1	***H. marismortui***	***H. walsbyi***	***N. pharaonis***
Salt optimum [M]	4-5 M	4.5M	3.3 M	3.5 M (pH 8.5)
Isolation	salted fish	Dead Sea (Israel)	solar saltern (Spain)	soda lake (Egypt)
Main research interests	rhodopsins, signal transduction	versatile nitrogen metabolism	square shape cells, halomucin	haloalkaliphilicity, respiratory chain
Genome size [Mb]	2.61/2.72	4.37	3.24	2.80
# Plasmids	2/4	8 (incl. CHRII)	1	2
%GC chromosome	68.0	62.4	47.9	63.4
rRNA operons	1	3	2	1
fla genes (motility)	yes	yes	no	yes
# transducer	18	21 (18)	0	19
# TBP	1 CHR + 5* PL	1	2	1
# TFB	5 CHR + 4* PL	7 CHR + 3 PL	9	8
# retinal proteins (Bop, Hop, Sop)	4 (1,1,2)	6 (3,1,2)	3 (2,1,0)	2 (0,1,1)

* - In *H. salinarum* str. R1, three plasmid-encoded TATA-box binding proteins (TBP) are duplicated, and two TBPs and one of the transcription factor B (TFB) paralogs are disrupted by ISH elements.

Halobacteriaceae have been thought to exhibit exclusively high GC base ratios of 62% to 68%, but surprisingly *H. walsbyi* exhibits only 48% GC. The GC content of plasmids is usually distinctly lower than the chromosome, e.g. 54-60% vs. 62% in *H. marismortui*. Notably, the high GC-content often complicates genome analysis of high GC species especially the accuracy of gene prediction (McHardy et al. 2004). The genomic organization of *Halobacterium* and *Haloarcula* is highly unusual in that large plasmids contain up to 25-30% of the total cellular DNA. These megaplasmids reveal chromosomal character because they encode functions such as RNAs *(H. marismortui)* and aminoacyl-tRNA ligases *(H. salinarum)* that are essential for survival. Plasmids in *H. marismortui* and *H. salinarum* inhabit also a third to a half of the halophilic insertion sequences (ISH) in the genome, respectively (Baliga et al. 2004; F. Pfeiffer, pers. comm.).

These ISH elements are highly mobile spreading rapidly through halobacterial genomes and are arranged in both, clustered and dispersed fashion. ISH elements account for the observed genetic instability in *Halobacterium* strains that are kept under the laboratory conditions. For example, spontaneous mutations cause gas vesicle defective mutants in a frequency of 10^{-2} and the synthesis of retinal or bacteriorhodopsin precursors is lost at a frequency of 10^{-4} (Oren 2002, pp. 324-331). Differences between the two *H. salinarum* strains, NRC-1 and R1, are also mainly due differences in the distribution of ISH elements in their genomes and not to other sequence differences (only 9 bp differ between the two chromosomes) (F. Pfeiffer, pers. comm.). *Halobacterium* strain GRB, however, is genetically much more stable as *H. salinarum*, since this strain lacks multiple copy families of ISH elements as well as a DNA restriction system that facilitates the introduction of foreign DNA (Ebert et al. 1984; Soppa and Oesterhelt 1989). *H. volcanii* and *H. mediterranei* are also thought to have much more stable genomes and are therefore more amenable to genetic analysis. Within all five completely sequenced haloarchaea, ISH elements are present though, and it is likely that these have been involved in horizontal gene transfer of bacterial genes in order to acquire new metabolic functions.

1.6 Motivation

In order to shed light onto life in halophilic environments, the completely sequenced halophilic genomes of *Halobacterium salinarum* str. R1, *Natronomonas pharaonis*, and *Haloquadratum walsbyi*, were analysed and compared amongst each other and with further haloarchaea. Though gene identification, function assignment, and metabolic pathway reconstruction might be performed straightforward by automatic means, the quality of data is usually not sufficient to establish cellular models thereafter. To avoid accumulation of prediction errors such as gene overprediction and cross-species transfer of misassigned functions (Table 1.3), the focus of this work was set on improving results of consecutive prediction steps that lead from complete genome sequences to biological models of the studied haloarchaea. This is reached by implementing computational procedures that post-process prediction results (e.g. for start codon selection, Chapter 1), by combining results from available tools (e.g. for enzyme assignment, Chapter 5), and by developing new prediction tools (e.g. for secretome analysis, Chapter 3). Apart from applying bioinformatics strategies, misprediction rates can also be reduced by integration of experimental and literature data (e.g. for metabolic pathway reconstruction, Chapter 5). Generated data from genomes and pathway analysis was stored and made available for other scientists, so that it might support the design of experiments and the generation of metabolic models for halophilic archaea in future.

Table 1.3: **Analysis of complete genomes from halophilic archaea.** Problems arising for consecutive prediction steps can be solved by computational approaches and by integration of experimental data. Solutions which have been applied within this project in order to improve prediction results and to obtain high-quality data for future modelling are marked by ticks.

Genome/ pathway analysis step	Problems of automatic prediction tools	Computational approaches	Integration of experimental data	Chapter
Gene annotation	▪ gene overprediction/start codon misassignments due to high GC content	✓ similarity search ✓ gene-context check ✓ protein feature check	✓ proteomics ▪ run-off assays	1
Function assignment	▪ cross-species transfer of misassigned functions ▪ high numbers of hypothetical proteins	✓ motif search ✓ profile search ✓ neighbourhood, fusion, phylogenetic profiles	✓ transcriptomics ✓ quantit. proteomics ▪ interactomics	2 3 4
Metabolic pathway reconstruction	▪ many pathway gaps in archaea	▪ reverse phylogenetic profiles	✓ enz. activity assays ✓ NMR abelling studies ▪ growth experiments	5 6

CHAPTER 2

Gene Prediction and Start Codon Selection in Halophilic Genomes

GC-rich genomes are characterized by a scarcity of stop codons which potentially results in gene overprediction and frequent start codon misassignments. Thus, gene finding results for the newly sequenced GC-rich genome *Natronomonas pharaonis* were processed in order to find valid genes and gene starts. For this purpose, N-terminal alignments with orthologs from the related haloarchaeon *Halobacterium salinarum* were classified and assessed. In case of alignment discrepancies or no orthologs, alignments from other similarity searches, internal features of halophilic proteins as well as the gene context were considered to generate an expert-validated set of genes and gene starts. Mass spectrometry results for *N. pharaonis* and *H. salinarum* were analysed to establish a set of identified *N. pharaonis* proteins and N-terminal peptides. The latter were found to be commonly processed by N-terminal methionine cleavage and to some extent subsequently modified by N-acetylation. The performance of several microbial gene finders for the GC-rich genome *N. pharaonis* was assessed by comparison to the expert- and proteomics-validated gene and gene start set. All gene prediction tools were found to predict a rather correct gene set but to produce insufficient results in respect to start codon assignments. This affects the overall performance of the tested gene finders, since these tend to extend genes above the validated gene lengths and, thus, frequently result in either gene overlap cases or false negative genes.

2.1 Introduction

Genomes with a high GC-content such as the halophilic archaeon *Halobacterium salinarum* (GC 68%) and the newly sequenced *Natronomonas pharaonis* (GC 63%) have a low frequency of stop codons (TAA, TGA, TAG) resulting in a severe overprediction of potential genes (Figure 2.1). A distinctive bias at the dinucleotide level further results in the under-representation of TA sequences and adds to the scarcity of stop codons. As an example, for every gene coding for a real protein in *H. salinarum* there are 1.7 additional spurious open

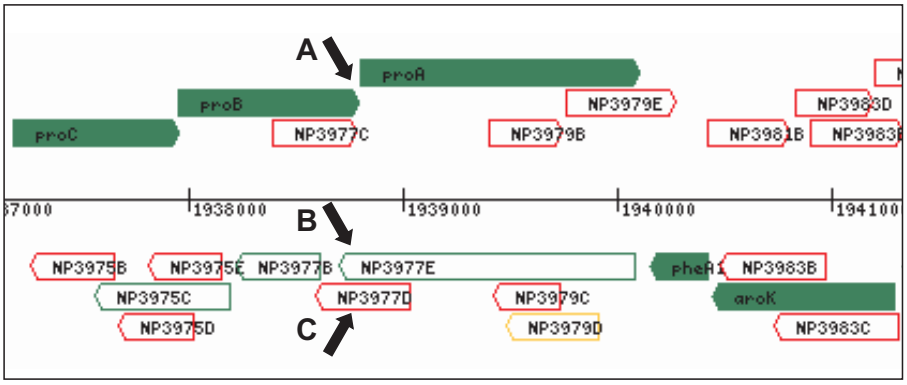


Figure 2.1: **Region view of the *N. pharaonis* proline synthesis cluster** in Halolex. Due to the high GC-content of most haloarchaeal species, several overlapping open reading frames (ORFs, horizontal boxes) were detected at each genome position. Gene finding programs need to assess whether these are potential genes (filled arrow) or spurious ORFs (open arrow). Additionally, the correct gene starts have to be selected. In case of incorrect gene prediction by the available tools, expert validation is required. Potential genes and gene starts of halophilic archaea can be distinguished e.g. by gene context analysis and by isoelectric point (pI) characteristics (green: pI < 6, red: pI > 10): **(A)** Functional genes are characterized by low pI values, and typical short overlaps within transcription units. ORFs that **(B)** extensively overlap with functional genes and/or **(C)** possess high pI values are excluded.

reading frames (ORFs) of at least 100 codons in other frames, reaching lengths of up to 1300 codons (Tebbe et al. 2005). Although circa 40% of all proteins have been already identified for *H. salinarum*, usage of overlapping alternate frames has not been found in a single case yet.

A further problem is the correct assignment of gene start codons, which applies to many microbial genomes. It is aggravated in archaea, since ribosome binding sites (Shine-Dalgarno sequence) around gene starts are poorly conserved and mainly precede genes within transcription units but are not found in case of leaderless mRNAs (Sartorius-Neef and Pfeifer 2004; Torarinsson et al. 2005). The latter are commonly produced from single genes and first genes of transcription units in archaea.

Halophilic archaea possess adapted proteins which carry acidic amino acids at their surfaces in order to remain soluble at high internal salt concentrations. Thus, cytoplasmic proteins show low isoelectric points (pI) and high levels of glutamate and aspartate (Figure 2.2). These intrinsic features of halophilic proteins and the gene context (Figure 2.1) can be analyzed in order to select correct genes and gene starts of the *N. pharaonis* and other haloarchaeal genomes from automatic gene prediction results. Further, by comparative analysis of the halophilic genomes on sequence and genomic levels as well as by the integration of proteomics data, a validated gene set was generated. This then enabled the assessment of microbial gene finders in respect to their gene identification and start codon assignment performance.

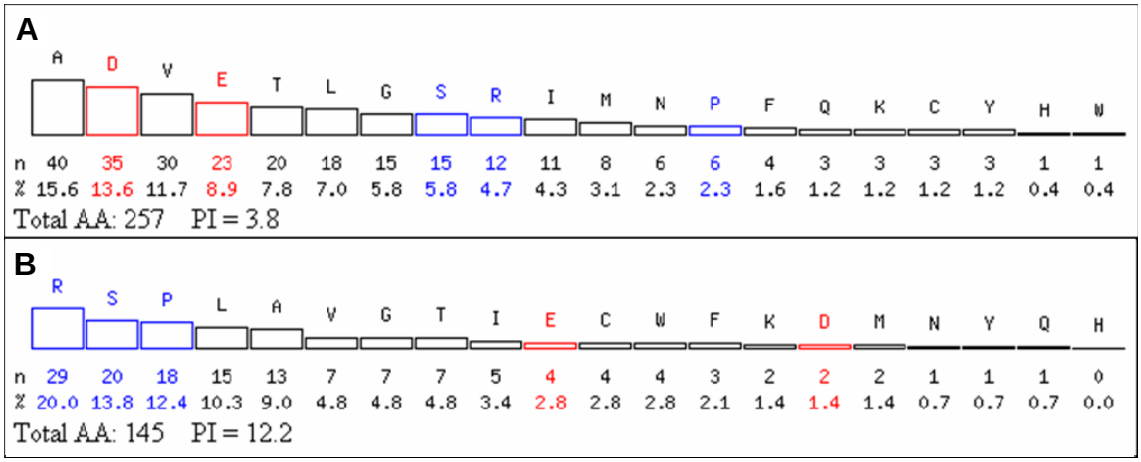


Figure 2.2: **Amino acid distribution and isoelectric points of two selected *N. pharaonis* ORFs.** Due to the high GC content, 16.2% (17.6%) of the *N. pharaonis (H. salinarum)* chromosome would translate *in silico* to Arg. However, validated proteins possess Arg contents around 6.5%, but high percentages of Asp and Glu. **(A)** The validated gene *proC* (see Figure 2.1) possesses many frequent amino acids of halophilic proteins (red). **(B)** In contrast, the spurious ORF NP3975D shows amino acids that would occur in *in silico* translated DNA (blue). Graphics were generated by using a web-based Halolex tool (F. Pfeiffer, pers. comm.).

2.2 Post-processing of gene prediction results by expert validation

In order to predict genes for the newly sequenced *N. pharaonis* genome, Reganor, which integrates two other gene finding programs, Critica and Glimmer, was used (McHardy et al. 2004). Reganor combines the strength of Critica found to be very specific in detection of similarity-supported genes and the more sensitive Glimmer tool that applies a sophisticated *ab initio* approach. However, Glimmer loses its performance for GC-rich genomes mainly due to a specifity loss, but, within Reganor, the specifity of Glimmer is enhanced by using Critica results as trainings set. Reganor further improves gene prediction by removing Glimmer ORFs that overlap Critica predictions and by flagging Glimmer ORFs below a given score (McHardy et al. 2004).

For post-processing of gene finding results from Reganor, sequence comparison data from blast searches, which are usually applied for function annotation of genes, proved to be useful. Predicted genes and start codons were checked in a sequential procedure analyzing blast results from searches against *H. salinarum,* the NCBI nr database, and *N. pharaonis* itself (Figure 2.3). In between the three distinct blast analysis steps, the remaining unchecked genes without homologs were automatically categorized into the set of spurious ORFs in case they overlapped already accepted genes. The so reduced set of yet unchecked genes remains then to be assessed by considering their pI values, amino acid distributions, and gene contexts.

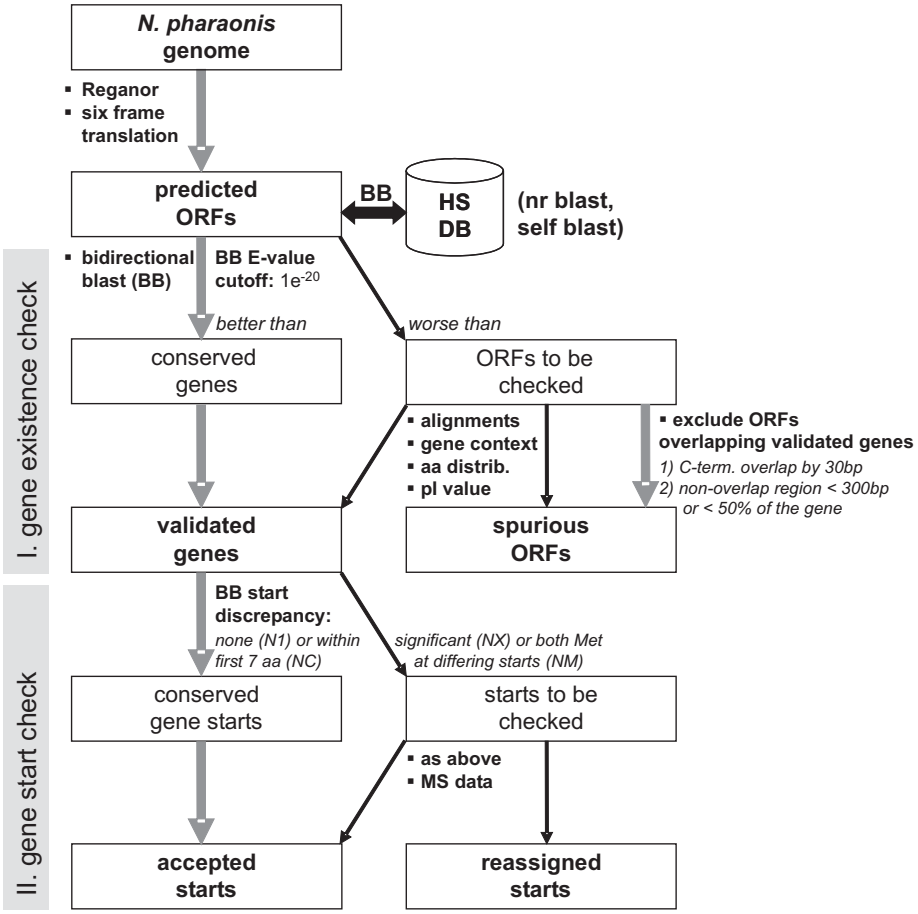


Figure 2.3: **Validation procedure for predicted *N. pharaonis* ORFs** checking the **(I)** existence of genes as well as **(II)** gene start codons by automatic (grey arrows) and manual means (black arrows). Intergenome comparison with the close relative *H. salinarum* (HS DB) by bidirectional blast (BB) enables to identify conserved genes by using an E-value cutoff. The BB alignments were further used to identify correctly assigned start codons which showed no or slight discrepancies between the related protein sequences. In case gene starts of *N. pharaonis* and *H. salinarum* orthologs differ significantly, expert validation was required, e.g. by using further alignments. Some decisions were further supported by mass spectrometry (MS) data (Figure 2.12 in Methods). ORFs that overlap already validated genes (with accepted gene starts) were excluded by an automatic routine. In a second and third assessment round, alignments from nr and self blast searches were analysed, again followed by an overlap check run. The remaining predicted gene set and their starts had to be categorized by assessing gene context and intrinsic features (pI value, amino acid distribution) of the predicted halophilic proteins (Figure 2.1).

The validation routine that utilizes results from bidirectional blastp (BB) search of predicted *N. pharaonis* genes against the *H. salinarum* gene set identified around 2000 orthologous gene pairs (E-value cutoff e^{-20}) between the two related haloarchaea. When comparing starts of the respective blast alignments, over 30% (632) of these gene pairs aligned from position 1 in both sequences (N1) and further 20% (390) of the alignments varied by only 7 amino acids for the two orthologous sequences (NC) (Supplemental Table 2.5). Thus, half of the *N. pharaonis* gene starts in respect to the final gene set could be validated by comparison with *H. salinarum* using the developed automatic procedure.

Over 10% (249) of the alignments between translated *N. pharaonis* and *H. salinarum* genes started from the initial methionine in one amino acid sequence, and an internal residue (M, V) from a potential ATG/GTG start codons in the other sequence. These cases indicate a start codon misassignment in one of the two orthologous sequences. Therefore, the longer sequence was used as query for tblastn and the resulting alignment analyzed. In case the homology extended over the previous start codon, the shorter sequence was extended, otherwise the longer sequence was shortened. Thus, the applied post-processing procedure led not only to the correction of *N. pharaonis* gene predictions, but also to over 200 start reassignments of the *H. salinarum* gene set.

The remaining *N. pharaonis* - *H. salinarum* gene pairs showed significant start discrepancies or aligned only within the C-terminal region of the orthologs. In some cases, this may be due to homology being restricted to a single domain, e.g. in transducer proteins. However, the majority of halophilic proteins have similar length, and gene starts and discrepancies therefore required manual assessment. *N. pharaonis* genes without a *Halobacterium* homolog were assessed by analysing nr and self blast results if available and/or by applying additional criteria such as gene context and predicted intrinsic proteins features.

After completion of post-processing the gene finder results, a final set of 2843 genes (2675 chromosomal genes) was established, to which will be referred to as expert-validated (EV) gene set in the following sections. The final gene set contained 91% ATG, 8% GTG, plus a few atypical starts. GTG starts were only considered in case of evidences from similarity searches or if no alternative ATG codon is detected in the sequence. Atypical start codons were found in circa one percent of the genes, most of them belong to probable pseudogenes or transposase fragments (one third are found on PL131).

The developed validation routine based on sequence comparison with a closely related species can be applied to further newly sequenced genomes, and was used to annotate *Haloquadratum walsbyi* genes by comparison against the gene sets of *N. pharaonis* and *H. salinarum*. However, the homology-based procedure is only useful if gene prediction quality for the related reference species is sufficient, and, thus, many genes can be automatically validated and few discrepancies have to be checked manually.

2.3 Intrinsic features of haloarchaeal proteins and gene context analysis

2.3.1 Isoelectric points and amino acid distribution of halophilic proteins

Known halobacterial proteins reveal characteristic isoelectric points and amino acid distributions, which differ significantly from the pI and amino acid distribution of *in silico* translated DNA from their genomes. Thus, considering the intrinsic protein features of halophilic proteins proved to be useful for the assessment of predicted *H. salinarum*, *N. pharaonis*, and *H. walsbyi* genes and gene starts. However, in certain cases, gene and start codon selection based on pI and amino acid distribution remains ambiguous (see next subchapter).

The relevance of start codon selection by pI values has already been shown previously for *H. salinarum*. There, misassigned start codons were found by detecting protein spots in the two-dimensional (2D) gel with deviations from their theoretical gel position (Figure 5 in Tebbe et al. 2005). A theoretical 2D gel for *N. pharaonis* (Figure 2.4) was predicted from the complete set of theoretical proteins which were translated from the EV gene set. It shows a typical spot pattern as found for *Halobacterium* with the majority of proteins distributed over the acidic region of the gel (pI 3-6). Halophilic membrane proteins tend to be more alkaliphilic than soluble proteins, though.

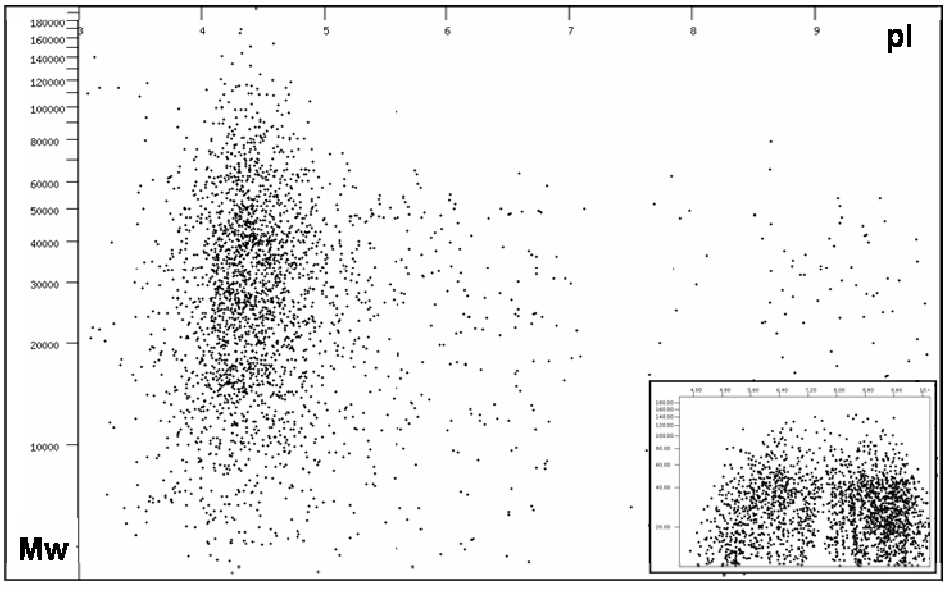


Figure 2.4: **Theoretical two-dimensional gel of *N. pharaonis*** for a pH range of 3 to 10. Nearly all proteins have an acidic isoelectric point (pI) as an adaptation to the high-salt content of the cytoplasm. The two-dimensional gels of other haloarchaeal species show an analogous pattern. Only protein spots for validated genes were included in the given diagram generated by a web-based Halolex tool ('Theoretical 2D gel') (F. Pfeiffer, pers. comm.). The inlay was generated by the JVirGel website and shows a theoretical two-dimensional gel for the archaeon *Pyrobaculum aerophilum* which reveals the typical butterfly-like pattern of a non-halophilic theoretical proteome.

Average pI values and amino acid distributions of proteins were calculated for the theoretical proteomes of haloarchaea as well as for a selection of other archaea and bacteria (Supplemental Table 2.6). Since codon availability for some amino acids strongly depends on the GC content of the genome, organisms need to adapt their proteins differently. While high GC species (>60%) are required to reduce arginine (by 7.5-10.5%) and proline (by 3.0-5.5%) levels in their proteins significantly compared to translated *in silico* DNA, adaptation of species with low GC contents (<43%) is less pronounced (Phe and Leu fractions reduced by 1.0-2.5%) (Supplemental Table 2.6).

For the four halophilic archaea, average isoelectric points and amino acid contents were highly similar although *H. walsbyi* shows a relatively low GC content (47.9%) compared to the high GC haloarchaea (61.1-65.7%). Since differences in GC contents result in different adaptation of the proteomes, arginine levels are lowered by 10.5% in the theoretical proteome of the high GC species *N. pharaonis* but only by 3.5% in *H. walsbyi*. However, independent from the GC content, all halophilic proteomes exhibit high glutamate and aspartate fractions, which add up to 17.5% in *N. pharaonis* (Figure 2.5). While high Glu levels are also found in non-halophilic several species, high concentrations of Asp are characteristic for haloarchaea. Furthermore, low contents of positively charged amino acids (~8.0%) were found in the predicted proteins of halophilic strains. Hence, all 4 haloarchaea

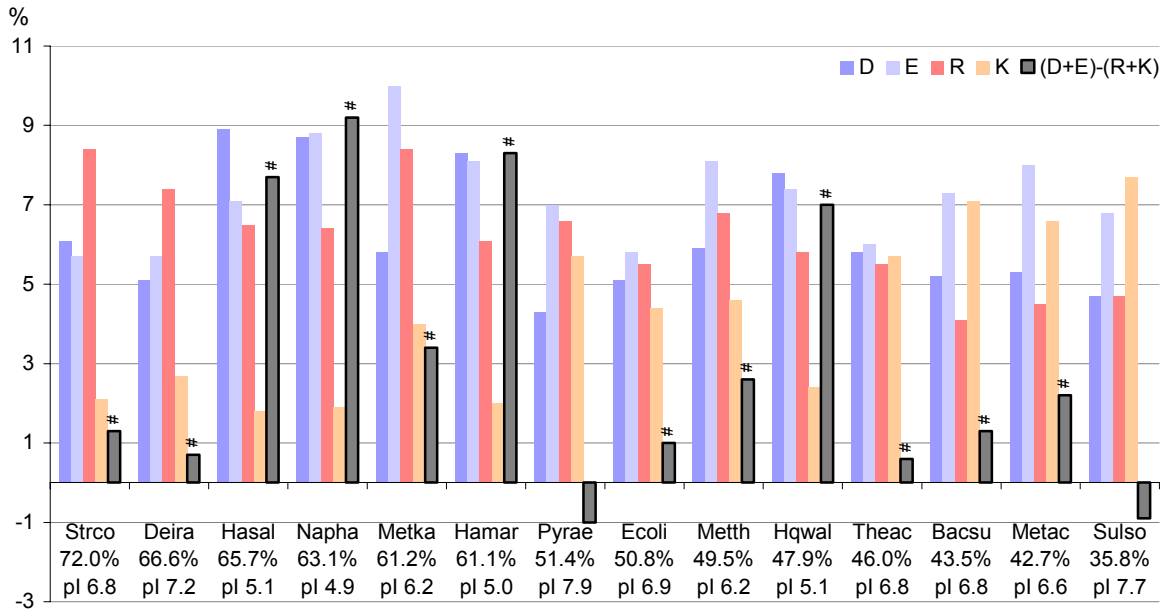


Figure 2.5: **Charged amino acids in the theoretical proteomes** of selected archaea and bacteria sorted by their GC contents in percent (second line). Negatively (blue) and positively (red) charged amino acids are balanced in most organisms (grey columns with hash signs) resulting in neutral average isoelectric points (pI). Halophilic proteomes (Hasal, Napha, Hamar, Hqwal) reveal large fractions of acidic amino acids and relatively low levels of positively-charged amino acids. Thus, haloarchaea exhibit acidic proteins which do not aggregate in their high-salt cytoplasm.

exhibit acidic proteins with average pI values around 5 that permit proteins to remain soluble in the high-salt cytoplasm. When analysing amino acid distributions and charge densities of protein surfaces for a subset of *H. salinarum* proteins that were modelled onto known protein structures from PDB, negative charges of acidic amino acids were found to be clustered at the halobacterial protein surfaces in comparison to the protein surfaces of non-halophilic species (L. O. Essen, unpublished data). On the other hand, positive charges from lysine, arginine, and histidine residues were reduced at halophilic protein surfaces. The average isoelectric point of the halophilic protein surfaces at pI 4.5 was considerably lower than the pI of the complete protein sequences, since acidic amino acids seem to be solely clustered at the protein surface while the inner region of haloarchaeal proteins is likely of mesophilic character.

In contrast, proteins of non-halophilic microorganisms reveal neutral pI levels with evenly distributed negatively charged (11.0-12.5% Glu and Asp) and positively charged amino acids (10.0-11.0% Arg and Lys). Theoretical proteomes of the crenarchaeota *S. solfataricus* and *P. aerophilum* show slightly alkaliphilic proteomes with isoelectric points around 8, since they possess higher levels of positively charged amino acids (~12%). Methanogens, which are close relatives of haloarchaea, have slightly acidic proteomes. Their proteins are characterized by high aspartate levels of up to 10%, but also high frequencies of positively charged amino acids (up to 12.4%).

2.3.2 Development of a pI scanning tool

The selection of start codons for halophilic proteins is supported by calculating theoretical isoelectric points and amino acid distributions for each of the alternative starts. Therefore, a pI scanning tool was developed which represents local isoelectric points of a given protein sequence. Local pI values are calculated from theoretical peptides in a window sliding technique, and are then represented versus sequence positions (Figure 2.6). In case the local isoelectric point profile shows a shift in pI value below a certain threshold, the closest start codon will be chosen as the most likely start. The pI scanner was applied as a post-processing tool for the assessment of gene and start codon predictions of halophilic proteins, especially in case various gene finders predicted different start codons.

Membrane and secretion proteins possess few acidic amino acids within their hydrophobic stretches, and positively charged amino acids are present in signal sequences (see Chapter 3). Thus, these proteins often have higher overall isoelectric points. However, the parts of the proteins, which are exposed to high salt concentrations, e.g. the ones following the signal sequence and loops in between transmembrane domains, have pI values and amino acid distributions that are typical for halophilic proteins.

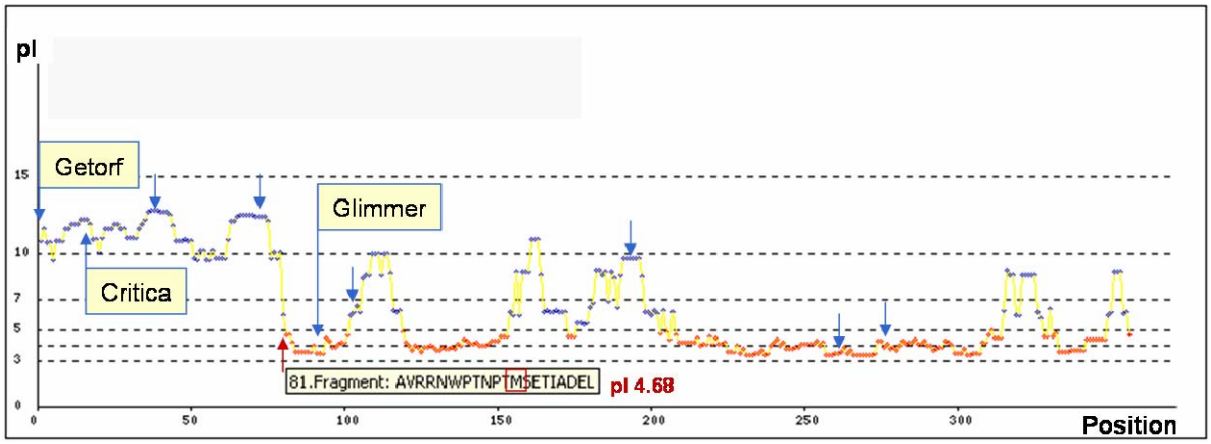


Figure 2.6: **Start codon selection for *proB* by using the isoelectric point profile.** The local pI for a peptide window of 20 aa shifts below the typical value for halophilic proteins (pI <5) at position 81 (red arrow). Alternative ATG or GTG start codon positions (blue arrows) and starts proposed by various gene finders (arrows with flag) are indicated in the diagram. Of the potential starts, the start codon at position 92 that directly follows the pI shift is selected. It coincides with the Glimmer prediction. This selection is supported by the gene context of *proB* (see Figure 2.1).

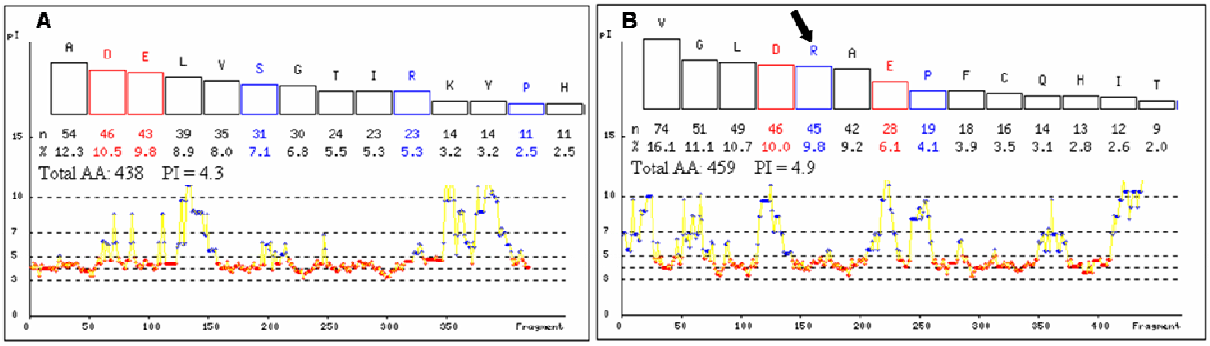


Figure 2.7: **Amino acid distribution plot and pI profile of ORFs in opposite frames.** The use of characteristic protein features for gene and start codon selection is limited in case of diverging ORFs in opposite open reading frames. Both, *proA* predicted for frame +1 **(A)** and NP3977E on frame -1 **(B)** (see Figure 2.1) have typical halophilic pI values below 5, high levels of Glu and Asp, and similar pI profiles. However, *proA* was assigned as the correct gene due to its similarity to proline synthesis genes. NP3977E shows a relatively high Arg content (arrow) compared to the average halophilic protein, and was annotated as spurious ORF. For colour codes see also legends of Figures 1.2 and 1.6.

A check of acidic amino acids and pI values does not resolve the question of correct genes and gene starts in all cases. If alternative genes with diverging gene context are predicted for opposite reading frames, both ORFs may show high levels of acidic amino acids as well as similar pI values and pI scans (Figure 2.7). Because of these similar intrinsic protein features, further criteria such as similarity search results are required to distinguish between diverging ORFs from opposite frames.

2.3.3 Gene distance analysis of haloarchaeal genomes

Prokaryotic genes often form transcription units so that these genes are transcribed on the same mRNA and are under control of the same promoter located in front of the first gene of the transcription unit. Genes within transcription units are found in serial gene context, and are separated by only few nucleotides or even overlap each other. In order to predict transcription units in *N. pharaonis* and two further halophilic genomes, a distance statistics for serial genes was compiled over the complete EV gene set (Figure 2.8). The distance statistics for all three halophiles reveal a maximum for a gene distance of -4 bp, and peak further for small overlaps by four and one nucleotides, respectively. Thus, most halophilic transcription units are characterized by typical gene overlaps (gene distances of -8, -4, and -1 bp), and gene distances can be applied as one criteria for start codon selection in haloarchaeal genomes. The observed gene overlaps arise due overlapping stop (TAA, TGA, TAG) and start codons (ATG, GTG) of genes in different reading frames, e.g. A*TG*A for a gene distance of -1 bp.

Distances of 10 to 35 nucleotides between serial genes are rare, and distance frequencies only increase slightly above 35 bp without another sharp peak. Independent transcription of the latter gene with its own promoter becomes more likely with increasing distance, since the promoter region, which is required to build up the transcription machinery, would fit in between the two serial genes. Here, a gene distance below 35 bp was assumed to indicate

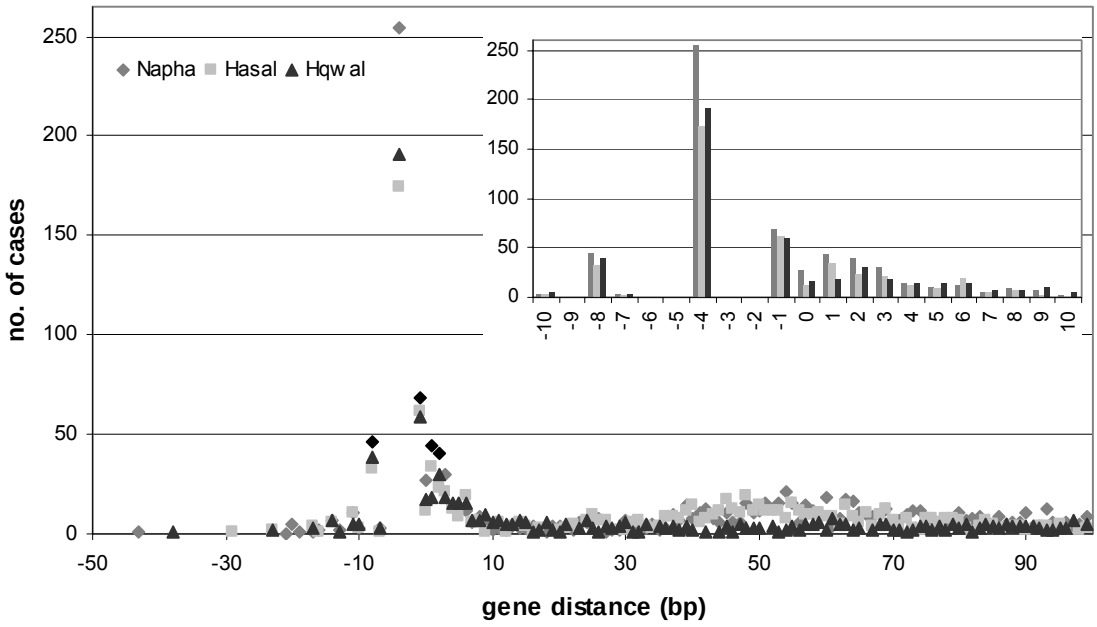


Figure 2.8: **Gene distance statistics for serial genes in three haloarchaeal genomes.** Overlapping genes were frequently observed, with typical gene distances of -1, -4, and -8 bp as shown in the histogram inlay. These are due to overlapping stop and start codons of genes in different frames, e.g. *TG*ATG for a distance of -1 bp and A*TG*A for a distance of -4 bp. Genes separated by less than 35 bp especially the ones with small overlaps are predicted to be as co-transcribed in transcription units.

transcription as a transcription unit. With this cutoff, 43% and 46% of the *N. pharaonis* and *H. salinarum* genes are predicted to be co-transcribed, while only one third of the *H. walsbyi* genes are probably found within transcription units.

The assumed minimal distance of 35 bp for genes that probably exhibit an independent promoter agrees with a recent study for which transcriptional and translational signals were predicted in archaeal genomes (Torarinsson et al. 2005). In this study, *Halobacterium* genes showed an AT peak at position -11 to -10 and a box A promoter motif at position -28, which is preceded by a predicted BRE signal for interaction with the archaeal transcription factor TFB located three nucleotides further upstream. No further bases between transcriptional start site and start codon have to be considered, since single genes and the first genes of transcription units, e.g. the *fdx* and *gvpO* genes of *H. salinarum*, often produce leaderless transcripts, which lack an 5' untranslated region and Shine-Dalgarno sequence (Sartorius-Neef and Pfeifer 2004). Thus, gene distances of at least 31 bp are required to befit a promoter region in *H. salinarum*. In most archaea, Shine-Dalgarno sequences (differing from the bacterial consensus) are frequently found for genes within transcription units. The prediction of Shine-Dalgarno sequences and conserved promoter regions in haloarchaeal genomes would validate proposed gene start codons in the future.

2.4 Validation of gene starts using proteomics data

2.4.1 Definition of a proteomics-verified gene and start codon set

Proteomics permits the large-scale identification of proteins by mass spectrometry. Purified proteins are digested and identified by their pattern of fragment masses (peptide mass fingerprint), which represents the peptides that result from digestion (MS). By another technique, proteins are digested and individual peptides of the protein are identified by fragmentation and subsequent mass-spectrometric identification of the fragmentation pattern representing the peptide sequence (MS/MS). While both techniques permit reliable identification of proteins, identification of individual peptides is much more reliable with MS/MS than MS. Therefore, when in this study it is referred to peptide identification by MS, the specific mass of the peptide must have occurred to a certain extent in the given set of MS spectra of the respective protein (see Methods).

Proteins of *N. pharaonis* and *H. salinarum* have already been identified by MS and MS/MS techniques (Tebbe et al. 2005; Klein et al. 2005; F. Siedler, pers. comm.). Thus, 41.2% (1170) of the EV gene set in *N. pharaonis* and 41.7% (1176) of the predicted gene set of *H. salinarum* are experimentally verified (Table 2.1). Here, the available proteomics data

Table 2.1: **Identification of proteins and N-terminal peptides by proteomics.** Circa 40% of the predicted genes and 6% of the start codons for *N. pharaonis* and *H. salinarum* were validated by analysing data from different mass spectrometry techniques (MS, MS/MS). Identified N-terminal peptides can be distinguished into unprocessed N-termini (N1-type) and processed N-termini (N2-type) for which the initial methionine residue was cleaved off. Percentages of the different peptide types in respect to the total number of identified N-terminal peptides are given in parentheses.

	Identified proteins	**N-terminal peptides**	**N1-type peptides**	**N2-type peptides**	**N1-/N2-type peptides**
N. pharaonis					
MS/MS	**1170**	**210**	72 (34.3%)	133 (63.3%)	5 (2.4%)
H. salinarum					
MS and MS/MS	**1176**	**198**	55 (27.8%)	140 (70.7%)	3 (1.5%)
MS/MS	**653**	**80**	11 (13.8%)	69 (86.2%)	0 (0.0%)
MS	**830**	**128**	46 (35.9%)	80 (62.5%)	2 (1.6%)

were further analysed in order to find identified N-terminal peptides that validate predicted gene starts. As a result, start codon assignments for 7.4% and 5.9% of the EV gene set of *N. pharaonis* and *H. salinarum* were verified by proteomics (Table 2.1). Thus, for the *N. pharaonis* chromosome, a set of 1145 experimentally verified genes and 206 verified gene starts to which will be later referred to as proteomics-verified (PV) gene and start set, was established.

2.4.2 Analysis of post-translational modifications in N-terminal peptides

MS spectra can be straightforwardly searched for post-translational modifications such as methionine cleavage and acetylation of N-terminal peptides, since the complete set of mass peaks for a single protein is available after analysis with MASCOT. Furthermore it can be easily checked whether the modification only occurs partially for a protein. However, analysis of MS data suffers from its rather high false positives rate when referring to an individual peptide. In contrast, data from MS/MS samples reliably identify peptides not only by correct peptide masses but also by the peptide sequence, but the MASCOT search has to be rerun with a new parameter set that considers post-translational modifications. In this analysis, MS data available for *H. salinarum* (Table 2.1) were analysed for various post-translational modifications.

Many prokaryotic and eukaryotic proteins are post-translationally modified through cleavage of the N-terminal methionine residue (ini-Met) by methionine aminopeptidase (Map). In the two halophilic archaea, around 60% of the identified N-terminal peptides were also found to exhibit a removed ini-Met. The observed post-translational modification of halophilic proteins was likely catalyzed by a homolog of Map (NP3190A, OE3623R). For each of the identified proteins, either an unprocessed (N1-type) or a processed (N2-type) N-terminal peptide was detected except for a few cases (5 cases in *N. pharaonis* and 3 cases in *H. salinarum*),

Table 2.2: **Amino acid statistics of processed *H. salinarum* N-terminal peptides** as identified by MS. The N-termini were analysed for two types of post-translational modifications, cleavage of the initial methionine residue (N2) and acetylation of the amino group (ac). A statistics of the amino acids that occur at the second position following the ini-Met was compiled (only amino acid with more than three cases listed, for the complete list see Supplemental Table 2.7). Amino acids with small radii of gyration are found at the start of N2-type N-termini, and larger residues follow the ini-Met of N1-type N-termini (NMet column). Only N-terminal peptides starting with serine or alanine residues were found to be acetylated (NAc column).

AA	ini-Met type	NAc	total	N1free	N1ac	N2free	N2ac
all AA			131	35.9%	0.8%	55.7%	6.1%
T	**N2**		30	3.3%	0%	96.7%	0%
S	**N2**	**X**	27	0%	0%	77.8%	22.2%
A	**N2**	**X**	17	17.6%	0%	64.7%	17.6%
D	**N1**		13	100%	0%	0%	0%
P	**N2**		11	0%	9.1%[a]	90.9%	0%
E	**N1**		6	83.3%	0%	16.7%[b]	0%
L	**N1**		6	100%	0%	0%	0%
Q	**N1**		6	100%	0%	0%	0%
I	**N1**		4	100%	0%	0%	0%

a, b - cases which do not fit into the observed pattern for N-terminal methionine cleavage and acetylation
(a - correct identification, b - false positive identification)

where both types of N-termini were observed in different samples. The type of amino acid following the ini-Met is strongly biased for N1-type and N2-type N-terminal peptides (Supplemental Table 2.7, Table 2.2). Ser, Thr, Ala and Pro are commonly found in cleaved peptides whereas Asx, Glx, Leu, and Ile are present in unprocessed N-termini. This finding is consistent with the described specifities of other methionine aminopeptidases, which were shown to act only on the seven amino acids with the smallest radii of gyration (Gly, Ala, Ser, Cys, Thr, Pro, Val). This substrate specifity can be explained by steric hindrance, as deduced from the crystal structure of Map (Bradshaw et al. 1998; Polevoda and Sherman 2003).

In eukaryotes, the set of processed/non-processed N-terminal residues coincides with the set of and stabilizing/destabilizing residues of the ubiquitin-dependent N-end rule pathway of protein turnover, where larger amino acids are recognized by ubiquitin ligase and the proteins subsequently degraded. Therefore it was suggested that the destabilizing ini-Met acts as a prophylactic cap that prevents premature degradation of proteins where a large residue follows the ini-Met (Bradshaw et al. 1998). However, the ubiquitin-dependent N-end rule pathway does not exist in archaea.

Data resulting from MS experiments for *H. salinarum* were further analysed for N-terminal acetylations of the α-amino group of the initial residue (Table 2.2). This type of post-translational modification is catalyzed by N-terminal acetyltransferases (Nat), which transfer an acetyl group from acetyl-CoA to the N-terminal residue. Around 7% of the N-terminal

peptides identified by MS were found to be acetylated at the α-amino group of the initial residue in *H. salinarum*. Again, for each of the identified proteins, the detected N-terminal peptide masses always either correlated with the theoretical mass of the free N-terminus or with the theoretical mass of the acetylated N-terminus. Spectra with the mass peak of the acetylated N-terminus as well as spectra exhibiting the mass of the non-acetylated N-terminus were only observed for OE3547F, hence, this protein might be only partially acetylated in *H. salinarum*. In contrast to eukaryotes, the ini-Met seems to be always cleaved off the N-terminus prior to the acetylation step except for one case (OE4339R). Only N2-type N-termini starting with a serine or alanine residue were found to be acetylated (Table 2.2). Thus, the specifity of the *H. salinarum* N-terminal acetyltransferase resembles the one of NatA in *Saccharomyces cerevisiae* (N2-type N-termini that start with Ser, Ala, Gly, or Thr) (Polevoda and Sherman 2003).

In contrast to the ini-Met cleavage, the extent of N-terminal acetylation differs greatly between prokaryotes and eukaryotes. While 80-90% of cytoplasmic proteins in mammals and 50% of yeast proteins are acetylated, acetylation in bacteria and archaea is rare. Three out of 810 proteins with verified N-terminal sequences are acetylated in *E. coli* (EcoGene database) and only three N-terminally acetylated (DHE2_SULSO, RS7/RL31_HALMA) but 97 non-acetylated proteins were described in archaea (Swiss-Prot database) (Polevoda and Sherman 2003). The presented data of acetylated and non-acetylated N-termini in *H. salinarum* is the first systematic large-scale study of N-acetylation in prokaryotes, and indicates that N-acetylation is indeed not as frequent in archaea but more common than previously estimated. Whereas many yeast and mammalian proteins are only partially acetylated, *H. salinarum* proteins seem to occur either in the acetylated or non-acetylated form. While *S. cerevisiae* employs three types of N-acetyltransferases (NatA, NatB, NatC) with different substrate specifity (Polevoda and Sherman 2003), the *H. salinarum* Nat seems to act similarly to NatA. It was suggested previously that cleavage of NatB-like substrates, which are characterized by an N1-type N-terminus that is followed by an acidic residue (Met-Asp or Met-Glu), are not recognized in prokaryotes (except for DHE2_SULSO). This is in contrast to yeast and mammals, which modify these types of substrates without exception. From the fact that NatB activity is missing in prokaryotes, it was concluded by Polevoda & Sherman that prokaryotic N-terminal acetylation is fundamentally different in prokaryotes. However, the finding that NatA-like acetylation is present rather suggests that N-acetylation in the different domains of life is similar but less complex in prokaryotes. Prokaryotes seem to lack several specific Nat isoforms featured by eukaryotes, but likely possess only a basal type of N-acetyltransferase.

The biological function of N-acetylation is not completely understood yet, and the role of N-terminal acetylation in the protection of eukaryotic proteins from degradation by the N-end rule pathway is arguable (Polevoda and Sherman 2003). It was suggested that acetylation might have only subtle effects upon most proteins, and that only few proteins require N-terminal acetylation for activity (e.g. actin in *Drosophila*) or stability (e.g. the glucose dehydrogenase in *Neurospora crassa*) (Perrier et al. 2005; Polevoda and Sherman 2003). This view is supported by deletion mutants of *S. cerevisiae* N-acetyltransferase genes, which were still viable but show defective phenotypes such as slow growth and reduced mating in the case of mutants for NatB (Polevoda and Sherman 2003).

Archaeal translation initiation resembles the one of eukaryotes, which is for example expressed by the usage of methionine as the initial amino acid of proteins. In contrast, bacteria and eukaryotic cell organelles use N-formyl-methionine for the translation initiation process, and, thus, newly synthesized bacterial proteins require to be deformylated prior to other post-translational modifications such as N-terminal methionine cleavage and acetylation. By analysing the MS data set, the usage of methionine for translation initiation in haloarchaea could be confirmed since no mass peak of any formylated N-terminus could be identified.

2.5 Performance of microbial gene finders for GC-rich genomes

2.5.1 Gene prediction in genomes with different GC contents

Gene prediction and start codon selection is especially difficult for GC-rich genomes. While circa one open reading frame is found at each position within microbial genomes with low GC contents, genome sequences with GC contents over 60% show multiple ORFs at each position (Figure 2.9). For example, the GC-rich chromosome of *Streptomyces coelicolor* (71.1% GC) encodes for 3.12 potential genes at each site, and the *N. pharaonis* chromosome (63.4% GC) shows a coding multiplicity of 2.19. Since microbial genes only overlap by some nucleotides within transcription units (Chapter 2.3.3), one gene has to be selected at each site of the genome to avoid gene overprediction. This could also be shown for *H. salinarum* and *N. pharaonis*, for which circa 40% of all proteins have been identified (Chapter 2.4) but usage of overlapping alternate frames has not been found in a single case although spurious ORFs were included in the MASCOT search database.

Multiplicity of coding in species with high GC contents results from a scarcity of stop codons, TAA, TGA, and TAG, in the genome. The under-representation of TA sequences due to a distinctive bias at the dinucleotide level adds further to the scarcity of stop codons in *Halo-*

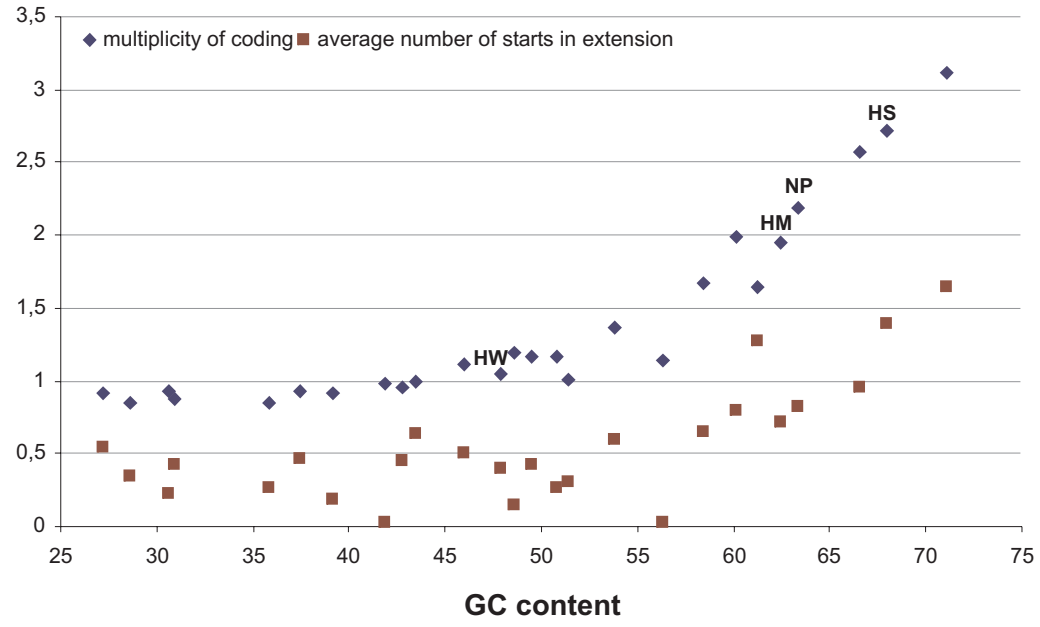


Figure 2.9: **Analysis of 26 genomes with different GC contents**: multiplicity of coding (blue) and average number of start codons in potential protein extensions (red). Due to the scarcity of stop codons (Supplemental Figure 2.13), GC-rich species potentially encode more than one gene per genome position and genes can potentially be extended due to the occurrence of multiple alternative start codons. Thus, gene overprediction and extensive start codon misassignments might occur.

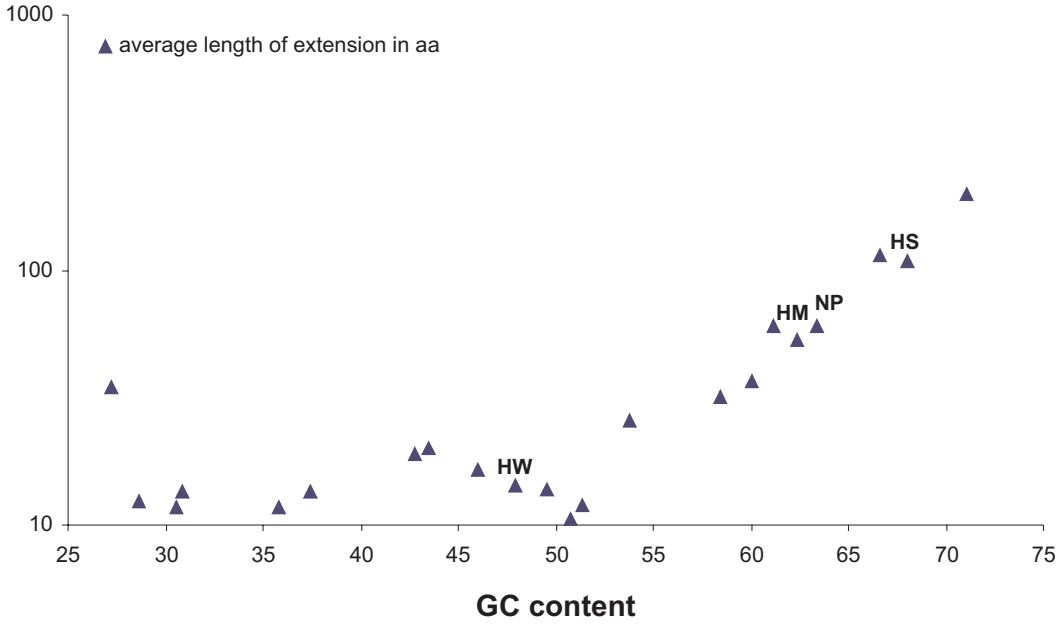


Figure 2.10: **Average length of potential gene extensions** for selected genomes with different GC content. Protein lengths in GC-low genomes only differ by up to 20 aa for alternative starts, whereas a reassignment of start codons in GC-rich genomes would lead to changes in protein lengths of up to 200 aa in average. In case maximal protein lengths are systematically selected in a GC-rich genome, high fractions of overlapping genes occur. For four genomes, average lengths were below 10 aa (not shown).

bacterium. While the *Sulfolobus solfataricus* genome (35.8% GC) shows 141 stops within 1000 nucleotides, only 23.5 stop codons per 1000 bases are found in the *S. coelicolor* genome (Supplemental Figure 2.13). Thus, it is more likely in GC-rich sequences that spurious ORFs with considerable length (more than 100 codons) arise between stop codons. When comparing the set of ORFs with the set of published genes to classify them into real genes and spurious ORFs, spurious ORFs of *S. coelicolor* showed an average length of 261 aa and even an ORF with 2800 aa length was excluded.

The low frequency of stop codons leads not only to a high number of ORFs which have to be distinguished into real genes and spurious ORFs but also in high numbers of potential gene starts to chose from. For the EV gene set of *N. pharaonis* (Chapter 2.2) and complete gene sets of selected other genomes, the average number of alternative start codons upstream of the assigned gene start to the previous in-frame stop codon was determined. As shown in Figure 2.9, the number of potential gene extensions rises significantly with increasing GC contents, starting from 0.27 (*S. solfataricus*) additional starts up to 1.64 (*S. coelicolor*) start codons per gene. Due to this observed multiplicity of starts, gene prediction in GC-rich species is prone to start codon misassignments.

It might be argued that one or two alternative starts per gene might frequently result in mispredicted start codons but these will not significantly affect the overall quality of gene prediction. This holds true for GC-low genomes, where choosing between alternative start codons results in protein variants with minor length differences between 12 aa (*S. solfataricus*) and 20 aa (*Bacillus subtilis*, 43.5% GC) (Figure 2.10). However, the multiplicity of start codons leads to a dramatic increase in length of potential gene extensions for GC-rich genomes, and translated genes of *S. coelicolor* can be extended in average by around 200 aa. Thus, theoretical proteins translated from GC-rich genomes apparently contain an additional N-terminal domain. Consequences of start codon misassignments upon overall gene prediction are discussed in more detail within the following subchapter.

For 3 of the 17 assessed genomes, relatively low average lengths of potential gene extensions were calculated (*Aeropyrum pernix*: 0.7 aa, *Pyrococcus horikoshii*: 0.8 aa, *Archaeoglobus fulgidus*: 4.8 aa, *Chlamydophila caviae*: 6.3 aa), indicating a systematic error in start codon assignments by favoring maximal gene lengths. However, since these species revealed GC contents below 60%, gene prediction quality is likely not strongly influenced by probable start codon misassignments.

2.5.2 Gene finder assessment using the validated *Natronomonas pharaonis* gene set

Three microbial gene finders, Reganor, Glimmer, and Critica, which have integrated into the GenDB annotation platform, were assessed in respect to their gene prediction performance for GC-rich genomes. For this purpose, predicted genes and gene starts for the *N. pharaonis* chromosome (GC 63.4%) were compared with the established validated gene set. Comparison against the EV gene set (Table 2.3) shows that all gene finders avoid extensive gene overprediction in contrast to six-frame translation (Getorf) (Figure 2.11). However, primary gene finders tend to over- (Glimmer) or underpredict (Critica) genes to some extent. Reganor combines the strength of Critica and Glimmer so that it has the best overall performance. Thus, Reganor is indeed qualified for gene prediction in GC-rich genomes as stated previously (McHardy et al. 2004). Further, the PV subset of genes was used to

Table 2.3: **Assessment of gene prediction results for several gene finders.** Analysis was done against the validated expert-validated (EV) and proteomics-verified (PV) gene sets of the *N. pharaonis* chromosome. All gene finders perform well in comparison to six-frame translation (Getorf) but overall results for Reganor are best. FP - false positives, FN - false negatives, NA - not applicable.

	Validated	**FN**	**%**	**Predicted**	**FP**	**%**
EV gene set	**2675**					
Getorf	2452	223	**8.3**	11062	8610	**351.1**
Glimmer	2612	63	**2.6**	2956	344	**13.2**
Critica	2426	249	**9.5**	2426	0	**0**
Reganor	2563	112	**4.6**	2595	32	**1.2**
PV gene subset	**1145**					
Getorf	1098	47	**4.1**	NA	NA	NA
Glimmer	1134	11	**1.0**	NA	NA	NA
Critica	1093	52	**4.6**	NA	NA	NA
Reganor	1113	32	**2.9**	NA	NA	NA

Table 2.4: **Assessment of start codon assignments for several gene finders.** Analysis was done against the validated expert-validated (EV) and proteomics-verified (PV) start sets of the *N. pharaonis* chromosome. For all gene finders significant start codon misassignments were observed with the tendency to predict genes that are too long. Especially start shifts of more than 50 aa result in overlapping genes and predicted proteins with additional N-terminal domains.

	Total	**Correct**	**%**	**Too long**	**%**	**Too short**	**%**	**Shift > 50aa**	**%**
EV start set	**2675**								
Glimmer	2612	1625	**62.2**	818	**31.3**	169	**6.5**	103	**3.9**
Critica	2426	2084	**85.9**	321	**13.2**	21	**0.9**	107	**4.4**
Reganor	2563	2179	**85.0**	357	**13.9**	27	**1.1**	109	**4.3**
PV start set	**206**								
Glimmer	205	139	**67.8**	53	**25.9**	13	**6.3**	1	**0.5**
Critica	200	175	**87.5**	25	**12.5**	0	**0**	0	**0**
Reganor	205	179	**87.3**	26	**12.7**	0	**0**	0	**0**

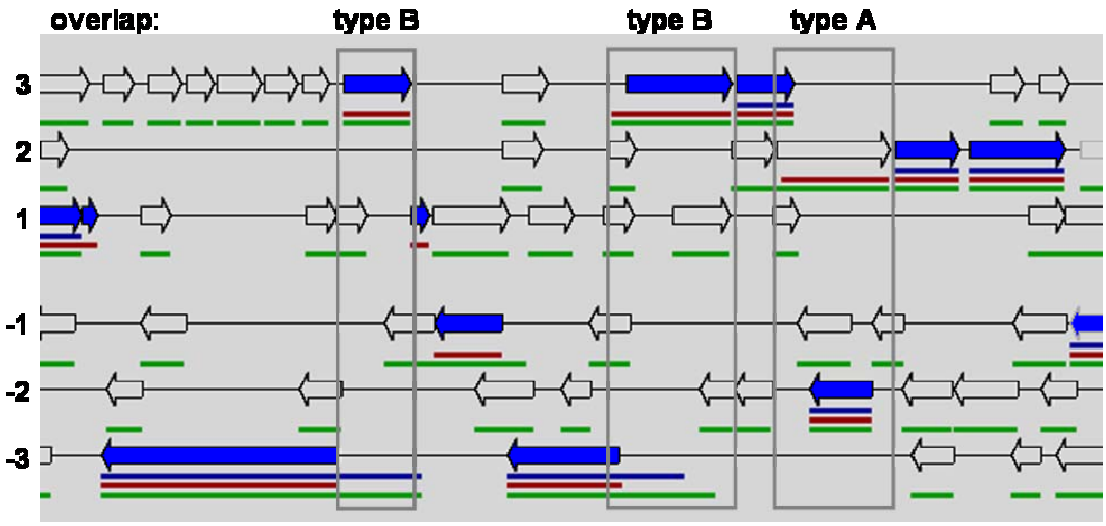


Figure 2.11: **Representation of gene prediction results for the *N. pharaonis* chromosome** (position 779,000 to 791,000) by the GenDB annotation platform. Predicted ORFs were classified into validated genes (blue arrows) and spurious ORFs (open arrows) by expert validation. At each genome position, several six-frame ORFs (Getorf, green lines) are present (multiplicity of coding) but all gene finders, Glimmer (red line), Critica (blue line), Reganor (not shown), avoid extensive gene overprediction. Cases of overlapping ORFs (grey boxes), which are often located in opposite frames, can either occur as a result of gene overprediction (type A) but also due to misassigned start codons (type B). While type B overlaps are resolved by shifting starts codons of one or both overlapping genes, type A overlaps require excluding one of the overlapping genes. Reganor and Critica reveal high frequencies of false negatives, since all overlap cases are resolved as type A overlaps (here: frame +3 genes in both type B boxes missed). Type B overlaps are common in GC-rich genomes due to longer potential gene extensions (Figure 2.10).

determine fractions of false negative genes. These were 1.5% to 5.0% below the rates for the EV gene set indicating that the EV gene set still contains some spurious ORFs in spite of the rigorous post-processing of gene prediction results.

Expert-validated gene starts as well as proteomics-verified N-termini were used to assess gene finder performance in respect to start codon assignments (Table 2.4). All three gene predictors tended to predict genes which were too long. For Critica and Reganor, the fraction of lengthy genes was around 13% percent but only about 1% of the genes were too short. Glimmer even predicted one third of the chromosomal *N. pharaonis* gene starts incorrectly.

Initially, the *N. pharaonis* gene set predicted by Reganor has been used for post-processing. Since Reganor chooses Critica over Glimmer starts when combining gene prediction results it might be argued that expert validation was biased in favour of Reganor/Critica starts over Glimmer start predictions. However, the PV start set which is even more reliable than the EV gene set also discriminates Glimmer as a gene prediction tool with a high frequency of misassigned starts.

Start codon missassignments do not necessarily affect overall gene prediction significantly in case of small start shifts. However, four percent of predicted gene starts for all tools were repositioned by more than 50 aa, mostly by shortening of the genes. This systematic

preference of longer gene versions also leads to misinterpretations in subsequent domain prediction. Signal sequences and LAGC motifs (Chapter 4), for example, will not be detected for translated genes with misassigned start codons, e.g. in one of the halophilic halocyanins (*H. marismortui* rrnAC1377). Another example is an additional large N-terminal cytoplasmic domain discussed for TatC1 (VNG2269G) of *H. salinarum* strain NRC-1 (Bolhuis 2002), which is probably solely the result of a misassigned start codon.

The start codon selection process also affects the overall gene selection results in GC-rich genomes. Due to potential gene extensions of considerable length (Figure 2.10), overlap cases of neighboring genes are likely to occur frequently, and have to be solved by expert-validation through shortening one or both overlapping genes. Overlapping genes are not solely the result of start codon misassignments (overlap type B) but can also arise from gene overprediction (overlap type A) (Figure 2.11). Thus, for each of the overlap cases the decision has to be made, whether genes should be shortened (type B) or whether one of the genes should be excluded (type A). This overlap ambiguity (misassigned gene or misassigned start codon) leaves room for improvement in the handling of overlaps by Reganor and Critica in GC-rich genomes. These tools select only one gene per genome position and generally assume type A overlaps. Since there is always one of the overlapping genes excluded even in case the overlap can be resolved by gene shortening, high rates of false negative genes are observed for Reganor and Critica (Table 2.3). Due to the integration of Critica and Glimmer data, the false negative rate for Reganor is not as high. Glimmer generally permits overlaps and Reganor excludes only Glimmer genes with extensive overlaps which are commonly true type A overlaps.

2.6 Conclusions

The quality of gene prediction by microbial gene finders strongly depends on the GC content of the genome. For genomes with high GC contents above 60% such as the halophilic archaea *N. pharaonis* and *H. salinarum*, there are not only high numbers of open reading frames per genome position potentially resulting in high numbers false positive genes, but also an increased number of alternative start codons per ORF. The microbial gene finders, Critica and Reganor, minimalize the number of false positive genes, and Reganor revealed best overall gene prediction due to lower false negative numbers than Critica. However, none of the three tested microbial gene finders, Glimmer, Critica, and Reganor, showed acceptable error rates with respect to start codon selection since genes were commonly too long. Due to the scarcity of stop codons, these gene extensions are furthermore of significant

length in GC-rich genomes so that signal sequence and domain predictions are affected. The lengthy gene extensions also frequently lead to overlapping genes so that two types of gene overlaps have to be considered, the ones caused by start codon misassignment and the ones resulting from gene overprediction. So far, gene prediction tools handle overlaps solely by choosing one of the overlapping genes and not through shortening of genes. Since the interconnection of gene and start codon selection processes for GC-rich genomes is not considered by these tools, high rates of false negatives occur that could be avoided by start codon optimization.

Future gene prediction approaches for GC-rich genomes should include procedures for start codon selection, e.g. by automatically analyzing N-terminal blast alignments of translated genes from related species. This approach already proved to be useful for choosing between alternative gene starts in *N. pharaonis*, where around one third of the initial gene starts agreed with *H. salinarum* starts. As further criteria for start codon selection, e.g. in case of alignment discrepancies, the distributions of certain GC-dependent amino acids such as arginine and proline, which differ greatly between genes and *in silico*-translated DNA in GC-rich genomes, might be assessed. Gene starts in halophiles can be optimized by checking for the switch of local isoelectric points to typical acidic values around the correct start position (cell sorting signals have to be taken into account) as well as by considering for typical transcription overlaps. Promoter analysis and prediction of ribosomal binding sites might also be performed in case reliable test data is already available for the genome.

Proteomics data can verify predicted genes and their gene starts. For *N. pharaonis* and *H. salinarum*, already 40% of the predicted proteins and 6% of the predicted N-termini were experimental identified by MS or MS/MS techniques so that promoter and signal sequence predictions are more reliable (Chapter 4). Furthermore, post-translational modifications such as N-terminal methionine cleavage and N-acetylation were analyzed establishing a first large-scale picture of these post-translational mechanisms in archaea. By the analysis of the complete MS/MS data set and the application of further procedures for the identification of N-terminal peptides, numbers of identified N-termini will likely increase in future.

2.7 Methods

2.7.1 Post-processing of gene prediction results

For gene prediction of the *N. pharaonis* genome, Reganor (McHardy et al. 2004) from the annotation package GenDB (Meyer et al. 2003) was used, which integrates results from Critica (Badger and Olsen 1999) and Glimmer (Delcher et al. 1999). In addition, sixframe translation (> 100 codons) was performed. From the resulting raw set of 11874 distinct ORFs, a set of 2843 validated genes was selected by the following procedure: Predicted amino acid sequences were bidirectionally compared to the *H. salinarum* strain R1 ORF set (Halolex website) using blast (Altschul et al. 1997) (Figure 2.3 in Chapter 2.2). The ORF set was also analyzed by blast against itself, and against the nr database. Overlapping ORFs were adjusted based on gene context as well as characteristic halophilic pI and amino acid distribution patterns (see next subchapter). In case of discrepant gene starts between *N. pharaonis* and *H. salinarum*, expert decisions were based on available mass spectrometry results (Klein et al. 2005; Tebbe et al. 2005) (Figure 2.12) and again on analysis of further alignments, gene context as well as protein characteristics. At the time of the performed expert validation, 950 proteins and 246 N-terminal peptides had been identified for *H. salinarum* by peptide mass fingerprint (July 2003).

The developed validation procedure was also used to improve gene prediction results of *H. salinarum* str. R1. Although the R1 gene set had before been derived by Orpheus, Glimmer (via strain comparison with the virtually identical strain NRC-1), six-frame translation as well as by expert validation (F. Pfeiffer, pers. comm.), more than 200 had to be corrected as a result of the bidirectional blast analysis with *N. pharaonis*. The developed procedure was also applied to create the validated gene set for the recently sequenced *H. walsbyi* genome (cooperation with H. Bolhuis). Here, interspecies comparison by bidirectional blast was performed against the two gene sets of *N. pharaonis* and *H. salinarum*.

Summary of identified peptides for OE3884F

MsStatus	score	sample	Nterm	
Trusted	193	g839t02a1ai6	N2	MAATTTPVIAAAYR TPQGKDGGVYADTR
Trusted	190	g839t01a1ai6	N2	MAATTTPVIAAAYR TPQGKDGGVYADTR
Trusted	159	g092t01a1ag13	N2	MAATTTPVIAAAYR TPQGKDGGVYADTR
Insecure	076	g092t02a3ag13	N2	MAATTTPVIAAAYR TPQGKDGGVYADTR
Questionable	054	g718t05a1aj11	N2	MAATTTPVIAAAYR TPQGKDGGVYADTR
Trusted	193	SUMMARY	N2	MAATTTPVIAAAYRR TPQGKDGGVYADT

Figure 2.12: **Identification of a *H. salinarum* protein and its N-terminal peptide by mass spectrometry** as given by the Halolex website. Protein samples were derived from two-dimensional gels and analysed by peptide mass fingerprint. Peptides of the protein sequence are colour-coded depending on their identification status (green: identified peptide, red: peptide not identified, grey: peptide not detectable, yellow: trypsin cleavage site). The red M in the N-terminal peptide sequences indicates that only N-terminal peptides without ini-Met were found.

2.7.2 Intrinsic features of haloarchaeal proteins and gene distance analysis

The pI value and amino acid distribution of a protein can be used for gene selection, since these features differ significantly between halophilic proteins (high Glu, Asp levels) which are encoded by functional genes and *in silico* translated DNA (high Arg levels) of spurious ORFs. Amino acid content and pI for each theoretical protein can be derived through the Halolex details page (Figure 2.2 in Chapter 2.1). Since non-translated N-terminal extensions of genes are also characterized by atypical high Arg contents and pI values, amino acid distribution and pI values further support start codon assessment. For this purpose, a scanning tool has been implemented that finds alternative start codons by protein extension or cleavage, and re-calculates the pI value and amino acid distribution for the protein starting from the modified start as well as for the extended/cleaved protein peptide (F. Pfeiffer, pers. comm.). Thus, an optimal start codon can be determined by choosing the start where pI value and amino acid distribution shift significantly from typical halophilic to atypical protein features. Using the validated gene set of halophilic genomes, average pI values and amino acid distributions over the complete predicted proteomes were calculated. For comparison, intrinsic protein features of several non-halophilic microorganisms were calculated using published amino sequences from NCBI.

Since the available Halolex tool supporting start codon selection of halophilic proteins did not graphically represent pI shifts in proteins, a new pI scanning tool was developed which draws local pI values of a given protein sequence which are calculated for a given peptide size by a simple sliding window method. Potential start positions are also displayed in the diagram. Thus, the correct start codon can be determined as the closest start codon following the position with a pI shift below a given threshold (pI <5).

As bacteria, archaeal genes commonly form transcription units as found in bacteria. The region viewer of Halolex marks overlapping genes, and gives a detailed view of the overlapping sequences. Thus, it can be assessed whether a start codon results in a typical overlap by a few nucleotides with the previous gene which indicates a transcription unit. After establishing the complete gene sets of *H. salinarum*, *N. pharaonis*, and *H. walsbyi*, an overall statistics of the gene distances in the haloarchaeal genomes was compiled.

2.7.3 Validation of gene starts using proteomics data by expert validation

For the identification of halophilic proteins, proteomics data resulting from MS and MS/MS techniques were processed as described previously (Klein et al. 2005; Tebbe et al. 2005). The available proteomics data were further analysed for identified N-terminal peptides in order to validate start codons of the genes that encode identified proteins. Available MS spectra were analysed for masses of the N-terminal peptide and modified forms thereof. A N-terminal peptide was defined as identified, in case its peptide mass occurred in at least in

3 spectra and in 25% of the available spectra for the given protein. The N-terminus was surely not identified, if at least 5 spectra exist for the protein with 0-10% identified N-terminal peptides (including modified variants). However, 377 N-terminal peptides of 814 identified proteins can not be detected by MS, since peptide masses are out of the scanned mass range (800-4000 Da). In case of MS/MS data, the N-terminal peptide is considered identified when the MASCOT score for the protein is at least 20 above the 5% significance level and the raw score of the N-terminal peptide is at least 20.

Theoretical masses of N-termini with a removed intitial methionine residue (ini-Met), an acetylated (+42.01056, monoisotopic mass) or formylated (+27.9949, monoisotopic mass) amino group were checked against the peaks of the MS spectra of the identified proteins. Thus, a statistics of identified N-termini including the ones with post-translational modification was derived. Identified N-termini were further analysed for the amino acid following the ini-Met in order to obtain substrate specifities of methionine aminopeptidase and N-terminal acetyltransferase in halophiles.

2.7.4 Performance of microbial gene finders for GC-rich genomes

Genome statistics (e.g. the number of stop codons) were compiled for the newly sequenced halophiles, *H. salinarum* str. R1, *H. walsbyi*, and *N. pharaonis*, as well as for a selection of published genomes with different GC contents (as downloaded from NCBI). Six-frame translation was performed (cutoff 100 aa) allowing ATG and GTG start codons in order to determine the multiplicity of coding on the chromosomes. Since ORFs are uniquely described by their stop position, the set of six-frame ORFs could be compared to the published gene set, and, thus, six-frame ORFs were subsequently classified into real genes and spurious ORFs. Starts of real genes were maximally extended up to the previous stop codon and the number of start codons in the extended region was determined.

Results of the microbial gene finders, Glimmer, Critica, and Reganor, as well as the six-frame translator Getorf were assessed against the expert-validated and proteomics-validated gene and gene start set of *N. pharaonis.* This enabled the assessment of gene finders for the GC-rich example genome by calculation of false positive/negative rates and the fraction of start codon misassignments.

2.8 Supplemental Material

Supplemental Table 2.5: **Results of the bidirectional blast (BB) analysis** between *N. pharaonis* (Napha) and *H. salinarum* (Hasal) genes. N-terminal alignments between symmetrical (sym) and asymmetrical (asym) gene pairs from BB alignments better than the cutoff of e^{-20} were categorized into:

- N1 alignment starts at position 1 in both organisms (identical starts)
- NM alignment starts at position 1 in *H. salinarum*, and with a Met/Val in *N. pharaonis* indicating that the *H. salinarum* gene is too short or the *N. pharaonis* gene is too long (or *vice versa*)
- NC alignment starts close to the N-terminus (up to position 7) with a minimal positional shift of 0-7 aa
- NX alignment does not start close to the N-terminus and/or the minimal positional shift of the alignment is more than 7 aa

Alignment category	Total	%	sym	%	asym Hasal	%	asym Napha	%
total	1978		1395		271		312	
N1	632	32.0	521	37.3	68	25.1	43	13.8
NM	249	12.6	204	14.6	24	8.9	21	6.7
NC	390	19.7	292	20.9	44	16.2	54	17.3
NX	707	35.7	378	27.1	135	49.8	194	62.2

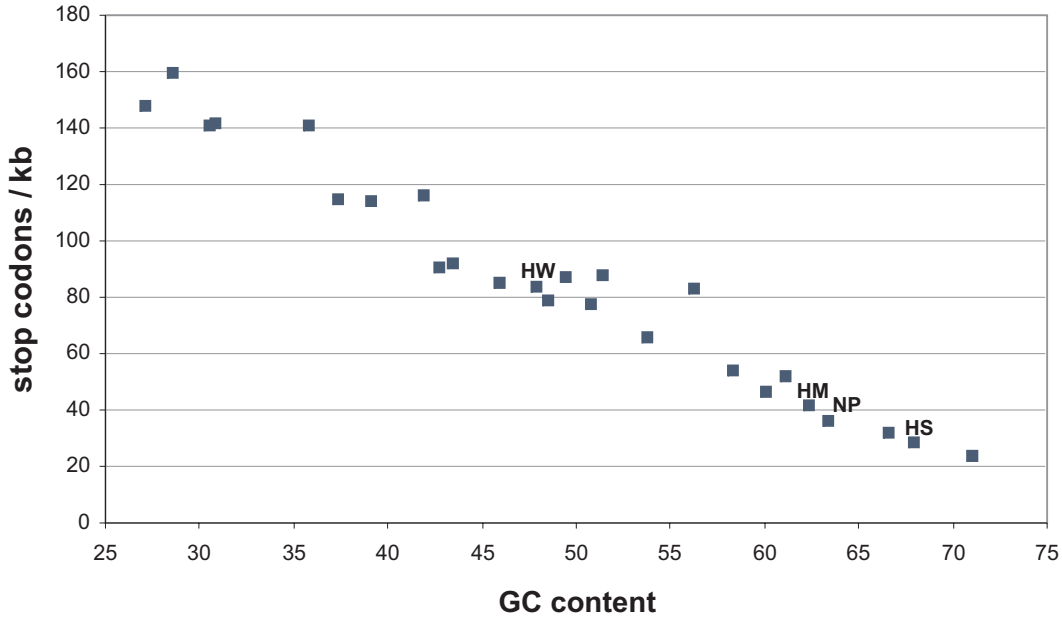
Supplemental Table 2.7: **Statistics of the second amino acid following the methionine** in N-terminal peptides of *N. pharaonis* (Napha) and *H. salinarum* (Hasal) as identified by MS and MS/MS. For each N-terminus, it was analysed whether the initial methionine residue is cleaved off (N2) or not (N1). Thus, the ini-Met type for each of the second amino acids could be proposed (given in parentheses if only few cases are available). In some instances, proteins were only partially processed (N1N2).

Second AA	ini-Met type	All		N1		N2		N1N2	
		Hasal	Napha	Hasal	Napha	Hasal	Napha	Hasal	Napha
A	**N1**	30	31	2	2	**27**	**29**	1[a]	0
C	**?**	0	0	0	0	0	0	0	0
D	**N1**	16	22	**16**	**22**	0	0	0	0
E	**N1**	8	8	**7**	**8**	1	0	0	0
F	**N1**	2	2	**2**	**2**	0	0	0	0
G	**N2**	2	7	0	0	**2**	**7**	0	0
H	**(N1)**	2	0	**2**	0	0	0	0	0
I	**N1**	4	6	**4**	**6**	0	0	0	0
K	**(N1)**	0	2	0	**2**	0	0	0	0
L	**N1**	7	9	**7**	**8**	0	0	0	1
M	**?**	1	1	1	0	0	1	0	0
N	**N1**	1	15	**1**	**15**	0	0	0	0
P	**N2**	16	9	0	0	**15**	**8**	1	1
Q	**N1**	8	6	**8**	**6**	0	0	0	0
R	**(N1)**	1	0	1	0	0	0	0	0
S	**N2**	48	47	0	0	**48**	**45**	0	2
T	**N2**	43	34	0	0	**42**	**34**	1	0
V	**N2**	5	10	0	0	**5**	**9**	0	1
W	**(N1)**	1	0	1	0	0	0	0	0
Y	**N1**	3	1	**3**	**1**	0	0	0	0

a - only the N2-type N-terminus is valid, since the N1-type N-terminus is a false positive MS hit

Supplemental Table 2.6: **Average amino acid distribution and isoelectric points derived from translated gene sets** of a selection of prokaryotes with GC contents ranging from 36% to 72%. Apart from the average percentage of amino acids in the predicted proteins, the amino acid distribution of *in silico* translated DNA of the genomes was additionally calculated, and the difference of the two distributions is given by +/- numbers. Due to differing codon availabilities for varying GC contents, species with high and low GC contents have to adapt their proteins to contain more P, R, and less F, L, respectively (grey). While several species reveal high E contents, high D levels are characteristic for the acidic proteins of haloarchaea (underlined) (Figure 2.5).

Spec	%GC	pI	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Bacsu	43.52	6.81	7.68 (1.37)	0.80 (-2.62)	5.19 (2.03)	7.25 (3.61)	4.50 (-2.00)	6.91 (2.13)	2.28 (-0.80)	7.36 (0.85)	7.05 (0.89)	9.65 (-0.13)	2.78 (0.76)	3.95 (-0.21)	3.68 (-1.14)	3.84 (-0.06)	4.12 (-3.96)	6.29 (-3.51)	5.42 (0.59)	6.75 (1.90)	1.03 (-0.21)	3.48 (0.52)
Deira	66.61	7.19	12.18 (-0.94)	0.68 (-2.41)	5.05 (2.38)	5.73 (2.66)	3.15 (0.86)	9.18 (-0.63)	2.10 (-0.75)	3.28 (1.68)	2.74 (0.67)	11.62 (4.71)	1.89 (0.95)	2.41 (0.92)	6.05 (-3.80)	4.13 (0.64)	7.42 (-7.28)	5.21 (-3.53)	5.81 (0.50)	7.68 (2.44)	1.39 (-0.52)	2.30 (1.45)
Ecoli	50.79	6.92	9.49 (0.88)	1.17 (-2.27)	5.14 (1.97)	5.76 (2.94)	3.90 (-0.44)	7.36 (1.30)	2.28 (-0.94)	6.01 (0.77)	4.41 (0.55)	10.66 (2.10)	2.85 (1.14)	3.94 (0.23)	4.43 (-1.66)	4.44 (0.37)	5.54 (-4.64)	5.81 (-3.11)	5.40 (-0.35)	7.06 (1.33)	1.53 (-0.38)	2.85 (0.25)
Hamar	**61.12**	**4.97**	**10.41 (1.80)**	**0.76 (-1.96)**	**8.32 (4.01)**	**8.09 (4.38)**	**3.26 (0.88)**	**8.27 (0.57)**	**2.01 (-0.70)**	**4.38 (1.61)**	**2.00 (0.38)**	**8.81 (2.07)**	**1.88 (0.83)**	**2.59 (0.50)**	**4.59 (-3.12)**	**3.13 (0.21)**	**6.12 (-8.87)**	**5.95 (-4.98)**	**6.91 (-0.04)**	**8.64 (1.72)**	**1.15 (-0.37)**	**2.72 (1.06)**
Hasal	**65.71**	**5.12**	**12.31 (2.06)**	**0.75 (-1.72)**	**8.91 (4.70)**	**7.10 (3.73)**	**3.14 (1.28)**	**8.14 (-0.47)**	**2.22 (-0.66)**	**3.82 (1.73)**	**1.80 (0.65)**	**8.65 (2.93)**	**1.68 (0.79)**	**2.26 (0.61)**	**4.60 (-3.99)**	**2.82 (0.48)**	**6.51 (-10.47)**	**5.45 (-4.79)**	**6.83 (-0.07)**	**9.25 (2.27)**	**1.11 (-0.55)**	**2.66 (1.53)**
Hqwal	**47.86**	**5.09**	**9.30 (3.65)**	**0.78 (-2.81)**	**7.79 (3.54)**	**7.44 (3.76)**	**3.20 (-0.22)**	**7.46 (2.59)**	**2.13 (-1.58)**	**6.17 (-0.42)**	**2.37 (-0.44)**	**8.19 (-0.68)**	**1.82 (-0.06)**	**3.45 (-0.42)**	**4.36 (-0.63)**	**3.43 (-0.65)**	**5.84 (-3.42)**	**6.85 (-4.20)**	**7.87 (1.43)**	**7.78 (1.38)**	**1.02 (-0.48)**	**2.75 (-0.33)**
Metac	42.68	6.59	6.88 (2.17)	1.26 (-1.80)	5.34 (2.49)	7.97 (3.80)	4.44 (-2.30)	7.24 (1.59)	1.67 (-0.85)	7.36 (0.89)	6.55 (0.38)	9.39 (-1.03)	2.46 (0.83)	4.48 (0.35)	3.99 (-1.69)	2.55 (-1.20)	4.49 (-3.04)	6.90 (-3.01)	5.43 (0.57)	6.82 (1.87)	1.06 (-0.26)	3.73 (0.28)
Metka	61.16	6.24	8.34 (1.22)	1.32 (-0.53)	5.80 (1.86)	10.00 (5.65)	2.87 (0.58)	8.05 (-1.13)	1.93 (-0.90)	4.83 (2.02)	4.02 (2.17)	10.08 (2.82)	1.90 (1.02)	1.92 (0.19)	5.46 (-3.85)	1.41 (-0.65)	8.35 (-6.24)	4.62 (-6.42)	4.60 (-2.15)	10.44 (3.80)	1.24 (-0.33)	2.82 (0.88)
Metth	49.54	6.18	7.33 (2.46)	1.21 (-1.95)	5.91 (2.61)	8.14 (3.97)	3.65 (-0.01)	7.97 (-0.03)	1.88 (-1.93)	7.71 (2.46)	4.57 (1.30)	9.42 (-0.33)	3.07 (0.80)	3.31 (0.33)	4.30 (-3.70)	1.90 (-1.88)	6.79 (-1.07)	6.14 (-3.75)	4.96 (-0.55)	7.67 (2.06)	0.84 (-1.09)	3.22 (0.28)
Napha	**63.12**	**4.85**	**11.38 (1.91)**	**0.78 (-1.74)**	**8.74 (4.46)**	**8.84 (5.10)**	**3.25 (0.88)**	**8.40 (0.31)**	**1.98 (-0.48)**	**4.20 (1.76)**	**1.90 (0.17)**	**8.75 (2.42)**	**1.73 (0.69)**	**2.22 (0.36)**	**4.61 (-3.34)**	**2.60 (0.11)**	**6.39 (-9.98)**	**5.28 (-5.61)**	**6.38 (-0.22)**	**8.88 (2.27)**	**1.06 (-0.23)**	**2.64 (1.16)**
Pyrae	51.36	7.86	9.89 (1.95)	0.87 (-1.69)	4.30 (1.95)	7.00 (3.27)	3.63 (-0.05)	7.67 (-0.03)	1.51 (-0.91)	6.29 (1.62)	5.68 (1.40)	10.57 (0.16)	1.93 (0.85)	2.60 (-0.46)	4.98 (-2.47)	2.08 (-1.04)	6.55 (-4.17)	4.93 (-3.95)	4.41 (-1.02)	9.32 (3.90)	1.47 (-0.29)	4.31 (0.96)
Strco	72.00	6.81	13.73 (1.04)	0.78 (-1.71)	6.14 (2.64)	5.69 (2.80)	2.64 (1.57)	9.60 (-1.95)	2.36 (-0.34)	2.87 (1.73)	2.05 (1.21)	10.21 (4.76)	1.70 (0.96)	1.70 (0.92)	6.20 (-5.32)	2.66 (0.24)	8.38 (-8.99)	4.98 (-3.73)	6.18 (0.36)	8.53 (2.72)	1.53 (-0.27)	2.05 (1.34)
Sulso	35.79	7.73	5.60 (2.34)	0.62 (-1.68)	4.68 (2.09)	6.81 (2.96)	4.44 (-1.66)	6.43 (2.12)	1.29 (-1.25)	9.44 (0.56)	7.74 (1.64)	10.35 (-2.41)	2.21 (0.52)	4.96 (-0.31)	3.81 (-0.45)	2.10 (-0.76)	4.72 (-1.61)	6.70 (-2.99)	4.73 (-0.46)	7.47 (2.18)	1.06 (-0.32)	4.83 (-0.53)
Theac	45.99	6.87	6.97 (1.24)	0.58 (-2.49)	5.76 (2.09)	6.02 (1.98)	4.69 (0.25)	7.26 (1.44)	1.63 (-1.78)	9.04 (1.77)	5.68 (1.63)	8.37 (-0.74)	3.20 (0.86)	4.25 (0.79)	3.94 (-1.82)	2.15 (-1.34)	5.48 (-3.54)	7.53 (-2.38)	4.78 (-0.13)	7.18 (2.32)	0.85 (-0.82)	4.64 (0.66)



Supplemental Figure 2.13: **Number of stop codons per 1000 nucleotides** for selected genomes with different GC contents. GC-rich species potentially possess fewer stop codons (TAA, TGA, TAG), a fact that potentially leads to gene overprediction and start codon misassignments (Figure 2.9).

CHAPTER 3

Living with two extremes: Conclusions from the genome sequence of *Natronomonas pharaonis*

Natronomonas pharaonis is an extremely haloalkaliphilic archaeon that was isolated from salt-saturated lakes of pH 11. We sequenced its 2.6-Mb GC-rich chromosome and two plasmids (131 and 23 kb). Genome analysis suggests that it is adapted to cope with severe ammonia and heavy metal deficiencies that arise at high pH values. A high degree of nutritional self-sufficiency was predicted and confirmed by growth in a minimal medium containing leucine but no other amino acids or vitamins. Genes for a complex III analog of the respiratory chain could not be identified in the _N. pharaonis_ genome, but respiration and oxidative phosphorylation was experimentally proven. These studies identified protons as coupling ion between respiratory chain and ATP synthase, in contrast to other alkaliphiles using sodium instead. Secretome analysis predicts many extracellular proteins with alkaline-resistant lipid anchors, which are predominantly exported through the twin-arginine pathway. In addition, a variety of glycosylated cell surface proteins probably form a protective complex cell envelope. _N. pharaonis_ is fully equipped with archaeal signal transduction and motility genes. Several receptors/transducers signalling to the flagellar motor display novel domain architectures. Clusters of signal transduction genes are rearranged in haloarchaeal genomes whereas those involved in information processing or energy metabolism show a highly conserved gene order.

3.1 Introduction

Strains of the *Natronomonas pharaonis* were first isolated from highly saline soda lakes in Egypt (Soliman and Truper 1982) and Kenya (Tindall et al. 1984) which show pH values around 11. Such alkaline brines are enriched with carbonate and chloride resulting in a scarcity of magnesium and calcium. The aerobic haloalkaliphilic euryarchaeon *N. pharaonis* thrives optimally in 3.5 M NaCl and at a pH of 8.5, but is sensitive to high magnesium concentrations.

Since plasma membranes and protoplasts lose their stability at high pH, it has been suggested that one of the key features of alkaliphilic specialization is associated with the cell envelope protecting the cell from alkaline conditions (Horikoshi 1999). *Bacillus* spp. contain acidic polymers which may support the adsorption of sodium and protons but repulse hydroxide ions. Haloalkaliphilic archaea have also been reported to possess unique cell walls consisting of glutaminylglycan polymers (Kandler and Konig 1998) as well as characteristic membranes containing C_{20}-C_{25} in addition to C_{20}-C_{20} diether core lipids (Tindall et al. 1984). Low extracellular proton concentrations further effect membrane-linked energetics due to immediate neutralization of extruded protons. In order to deal with alkaline conditions, some bacteria replace protons by sodium ions as the coupling ion rather than increasing $\Delta\Psi$ (Skulachev et al. 1999).

A wide range of extracellular enzymes such as alkaline proteases, amylases, and cellulases, has been isolated from alkaliphiles, and were utilized for industrial production of laundry detergent additives and cyclodextrins (Horikoshi 1999). Most alkaline enzymes have been described in *Bacillus* spp., but the haloalkaliphilic archaeon *Natronococcus occultus* was also found to produce an haloarchael α-amylase and to exhibit extracellular proteolytic activity (Horikoshi 1999).

N. pharaonis has been phylogenetically classified within the order Halobacteriales, which includes the intensively studied *Halobacterium salinarum*. *Natronomonas* cells are motile (Soliman and Truper 1982) and actively search for optimal growth conditions with the help of retinal proteins responsible for light-dependent ion transport and sensory functions. For *N. pharaonis* (strain SP1), the chloride pump halorhodopsin (Lanyi et al. 1990) and sensory rhodopsin II (Seidel et al. 1995) with its transducer HtrII (Klare et al. 2004) have been described in detail.

Here we report the complete *N. pharaonis* genome and complementing experimental results. This study identified novel adaptation strategies of alkaliphiles regarding its respiratory chain, nitrogen metabolism, and its cell envelope.

3.2 Results and discussion

3.2.1 Genome and gene statistics

The genome of *Natronomonas pharaonis* type strain Gabara (DSM 2160) consists of three circular replicons, the 2.6 Mb chromosome, a typical haloarchaeal 131 kb plasmid (PL131) and an unique multicopy 23 kb plasmid (PL23) (Table 3.1). The GC-rich chromosome (63.4% GC) contains an integrated copy of PL23, and features four regions of reduced GC

Table 3.1: **Basic data for the *N. pharaonis* replicons.**

	CHR	PL131	PL23
Length (bp)	2595221	130989	23486
GC content	63.4%	57.2%	60.6%
Sequence coverage (normalized values)	5.8 (1.0)	3.9 (0.67)	100.5 (17.3)
% coding	90.8%	82.3%	83.9%
Encoded proteins	2675	132	36
Avg. protein length (aa)	293	271	183
Encoded stable RNAs	51	-	-

content (GC-poor regions I - IV) as well as several transposases (illustrated in Figure 3.1) (Additional information on transposases, plasmids, and GC-poor regions is provided as Supplemental text 3.5.1). The replication origin of *N. pharaonis* is delineated by a 30 bp inverted repeat (302280-302311, 302712-302681) and an adjacent Cdc6 homolog (*cdc6_1*, NP0596A), that are very similar to the recently identified inverted repeat and the *orc7* gene of the *Halobacterium* NRC1, respectively (Berquist and DasSarma 2003). Interestingly, haloarchaeal origins are found around the maximum, not the minimum, of the cumulative GC skew plot.

By rigourous evaluation of the automatic gene finder data, 2843 protein-coding and 51 RNA genes were predicted for the GC-rich *N. pharaonis* genome. This process was greatly facilitated by the close relationship between *Natronomonas* and other halophilic archaea with completely sequenced genomes, *Halobacterium salinarum* (of which two strains have been sequenced with minimal sequence deviation: strain R1 (Oesterhelt et al., unpublished, www.halolex.mpg.de) and strain NRC-1 (published as *Halobacterium sp.* NRC-1 (Ng et al. 2000))) as well as *Haloarcula marismortui* (Baliga et al. 2004). The accuracy of our gene selection is further enhanced by the availability of genome-scale proteomic data for *H. salinarum* strain R1 (Klein et al. 2004; Tebbe et al. 2004) (www.halolex.mpg.de) and confirmed by genome-wide proteomic data for *N. pharaonis* (F. Siedler, unpublished; a single false negative was uncovered in a set of several hundred identified proteins).

The subsequent assessment of the applied gene prediction tools by comparison with the validated *N. pharaonis* gene set revealed significant improvement of gene selection by the REGANOR approach as compared to that of the underlying programs CRITICA and adequately trained GLIMMER (see Methods). However, about 400 of the start codons predicted by REGANOR and CRITICA (14-15%) and even a third of the GLIMMER starts were reassigned, mostly because predicted genes were too long.

A total of 43% of the *N. pharaonis* genes are likely to be co-transcribed, and 322 gene pairs revealed typical transcription unit overlaps of 1 or 4 bases resulting from a shift of open reading frames by -1. The gene order of transcription units coding for ribosomal proteins,

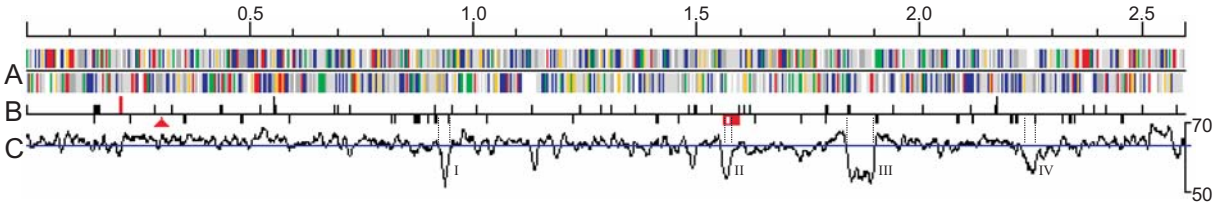


Figure 3.1: **Linear representation of the *N. pharaonis* chromosome.** The scaling on top is in Mb.
(A) Proteins (above line: forward strand, below line: reverse strand) are coloured by function category (blue: metabolism, red: genetic information processing, green: transport and cellular processes, yellow: environmental response).
(B) Above line RNA genes (long red: rRNAs of the single rRNA operon, long black: RNAseP RNA and 7S RNA, short: tRNAs), and below line transposase genes, the replication origin (red triangle), and an integrated copy of PL23 with its 13 kb insertion (red bar with white box) are shown. Four GC-poor regions are marked by vertical lines.
(C) GC content of the coding regions (black, computed with a window of 30 proteins) and average GC content (blue line, 63.4% GC). GC-poor regions I to IV are marked.

RNA polymerase and membrane complexes involved in energy metabolism (*atp*, *nuo* clusters) and pH adaptation (*pha* cluster) is very well conserved between *Halobacterium* and *Natronomonas*. In contrast, gene clusters involved in signal transduction and motility are extensively rearranged between the genomes of halophiles (Figure 3.5 in Chapter 3.2.8).
Cytoplasmic *N. pharaonis* proteins contain a high proportion of acidic amino acids (average 19.3%) resulting in low isoelectric points (average pI 4.6). These are typical adaptative features of haloarchaea, which are known to apply the salt-in strategy (high internal salt concentrations) in order to cope with their hypersaline environment. In addition, *N. pharaonis* was reported to produce the compatible osmolyte 2-sulfotrehalose (Desmarais et al. 1997), and its genome codes for a typical compatible solute transporter (*tp58*, NP3588A). However, homologs of genes for the *de novo* synthesis of common compatible solutes such as glycine betaine (*betAB*) and trehalose (*otsAB*) or an osmolyte-binding protein (*cosB*) could not be detected.

3.2.2 Function analysis

Of the 2843 proteins, 65% were grouped to an orthologous cluster (COG), and specific and general functions could be assigned for 45% and 12% of the predicted proteome, respectively. Proteins with specific function belong to the functional categories metabolism (17%), transport and cellular processes (9%), genetic information processing (7%), environmental information processing (5%), and miscellaneous (7%) (see also Table S2). For the remaining proteins only general functions (12%), partly derived through gene-context analyses, or no functions (43%) were assigned. One fifth of the *N. pharaonis* proteins have no homologs in other species (singletons). Amongst them are 2 probable *Natronomonas*-specific membrane complexes, each encoded by nine adjacent genes (NP2336A - NP2352A, NP5354A - NP5338A), which are highly similar and arranged in the same gene

order. All fifteen previously published *N. pharaonis* genes could be identified within the genome, although minor sequence variations were detected usually due to strain differences. However, a partially-sequenced protease from *N. pharaonis,* which is extremely similar to vertebrate chymotrypsin (Stan-Lotter et al. 1999), could not be found.

3.2.3 Central metabolism and transport

Metabolic enzymes comprise a large number of fatty-acid degradation genes, and the complete set of enzymes involved in biosynthetic pathways leading to amino acids and coenzymes. Thus, the chemoorganotrophic *N. pharaonis*, which is usually grown on media with amino acids as carbon source, has a high degree of nutritional self-sufficiency. In agreement with this, we were able to simplify the synthetic medium for this species omitting all amino acids except leucine. Requirement of this amino acid might be caused by disruption of the isopropylmalate synthase gene (*leuA_1*, NP2206A) in the 5' region. As *H. salinarum*, *N. pharaonis* is likely not capable of sugar utilization due to the lack of genes encoding key enzymes of glycolytic pathways. A gene cluster (NP4962A – NP4944A) encoding a set of probable anaerobic dehydrogenase subunits might indicate growth of *N. pharaonis* cells under anaerobic or micro-aerophilic conditions. Genes encoding mevalonate pathway enzymes are present, as other genes required for the synthesis of membrane lipids and other components such as menaquinones (Soliman and Truper 1982) and retinal pigments (Seidel et al. 1995), which are derived from prenyl-precursors. A large set of predicted transporter genes was found, including homologs of transporters for metal ions such as iron, manganese, copper, and cobalt, which are scarce in a highly alkaline environment. A more extensive description of the predicted metabolism and transport of *N. pharaonis* is given as Supplemental text 3.5.2.

3.2.4 Nitrogen metabolism

Natronomonas grows under highly alkaline conditions in brines of pH around 11 (Soliman and Truper 1982; Tindall et al. 1984). These extreme pH conditions cause not only low availability of metal ions, but also reduced levels of ammonium ions. For the uptake of ammonium, which can be assimilated in the central metabolite glutamate, *N. pharaonis* possesses transporters for several exogenous nitrogen sources, ammonium (AmtB, NP0922A), nitrate/nitrite (NarK, NP4228A), and urea (ABC transporter UrtABCDE, NP1996A-NP2004A) (Figure 3.2). Genes involved in nitrate reduction (*narB_1* (NP4226A), *nirA_1* (NP4224A)) and urea conversion (*ure* cluster (NP2008A-NP2020A)) to ammonium were found to be clustered with their respective transporter genes. The observed nitrate assimilation pathway has also been described for cyanobacteria (Hirasawa et al. 2004), and the enzymes involved show 34-51% sequence identity to those present in *N. pharaonis*. It is

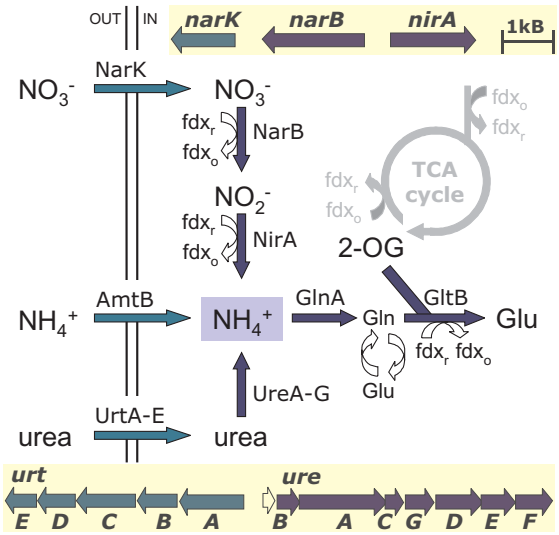


Figure 3.2: **Schematic representation of the nitrogen metabolism of *N. pharaonis*.** The figure illustrates proposed transport (green arrows) and metabolic processes (blue arrows) for nitrogen compounds and the associated gene clusters (*nark-narB-nirA*, *urtABCDE*, *ureBACGDEF*). Ammonia can be supplied by (a) direct uptake, (b) uptake and reduction of nitrate (NO_3^-) via nitrite (NO_2^-), and (c) uptake and hydrolysis of urea. It is utilized by a two-step reductive conversion of 2-oxoglutarate (2-OG) to glutamate (Glu) involving ferredoxin (fdx). The two *glnA* paralogs, *gltB*, and *amtB* were found distributed over the genome.

likely that *Natronomonas* uses ferredoxin and not NADH as the electron donor for all three reductive conversions. This view is supported by the occurrence of conserved ferredoxin-binding residues within the *N. pharaonis* NirA protein (Hirasawa et al. 1998) and ferredoxin-dependence of nitrate and nitrite reductases in the halophile *Haloferax mediterranei* (Martinez-Espinosa et al. 2001). Nine ferredoxins of 4 orthologous groups (COG0633, COG1141, COG1146, COG3411) are present in *N. pharaonis*, and ferredoxin appears to be the common proteinaceous electron carrier for functional N-assimilation as well as the conversion of 2-oxoacids and aldehydes.

While *H. marismortui* also possesses all necessary genes for urea conversion and nitrate assimilation but also for nitrate respiration, *H. salinarum* lacks these genes. Urea cycle enzymes for the conversion of ornithine to arginine are encoded in all 3 halophilic strains. With respect to arginine degradation, halophiles adopted different strategies. *H. marismortui* splits arginine into ornithine and urea by arginase (EC 3.5.3.1) (rrnAC0383, rrnAC0453) while *H. salinarum* employs plasmid-encoded enzymes of the arginine deiminase pathway for the fermentation of arginine. In this pathway arginine is converted to ornithine and carbamoylphosphate, which is further degraded to CO_2 and NH_3 with concomitant ATP generation (Ruepp and Soppa 1996). *N. pharaonis* lacks both pathways for arginine utilization.

3.2.5 Respiratory chain

The available biochemical data on electron transport chain components for *N. pharaonis* (Scharf et al. 1997) and other respiratory archaea (Schafer et al. 1996) as well as the genomic data from Halolex (www.halolex.mpg.de) and STRING databases (von Mering et al. 2003) were used to generate an archaeal profile of the electron transport chain (Figure 3.3A). The profile revealed a high degree of plasticity in the composition of the

respiratory chain often differing greatly from the 'classical' five complex systems found in mitochondria. Notably, type, number, and composition of terminal oxidases vary widely in the archaeal domain of life. Furthermore, NADH is not oxidized by proton-pumping type I NADH dehydrogenase complexes because the NADH acceptor module (*nuoEFG*) is absent. In *N. pharaonis* NADH dehydrogenation is likely to occur via the non-proton-pumping type II NADH dehydrogenase encoded by a homolog (NP3508A) of the *ndh* gene characterized in *Acidianus ambivalens* (Gomes et al. 2001). All subunits of the succinate dehydrogenase (*sdhCDBA* (NP4264A-NP4270A)) are present and menaquinone biosynthesis genes were found supporting its proposed function as mobile carrier (Scharf et al. 1997). However, no complex III subunits could be identified in the genome in spite of searching with the respective *H. salinarum* genes *(petABC)*. The crenarchaeote *A. ambivalens* is able to channel electrons from complex I/II directly into a terminal quinole oxidase using a mobile quinol carrier (Gomes et al. 2001). None of the three terminal oxidases in *N. pharaonis* though, shows sufficient sequence similarity to support the idea that one of them is a quinole oxidase. The *N. pharaonis* cytochrome ba_3-type oxidase (*cba* cluster (NP2960A-NP2968A)) (Mattar and Engelhard 1997) interacts with the blue copper protein halocyanin (*hcp_1* (NP3954A)), and the *cbaD* subunit of the orthologous complex in *H. salinarum* occurs as a fusion protein with halocyanin. The propable electron transport from halocyanin to the terminal oxidase complex in haloarchaea, indicates a novel yet unknown type of complex III, which mediates the electron transfer step between menaquinone to halocyanin.

Through physiological experiments we showed that *N. pharaonis* is able to eject protons during respiration (oxygen-induced acidification) with a subsequent increase in ATP levels (oxidative phosphorylation) (Figure 3.3B,C). Since both effects are sensitive to the protonophore CCCP, *N. pharaonis* needs to possess a functional respiratory chain including a component responsible for proton pumping upon electron transfer (either one of the oxidases or the postulated complex III analog). Furthermore, CCCP did not prevent light-induced alkalinization* by the light-driven chloride pump halorhodopsin, but did prevent the subsequent increase in ATP levels. Thus, the *N. pharaonis* ATP synthase also operates with protons, and completes a full proton circuit. This finding shows that *N. pharaonis* does not replace protons by sodium ions as coupling ion between respiratory chain and ATP synthase. In contrast, the alkaliphilic bacterium *Bacillus halodurans* FTU switches from proton to sodium energetics in case of alkaline conditions or the addition of a protonophore (Skulachev 1992). Induction of respiratory complexes and ATPases with altered ion-specifity

* Pumping of chloride ions into the cell causes a passive cation flow. The relative contribution of the different cations depends on their membrane permeability. The protonophore CCCP selectively increased membrane permeability for protons and accordingly results in increased alkalinization.

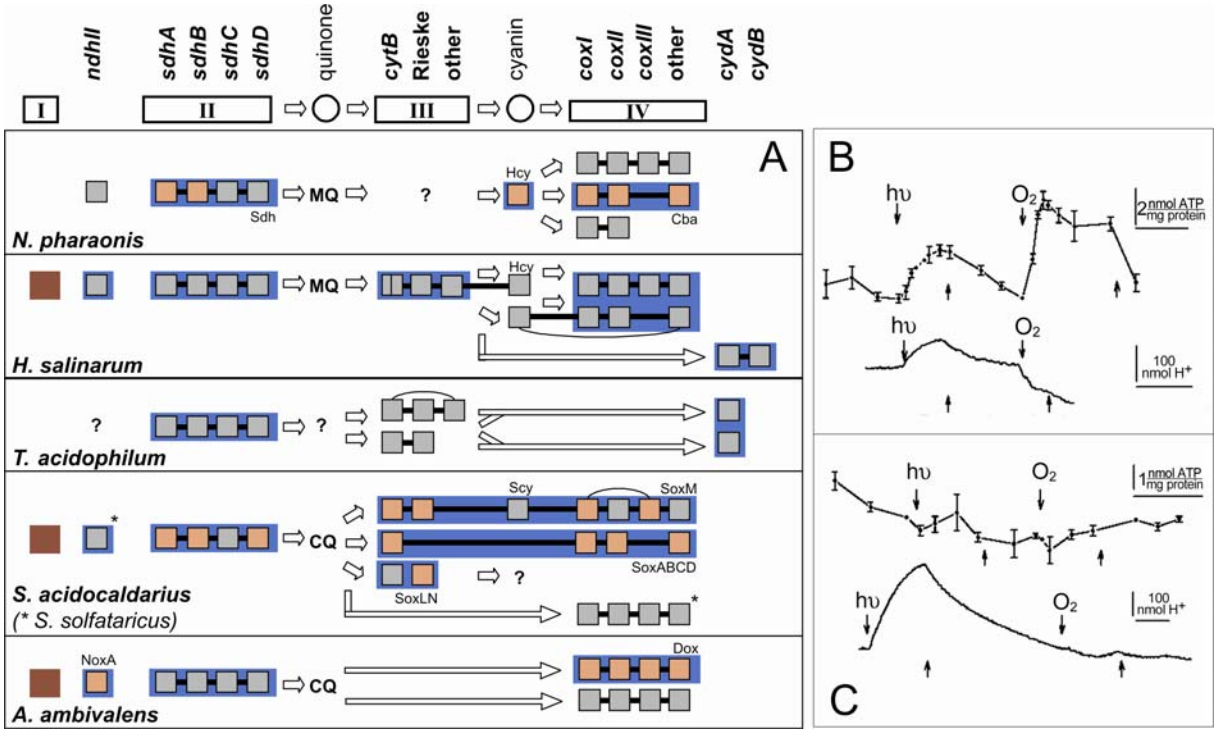


Figure 3.3: **Genomic and experimental studies of the *N. pharaonis* electron transport chain.**
(A) The electron transport chain profile for several respiratory archaea displays subunits (squares) of respiratory complexes (boxes I-IV). These subunits are often encoded adjacent in the archaeal genomes (straight connections) and can be fused to each other (curvy connections). The proposed electron flow between respiratory complexes is indicated by arrows. Complexes, which have been characterized experimentally, are indicated in blue, and protein-sequenced subunits of isolated complexes are coloured yellow. The profile for *Sulfolobus acidocaldarius* (Sa) was established by complementation with genetic data from the completely sequenced *Sulfolobus solfataricus* (Ss) (asterisks). Menaquinone (MQ), caldariellaquinone (CQ), halocyanin (Hcy), and sulfocyanin (Scy) are shown or predicted to function as mobile carriers in archaeal respiratory chains. Homologs to complex I subunits have been found in archaeal genomes, but genes encoding the NADH acceptor module are missing. A functional NADH-dehydrogenating complex I has been experimentally excluded for three species (red). Instead, it is replaced by NADH dehydrogenase type II, which is not capable of proton translocation.
(B), (C) Oxidative and photo-phosphorylation processes in *N. pharaonis* cells were investigated through measurements of ATP levels (upper curves, mean values with error bars from triplicates) and extracellular pH (lower curves, continuous recording at a rate of 10/s) in the (B) absence or (C) presence of the protonophore CCCP (0.2 mM). The effects (ON: above curve arrow, OFF: below curve arrow) of light ("hv", λ >515 nm, 32 mW/cm^2) and aeration ("O_2") were determined. All experiments were performed at pH 8.1. Vertical scaling bars indicate ATP level and amount of proton uptake, whereas horizontal scaling bars indicate a 10 min time interval.

permits alkaliphiles to cope with inversed proton gradients requiring no increase of ΔΨ and avoiding reduced ATP yield as well as the risk of membrane electric breakdown (Skulachev et al. 1999). However, dependent on the external pH, varying internal pH values were observed for alkaliphilic *Bacillus* strains, and the internal pH reaches values over pH 9 (Horikoshi 1999). For *N. pharaonis* we measured internal pH values up to 9.3. By accepting high pH values in the cytoplasm, the difference between intra- and extracellular pH remains moderate, and consequently protons are permissible as coupling ion. It should be noted, that

our data are not in agreement with the previous suggestion that chloride is used as alternative coupling ion by *N. pharaonis* (Avetisyan et al. 1998).

3.2.6 Secretion and membrane anchoring

N. pharaonis encodes the same probable components of the Sec and Tat (twin-arginine) protein translocation pathways which were found in the *Halobacterium* genome (Pohlschroder et al. 2004). For the transport of archaeal flagellins involvement of *flaI* and *flaJ* genes was proposed (Thomas et al. 2001), since those show similarity to components of type II/IV secretion systems and were described to influence flagella formation (Patenge et al. 2001). To estimate the contribution of different systems for protein secretion in *N. pharaonis*, the complete secretome was predicted as described in Materials and Methods (Figure 3A, see also Table S3). This and previous genome surveys for *H. salinarum* strain NRC-1 (Bolhuis 2002; Rose et al. 2002) found that the Tat system, which secretes folded proteins, might be extensively used in haloarchaea compared to non-halophilic archaea. Haloarchaea probably utilize the Tat pathway not only for coenzyme-containing redox components like halocyanins (e.g. *hcp_4* (NP0050A)) and thioredoxins (e.g. *trx_1* (NP3914A)), but also for the export of non-redox proteins such as substrate-binding proteins of ABC transport systems (e.g. *dppA_2* (NP3578A)) and chemotactic signal transduction (e.g. *H. salinarum cosB* (OE3476R)). Thus, an adaptation to the high-salt and alkaline environment may involve the avoidance of protein folding in the extracellular space where chaperones are absent (Rose et al. 2002). However, there are also several proteins in *Natronomonas*, which are likely co-translationally delivered to and exported through the Sec system, among them 3 extracellular subtilisin-like proteases (NP1682A, NP2654A, NP4628A).

For the *N. pharaonis* halocyanin (*hcp_1* (NP3954A)), attachment of a diphytanyl chain to a N-terminal cysteine residue has been shown previously (Mattar et al. 1994). Prokaryotic lipoproteins are cleaved by signal peptidase II upstream of a cysteine residue that is part of a conserved motif (lipobox), and subsequently modified by lipid anchor attachment to the cysteine residue of the processed N-terminus (Hayashi and Wu 1990). *N. pharaonis* halocyanin exhibits a probable lipobox motif (‘LAGC’) indicating signal peptidase II-like processing before the observed lipid attachment. However, a signal peptidase II homolog is absent as in other archaeal genomes, and an alternative signal peptidase II-like protease remains to be identified. The *N. pharaonis* genome encodes a large number of proteins (91) containing lipobox motifs as adapted from PROSITE (PS00013) (see Methods). Probable N-terminal anchored lipoproteins comprise a third of the predicted secretome, and seem to be commonly translocated via the Tat pathway (85 proteins) (Figure 3.4A). More than 20 transporter subunits, 6 of the halocyanin homologs, and many *Natronomonas*-specific

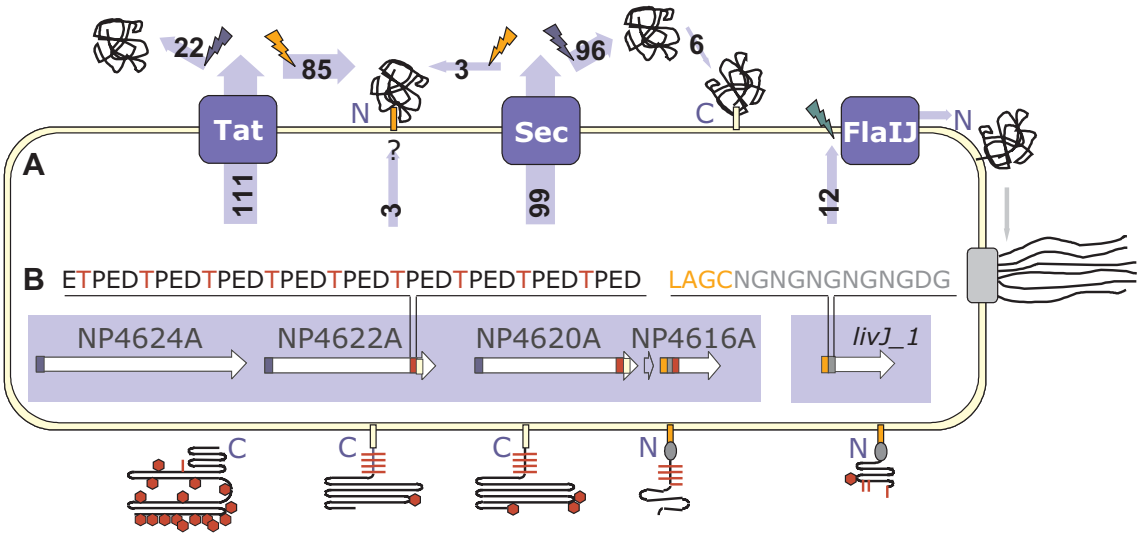


Figure 3: Schematic representation of protein secretion, anchoring, and glycosylation in *N. pharaonis*.
(A) Substrates of the Tat, Sec and flagellin-specific protein translocation systems (blue boxes) are cleaved by signal peptidases (flash signs), and partly remain C- or N-terminally anchored to the cell membrane. Secreted proteins are cleaved by signal peptidase type I (blue), whereas lipobox-containing proteins are cleaved by signal peptidase type II (orange) and N-terminally attached to a lipid-anchor (orange box). Lipoproteins are frequently transported via the Tat pathway (substrates numbers indicated in the light-blue arrows). For 3 lipobox-containing proteins the export pathway remains as yet unassigned. Furthermore, 6 proteins are likely to be modified by a C-terminally attached lipid anchor (yellow box). After cleavage by membrane-bound preflagellin peptidase (green), the substrates of the flagellin-specific export pathway reveal an N-terminal hydrophobic stretch possibly involved in membrane retention.
(B) Signal sequence and peptide repeat modules (marked by coloured boxes) for a representative gene cluster (white arrows) are presented diagrammatically in models of the cell surface proteins. A Thr-rich tetrapeptide repeat (red box in genes, red line in proteins), likely to be O-glycosylated, occurs in several cell surface proteins adjacent to the C-terminal or N-terminal lipid anchor. An Asn-Gln dipeptide repeat (grey box, oval) follows directly after the lipobox-containing Tat-related signal sequence (orange box) of several membrane components. Other indicated features are Sec-related signal sequences (blue box in genes) and N-glycosylation sites (red hexagons in proteins).

proteins of unknown function are found amongst predicted lipoproteins. Halophilic archaea have a higher fraction of lipobox-containing proteins than non-halophilic archaea, the highest fraction being found in *N. pharaonis* (*N. pharaonis*: 91 (3.20% of all proteins), *H. maris-mortui:* 116 (2.74%), *H. salinarum*: 49 (1.74%) vs. *A. fulgidus:* 16 (0.66%), *M. mazei:* 23 (0.68%), *P. furiosus:* 9 (0.42%)). Lipoproteins seem to be more common in bacteria (*B. subtilis:* 57 (1.39%), *E. coli:* 90 (2.12%), *C. glutamicum:* 77 (2.57%)). The retention of 34% of secreted *N. pharaonis* proteins by lipid anchors may reflect a protection mechanism against alkaline extraction of proteins from the cell membrane, an effect commonly exploited to deplete membrane preparations of peripheral membrane proteins (Klein et al. 2004).
Interestingly, 8 of the putative lipoproteins have Asn-Gly dipeptide repeats directly following the lipobox (Figure 3B), amongst them 2 halocyanins (*hcp_1* (NP3954A), *hcp_2* (NP4744AA)) and 3 substrate-binding proteins (*dppA_1* (NP0758A), *livJ_1* (NP4140A), *sfuA*

(NP5000A)). These proteins are likely involved in interactions with membrane protein complexes (respiratory complexes, ABC transporters), thus, the repeat regions (length up to 24 aa) might function as flexible hinges promoting protein interactions.
Apart from the putative lipobox-containing proteins, 12 proteins with probable cleavage sites ('RGQ') for preflagellin-peptidase (*flaK* (NP1276A)), as previously proposed for halobacterial flagellins (Thomas et al., 2001), were identified. These proteins include not only the 3 flagellins (NP2086A-NP2090A), but also proteins with unknown functions, most of them encoded adjacent to each other and to genes similar to components of type II/IV secretion systems.

3.2.7 Cell envelope

Thr-rich tetrapeptide repeats varying in length and amino acid composition were detected in 9 *Natronomonas*-specific proteins, three of whose genes (NP4616A, NP4620A, NP4622A) clustered within one genomic region (Figure 3.4B). Four of the repeat-containing proteins revealed high regional similarity to the N- and C-termini of halophilic cell surface glycoproteins (*csg*) which form regular S-layer cell envelopes. Within these similar regions three topological features have been described for the *H. salinarum* Csg; an N-terminal signal sequence, a C-terminal lipid-anchor region (Kikuchi et al. 1999) and - directly in front of this membrane anchor - a pattern of O-glycosylated threonines (Lechner and Sumper 1987). In all *N. pharaonis* proteins with Thr-rich tetrapeptide repeats, secretion, lipid-retention and glycosylation signals were present (Figure 3B). Interestingly, the Thr-rich repeating units are located not only next to the C-terminal lipid anchor regions as in *H. salinarum*, but also adjacent to N-terminal lipid anchors. This finding suggests that the likely function of observed Thr-rich tetrapeptide repeat patterns is as a glycosylated spacing region that forms a periplasma-like reaction space between cell envelope and membrane. Because *N. pharaonis* encodes several proteins with Csg-like features, it might not possess a typical S-layer, but instead form a more complex cell envelope consisting of various glycoprotein species with distinct saccharides.

3.2.8 Motility and signal transduction

The proposed signal transduction cascade of *Natronomonas* and *Haloarcula* is very similar to that described for *Halobacterium* (Rudolph and Oesterhelt 1996), and consists of signal receptors/transducers, the two-component regulatory system (*cheA/Y* (NP2172A, NP2102A)), an adaptation module (*cheR/B* (NP2170A, NP2174A)), and the flagella with its motor. However, there are many differences in the organization of the genes in the genome and the domain architecture of individual components (Figure 3.5, further detailed as Supplemental text 3.5.3).

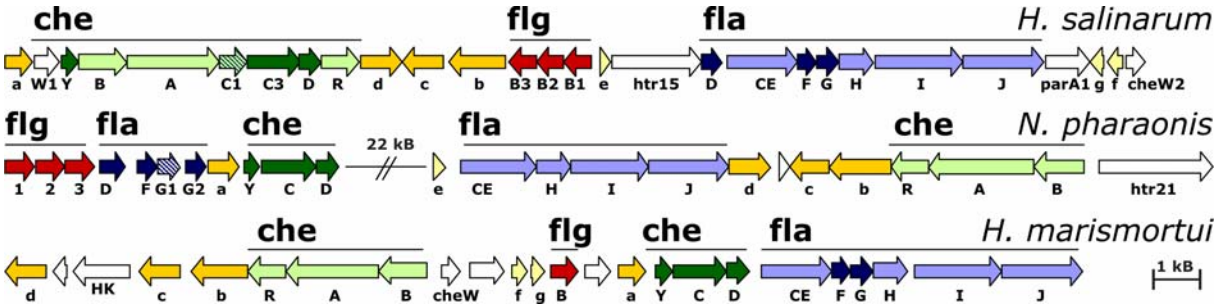


Figure 3.5: **The MO-ST cluster in three halophilic genomes.** Flagellin (flg) genes are shown in red, genes for flagellin-associated proteins (fla) in blue, chemotaxis (che) genes in green, and genes for additional conserved proteins in yellow. Che genes are separated in two subsets (light and dark green) in *Natronomonas* and *Haloarcula* but are arranged into a single cluster in *Halobacterium*. Similarly, fla genes are separated in two subsets (light and dark blue) in *Natronomonas* but again are arranged into a single cluster in *Halobacterium* and *Haloarcula*. Hatched arrows indicate species-specific duplicated neighbouring genes. Additional conserved genes are dark yellow (present in all species) or light yellow (present in two species) and are given in lowercase letters. The MO-ST cluster of *Natronomonas* is interrupted by a 22 kb insertion. Genes involved in the environmental response are marked by their gene names. Several *Halobacterium* genes are renamed as compared to our previous publication (Rudolph and Oesterhelt 1996): *cheC1 (cheJ), flaCE (flaE), parA1 (flaK)*.

Major domain shuffling is also evident when comparing the 19 *Natronomonas* and the 18 *Halobacterium* transducers (also known as methyl-accepting chemotactic proteins) that participate in the signal transduction cascade. An example are the transducers mediating the phototactic response. *Natronomonas* has a single blue-light photoreceptor (sensory rhodopsin II, SRII (NP4834A)) which is photochemically very similar to the blue-light photoreceptor SRII from *Halobacterium* but rather different from the orange/UV light photoreceptor SRI (Lutz et al. 2001). Although the transducers HtrII from these two archaea form a complex with their respective blue-light photoreceptors SRII (thus mediating the same photophobic response) and genes for receptor and transducer are cotranscribed (Seidel et al. 1995; Zhang et al. 1996), their domain architecture differs. *Natronomonas* HtrII (NP4832A) is characterized by a short loop between its two transmembrane domains (TM2SL) and shares its domain architecture with that of *Halobacterium* HtrI and three paralogs from *Natronomonas* (NP1756A, NP3122A, NP3134A) (Figure 3.6). Four of the five TM2SL-type transducers also show similarity in their gene context. Their genes are shown (Seidel et al. 1995; Yao and Spudich 1992; Zhang et al. 1996) or predicted to be co-transcribed by forming a transcription unit with a retinal protein or a distant homolog thereof (NP1758A, NP3132A). In contrast, HtrII of *Halobacterium* shares its domain architecture, characterized by a long extracellular domain between its two transmembrane domains (TM2ED), with six paralogs, several of which are involved in chemotaxis as determined experimentally (Kokoeva et al. 2002) or indicated by gene context. Whereas *Halobacterium* and *Haloarcula* have multiple TM2ED-type transducers and share several orthologous gene pairs, *Natromononas* has mainly TM2SL-type transducers, which lack an

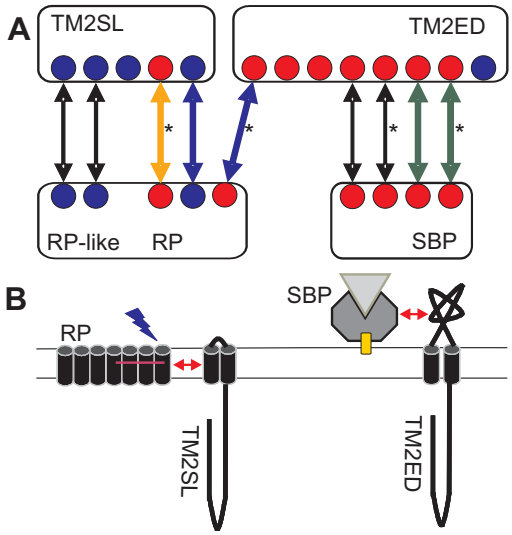


Figure 3.6: **Domain architecture and gene context of several transducers from *Natronomas* and *Halobacterium*. (A)** Two groups of transducers (upper boxes) with distinct domain architectures and their adjacent genes (arrows and lower boxes) from *Natronomonas* (blue dots) and *Halobacterium* (red dots) are schematically represented. Transducers with a long extracellular domain between their two transmembrane domains (box "TM2ED") are frequently involved in chemotaxis, and are co-transcribed with the genes for periplasmic substrate-binding proteins (box "SBP"). Several of the transducers with a short loop between their two transmembrane domains (short-loop transducers, box "TM2SL") are co-transcribed with retinal-containing photoreceptors (box "RP") or distant homologs thereof. Experimental environmental response data are indicated by coloured arrows (green: chemotaxis (Kokoeva et al. 2002), blue: blue-light phototaxis (Seidel et al. 1995; Zhang et al. 1996), orange: orange-light phototaxis (Yao and Spudich 1992). The existence of orthologous gene pairs in *Haloarcula* is indicated by asterisks.
(B) The interaction between a short-loop transducer (TM2SL) and a retinal protein (RP) occurs within the membrane. The interaction between a chemotactic extracellular-domain transducer (TM2ED) with a lipid-anchored (yellow box) periplasmic substrate-binding protein (SBP) may occur outside of the membrane.

extracellular domain. As a result, *Natronomonas* may have either reduced chemotactic capabilities or chemotaxis may be mediated by other transducers with different domain architectures. A likely candidate for a chemotactic transducer is Htr32 (NP6128A), which has a substrate-binding domain directly fused to its signalling domain. Further transducers with unusual domain architectures are Htr34 (NP1642A) and Htr35 (NP1486A), which have an N-terminal signalling domain with a long C-terminal extension.

3.3 Conclusions

In conclusion, we describe a number of features which permit *N. pharaonis* to cope with with its extremely salty and alkaline environment. As protection, *N. pharaonis* may form a complex cell envelope consisting of different types of glycoproteins, probably glycosylated at Thr-rich tetrapeptide repeats. A large fraction of the extracellular proteins, being directly exposed to these extreme conditions, is predicted to be retained to the cell membrane by N- or C-terminal lipid anchors to prevent their alkaline extraction. As in other halophilic archaea, these are mainly secreted in the folded state through the Tat pathway as to avoid folding in a hostile and chaperone-free environment. *N. pharaonis* further encodes three potential alkaline proteases, which are probable secreted into the extracellular space by the Sec system.

Several key components of the energy metabolism are present in the cell membrane, in particular the ATP synthase and the respiratory chain complexes, of which a complex III analog remains to be detected in the genome. However, a functional respiratory chain was experimentally validated for *N. pharaonis*. Its determined ion specifity differs from that of other alkaliphiles using protons rather than sodium as coupling ion.

A versatile nitrogen metabolism and a large number of transport systems for nitrogenous compounds and heavy metals, which are scarce in the habitat, reflects metabolic adaptation to the alkaline environment.

Genome annotation showed that *N. pharaonis* possesses high biosynthetic capabilities, which was exploited to develop a simple synthetic growth medium. Halophilic archaea have a highly similar signal transduction cascade for chemo- and phototaxis but signal transduction and motility genes differ with respect to gene organization and domain architecture. Thus, high plasticity of environmental responses can be predicted reflecting variable halophilic environments.

3.4 Materials and methods

3.4.1 Genome sequencing and assembly

N. pharaonis type strain Gabara (DSM 2160) was sequenced with 5.8-fold sequence coverage using a shotgun clone library (average insert size of 2 kb), and assembled with the PHRED-PHRAP-CONSED package (Gordon et al. 1998). The coverage for the two plasmids differed from that of the chromosome (100.5 for PL23, and 3.9 for PL131). The sequence is of high quality (0.01 Errors/10 kb for the chromosome, 0.00 for PL23, and 0.04 for PL131). A copy of PL23 is integrated into the chromosome and this copy was found to carry itself

a 13 kb GC-poor insert (confirmed by clones bridging the 2.8 kb between the integration points and by a coverage typical for the chromosome). Apart from this integration, no other polymorphism between integrated and free plasmid could be detected.

3.4.2 Gene prediction and annotation

For gene prediction, REGANOR (McHardy et al. 2004) from the annotation package GENDB (Meyer et al. 2003) was used, which integrates results from CRITICA (Badger and Olsen 1999) and GLIMMER (Delcher et al. 1999). In addition, sixframe translation (>100 codons) was performed. From the resulting raw set of 11874 distinct ORFs, a set of 2843 validated genes was selected by the following procedure: Using BLAST (Altschul et al. 1997), predicted amino acid sequences were bidirectionally compared to the *H. salinarum* strain R1 ORF set (www.halolex.mpg.de) containing 1060 genes experimentally verified by proteomic analysis (Klein et al. 2004; Tebbe et al. 2004). This allowed identifications of undetected small genes, the discrimination between real proteins and spurious ORFs, the improvement of start codon selection, and the initial assignment of protein function. The ORF set was also analyzed by BLAST against itself, and against the NR database. Overlapping ORFs were adjusted based on gene context as well as characteristic halophilic pI and amino acid distribution patterns. tRNAs and other RNAs were predicted using tRNAscan (Lowe and Eddy 1997) and BLAST against *H. salinarum*, respectively. In the final validated gene set, 90.9% of the start codons were ATG, 7.9% GTG, and the residual 1.2% are pseudogenes, e.g. due to interruption by ISH elements.

The generated validated gene set was utilized to assess performance of the 3 gene predictors, REGANOR, CRITICA, and GLIMMER (trained by CRITICA ORF set for optimal results) (McHardy et al. 2004). For the chromosome, GLIMMER predicted 12.9% false positive (FP) and 2.4% false negative (FN) while CRITICA predicted no FP and 9.3% FN ORFs. The REGANOR gene finder performed best with 1.2% FP and 4.2% FN. One third of the GLIMMER starts were reassigned (31.3% of the genes were shortened, 6.5% extended) while 13.2% (13.9%) of the genes predicted by CRITICA (REGANOR) were shortened and 0.9% (1.1%) were extended.

Each *N. pharaonis* protein was assigned to a cluster of orthologous groups (COG) and to a functional category (Tatusov et al. 1997), with a minimal BLAST e-value of e^{-05}. BLAST results against *H. salinarum* and other databases were carefully evaluated for annotation of gene functions or descriptions. Intergene distances and configurations between gene pairs were analysed, and genes less than 35 bases apart were considered to be co-transcribed. Regions were defined as GC-poor when the GC content of 30 adjacent coding regions was 5% below the replicon average.

3.4.3 Motif searches

The PROSITE pattern PS00013 (Hulo et al. 2004) was used to search for lipid attachment sites (lipobox) within the first 50aa of predicted proteins of *N. pharaonis* and other prokaryotes. The lipobox is highly variable, resulting in a rather unspecific sequence motif, but with organism-specific preferences (e.g. LAGC in *E. coli* and *N. pharaonis*). We counted only those lipobox variations which occur frequently in a given species (i.e. at least 3 times) to reduce the number of false positives. Proteins with a predicted lipobox, which contain an additional twin-arginine motif and proteins identified through predictions from TATFIND2.2 (Rose et al. 2002) were used to determine the total number of twin-arginine translocation pathway (Tat) substrates. The number of general secretion pathway (Sec) substrates results from SIGNALP3.0 (Nielsen et al. 1997), excluding specific Tat substrates and proteins with TMHMM-based transmembrane helix predictions past the first 50 residues (Krogh et al. 2001). Proteins with signal peptide predictions were searched for the proposed preflagellin peptidase cleavage site in haloalkaliphilic archaea ('RGQ') (Thomas et al. 2001). N-/O-glycosylation sites were predicted using NetNGlyc and NetOGlyc tools (http://www.cbs.dtu.dk/services/). NG-dipeptide repeats/T-rich tetrapeptide repeats ('[VP][TE][ED]T') with 4/2 repeating units were detected by pattern searches within a sequence window of 30/50 aa.

3.4.4 ATP and pH measurements

N. pharaonis cells grown to late logarithmic phase in DSM medium 205 were harvested and resuspended in basal salt (medium without casamino acids) to result in an OD_{600} of 4. Cells were kept under a continuous nitrogen flow in a thermostated (20^{o}C) glass vessel. For illumination, a 100-W mercury lamp (HBO 100 W-2, Oriel, Stratford, CT) was used, fitted with a heat protection filter (Calflex 3000, Balzers, Lichtenstein), and a yellow cutoff filter (OG 515,Schott, Mainz, Germany) resulting in an irradiance of 32 mW/cm^2. Air was flushed through the medium for oxygenation. The pH traces were recorded with a standard glass electrode. For ATP determination by a luciferin/luciferase assay, 0.1 ml cells were lysed in 5 ml ice-cold buffer (10 mM $MgCl_2$, 0.02% NaN_3, 0.1 mM EDTA, 25 mM Hepes, pH 7.5). Luminescence was measured (Lumac Biocounter, Abimed, Germany) in triplicate by adding 0.1 ml of a 1:2 (w/w) mixture of D-Luciferin (0.1 mg/ml) and *Photinus pyralis* luciferase (0.2 mg/ml, Sigma-Aldrich) to 0.5 ml lysed cells. The protonophore CCCP (carbonyl cyanide m-chloro-phenylhydrazone) was added to a final concentration of 0.2 mM.

For measurement of the internal pH, freshly harvested cells were disrupted by sonification after washing twice with unbuffered basal salt without citrate (Koch and Oesterhelt 2005). When grown between pH 9.0 and 9.5, the internal pH was similar to that of the medium (external/internal: 9.0/9.3 and 9.5/9.2).

3.4.5 Synthetic medium

A synthetic medium for *N. pharaonis* (M. Engelhard, pers. comm.) was further simplified by omitting amino acids and vitamins. The minimal medium for *N. pharaonis* consists of: 20 mM sodium acetate, 10 mM sodium pyruvate, 12 mM NH_4Cl, 5 mM leucine, 3.4 M NaCl, 27 mM KCl, 175 mM Na_2CO_3, 1 mM $MgSO_4$, 2 mM Na_2HPO_4, 2 mM NaH_2PO_4, 5 µm $FeSO_4$, 4 µM $CuSO_4$, 4 µM $MnCl_2$, 3 µM $ZnSO_4$, 3 µM $CaCl_2$, pH 9.2.

3.5 Supplemental material

Supplemental Table 3.2: **Function categories (bold) and function classes** for the classification of *N. pharaonis* proteins.

FC abbreviation	*N. pharaonis* proteins	FC description
	2843	
MET	**475**	**metabolism**
EM	43	energy metabolism
CIM	45	central intermediary metabolism
AA	113	amino acid metabolism
COM	101	coenzyme metabolism
NUM	53	nucleotide metabolism
LIP	87	lipid metabolism
CHM	33	carbohydrate metabolism
TP_CP	**264**	**transport and cellular processes**
MOT	12	motility
SEC	13	protein secretion
TP	199	small molecule transport
CE	13	cell envelope
CP	27	cellular processes
ENV	**155**	**environmental information processing**
SIG	91	signal transduction
REG	64	gene regulation
GIP	**203**	**genetic information processing**
TL	98	translation
TC	25	transcription
RRR	51	replication, repair, recombination
RMT	18	RNA maturation
CHP	11	chaperones
MIS	**553**	**miscellaneous**
ISH	50	ISH-encoded transposases
MIS	158	miscellaneous
GEN	345	general function
UNASS	**1193**	**unassigned**
CHY	661	conserved hypothetical protein
HY	532	hypothetical protein

3.5.1 Transposases, plasmids, and regions with reduced GC content

Transposases. Besides the 35 copies of the IS1341-type transposase spread throughout the genome (6 of these on PL131), *Natronomonas* has only a few additional transposases. Transposases from four *Halobacterium*-specific insertion elements (ISH3, ISH4, ISH6, ISH9) were identified on PL131 of *N. pharaonis* and in GC-poor region I, but only ISH4 occurs more than once. There are three IS200-type transposases (NP3910A, NP4630A, NP4812A), each encoded next to a IS1341-type transposase but with variable relative gene orientation (as described for other species (Mahillon and Chandler 1998)). Each of the IS200/IS1341 gene pairs has an orthologous gene pair in *Halobacterium* with identical gene orientation and high sequence conservation.

Plasmids. Plasmid PL131 (131 kb) has a GC content of 57% which is lower than that of the chromosome (63%) (Table 1). It contains genes for proteins involved in transport, signal transduction, and regulation, but only few enzymes. The majority of the *H. salinarum* proteins with high sequence similiarity to PL131-encoded proteins are themselves encoded on the *Halobacterium* plasmids. PL131 shows signs of DNA rearrangements as indicated by the presence of several complete or disrupted transposase genes and by other disrupted or truncated genes. The sequence coverage for PL131 was below that of the chromosome which may indicate a lower copy number or plasmid elimination from a subpopulation of *Natronomonas*.

Plasmid PL23 (23 kb) occurs in free form with a high copy number (sequence coverage was 17 times that of the chromosome) (Table 1). PL23 has a remarkably compact and ordered organization of genes arranged into a few long transcription units. PL23 lacks transposases but encodes a probable recombinase (NP7062A) which is a member of the phage integrase family (see PFAM:PF00589). Only a few of the PL23-encoded genes show similarity to genes from other species. A copy of PL23 has been integrated into the chromosome next to the valine tRNA gene with the plasmid-encoded probable recombinase adjacent to the integration point. The last 25 bp of the integrated plasmid and of the tRNA form an imperfect direct repeat. Notably, the integrated copy of PL23 contains a 13 kb insert with reduced GC content (GC-poor region II) which is not present in the free plasmid.

Regions with reduced GC content. Four main regions with reduced GC content were identified in the chromosome (Figure 3.1 in Chapter 3.2.1). The 14 kb GC-poor region I reveals a gene cluster involved in sugar metabolism. GC-poor region II (54% GC) corresponds to the 13 kb insert in the integrated copy of PL23 (60% GC), and contains mainly hypothetical proteins and a disrupted endonuclease (NP3256A) similar to endonuclease HNH from *Natrialba* virus PhiCh1. The 21 kb GC-poor region III (GC content

partly below 50%) contains 39 genes including a number of remarkably long proteins (up to 2000 aa) some of them containing helicase domains. GC-poor region IV (25 kb) contains a gene cluster presumably involved in cell envelope formation (NP4624A to NP4616A) (Figure 3B) which is followed by a subtilisin-like serine protease (NP4628A).

3.5.2 Physiological capabilities

Central metabolism. *N. pharaonis* is an aerobic chemoorganotroph that can be grown on complex or synthetic media with amino acids as a carbon source. Several intermediary carbonic acids such as pyruvate, succinate and propionate promote growth (Soliman and Truper 1982). Consistent with these observations, all genes encoding enzymes required for the citric acid cycle and pyruvate utilization as well as extracellular proteases and probable amino acid transporters are present in the genome. For oxidative decarboxylation of 2-oxoglutarate and pyruvate, 2-oxoacid-ferredoxin oxidoreductases are likely used as experimentally shown for *H. salinarum* (Kerscher and Oesterhelt 1981). All gluconeogenic reactions enabling sugar biosynthesis from different carbon sources are present, but glycolytic pathways such as the Embden-Meyerhof pathway and semi-phosphorylated Entner-Doudoroff pathways found in sugar-utilizing halophiles are likely not present as key enzymes are absent.

A large group of genes involved in the fatty-acid degradation pathway was identified, amongst them a gene cluster encoding all beta-oxidation pathway enzymes and genes encoding enzymes involved in the degradation of uneven-numbered fatty acids. These enzymes require biotin and coenzyme B_{12}, respectively, which are synthesized in *N. pharaonis* by enzymes encoded in the biotin and cobalamine operons. As expected from the prenyl-based lipid composition, the fatty acid biosynthetic pathway is absent.

N. pharaonis has been described as a strict aerobic species and the genome contains components of the electron transport chain, a typical archaeal ATP synthase complex, and enzymes providing protection against oxidative damage such as catalase/peroxidase and superoxide dismutase. However, a *N. pharaonis* gene cluster was identified containing genes (NP4962A, NP4960A) homologous to predicted formate dehydrogenase subunits from *H. marismortui* (rrnAC1332, rrnAC1333), and genes (NP4946A, NP4944A) with 26-32% similarity to *H. salinarum* DMSO reductase subunits (OE2223F, OE2225F) (Muller and DasSarma 2005). In contrast to *H. mediterranei* and *H. marismortui*, only a gene encoding assimilatory nitrate reductase (NP4226A), but no respiratory nitrate reductase homolog was found.

Biosynthetic capabilities. *Natronomonas* contains several gene clusters involved in multistep pathways leading to the synthesis of arginine, lysine and branched-chain amino acids. Thus, in contrast to *H. salinarum* which lacks these gene clusters and whose growth is dependent on exogenous sources of these amino acids, *N. pharaonis* should exhibit a greater degree of nutritional self-sufficiency, as does *Haloarcula hispanica* (Hochuli et al. 1999). Genes for the complete synthesis of nicotinate, folate, thiamine, biotin, molybdopterin, cobalamine, hemes, and menaquinones are also present, sometimes clustered, and their presence confirms the observation that *N. pharaonis* is not dependent on exogenous vitamins for growth (Soliman and Truper 1982). Based on these findings, we have developed a very simple synthetic growth medium containing acetate and pyruvate as the sole carbon sources. Starting from a rich synthetic medium (M. Engelhardt, personal communication), we omitted those amino acids and vitamins for which we identified biosynthetic pathways in our genome analysis. Consistent with this, *N. pharaonis* is able to grow without external amino acids except leucine. Although all genes required for leucine synthesis are present, the 2-isopropylmalate synthase might not be fully functional, and leucine synthesis subsequently impaired. The 2-isopropylmalate synthase gene aligns to orthologous genes of several other species in its 5'-region but the upstream sequence does not contain a valid start codon (ATG or GTG).

Interestingly, a gene fusion *purNH* occurs in *N. pharaonis* (NP1662A) and *H. salinarum* (OE1620R) encoding enzymes potentially catalyzing steps 3 and 9 of purine biosynthesis from PRPP (phosphoribosyl-pyrophosphate) to IMP (inosine monophosphate). In contrast, most other organisms encode bifunctional proteins (*purH*) that perform biosynthesis steps 9 and 10. The IMP cyclohydrolase domain (step 10) is missing in some archaea, and is replaced by a non-orthologous archaeal type of this gene (*purO*, NP0732A) (Graupner et al. 2002).

Membrane lipids of archaea consist of glycerol diether lipids with prenyl side chains in contrast to bacterial diacylglycerol esters. Halophiles further produce many other isoprenoid-derived compounds such as squalenes, phytoenes, menaquinones, dolichol, and carotenoids (Oesterhelt 1976). Thus, the mevalonate pathway for the *de novo* synthesis of isoprenoid precursors is of major importance in archaea. In the *N. pharaonis* genome, the complete gene equipment for this pathway was found including two non-orthologous genes (*idiA* (NP4826A)*, idiB_1* (NP0360A)*, idiB_2* (NP5124A)) encoding probable IPP isomerases. Mevalonate phosphokinase is probably replaced by a predicted kinase (NP2852A) found adjacent to the mevalonate kinase gene (*mvk* (NP2850A)) in halophiles and other archaea. Several E- and Z-prenyltransferases involved in the synthesis of membrane lipids, quinone side chains and dolichols from isoprenoid precursors are present. Dolichyl-PP and dolichyl-P might be used as carriers of sugar moieties for the synthesis of glycoproteins as described for *H. salinarum* (Sumper 1987). Identified carotene biosynthesis genes are required not only

for retinal synthesis (15,15'-dioxygenase homolog absent), but also for production of bacterioruberin pigments which protect the organism against high irradiance.
Seven sugar nucleotidyltransferase homologs are present in *Natronomonas*, six of them distributed in two sugar metabolism gene clusters. These two regions contain further homologs similar to nucleotide sugar metabolism enzymes and polysaccharide transport proteins, and are likely involved in the biosynthesis of polysaccharide-containing cell wall components using nucleotide sugars as precursors. The two gene clusters were found within or adjacent to GC-poor regions indicating that they were acquired by horizontal gene transfer.

Transport. A total of 31 gene clusters encoding ABC-type transport systems were identified in *N. pharaonis*, which show sequence similarity to amino acid and peptide transporters as well as to systems involved in the transport of sulphate, phosphate, and many metal ions. Metal transporters e.g. for iron, manganese, copper, and cobalt uptake are especially important for *Natronomonas*, since these are scarce in a highly alkaline environment. However, specific substrates of the ABC transport systems usually cannot be reliably predicted by sequence homology. However, additional evidence, e.g. gene neighbourhood analysis, permitted more specific assignments as in the case of the urea gene cluster which includes urea transporter and urease genes.
Analysis of the genome also revealed an operon consisting of nine genes *phaEFGB1B2CD1D2D3* (NP0840A – NP0826A) as well as *phaA* (NP5056A) located elsewhere, which are homologous to the genes for the subunits of the pH adaptative potassium efflux system described for *Rhizobium meliloti*. This transport system might be involved in pH adaptation of *N. pharaonis*. We could also identify similar systems in *H. salinarum* (OE1954R – OE1944R, no *phaA* homolog) and *H. marismortui* (rrnAC3537 - rrnAC3529, pNG7037) with conserved gene order and 40-50% protein sequence identity.

3.5.3 Motility and signal transduction cluster

The proposed signal transduction cascade of *Natronomonas* and *Haloarcula* is very similar to that described for *Halobacterium* (Rudolph and Oesterhelt 1996), and consists of signal receptors/transducers, the two-component regulatory system (*cheA/Y* (NP2172A, NP2102A)), an adaptation module (*cheR/B* (NP2170A, NP2174A)), and the flagella with its motor. However, there are many differences in the organization of the genes in the genome and the domain architecture of individual components. In all three species, most genes coding for flagellins (*flg*), flagella-associated proteins (*fla*), and chemotaxis-related proteins (*che*), are clustered within one genomic region, the motility and signal transduction (MO-ST) cluster, but major alterations in the organization of the genes are evident (Figure 3.5 in Chapter 3.2.8). Apart from the highly conserved CheY (NP2102A), FlaH (NP2156A), and FlaI

(NP2158A) proteins (64%-75% sequence identity), orthologous MO-ST cluster proteins in these species reveal only 40-50% sequence identity. Four as yet uncharacterized proteins (NP2100A, NP2162A, NP2166A, NP2168A) were found to be conserved in all 3 MO-ST clusters and thus may be involved in signal transduction or motility. Interestingly, there are several gene and domain duplications (*flaG, cheC*, flagellins) and genes with common domains (*cheY/ cheB, flaD/ flaCE*). In addition, the N-terminal region of the mature flagellins reveals similarity to the N-terminal region of the FlaF (NP2094A) and FlaG (NP2096A, NP2098A) proteins, which lack a flagellin-specific signal sequence.

The *Natronomonas* genome is accessible through HaloLex (www.halolex.mpg.de). It has been submitted to the EMBL database under the accession numbers CR936257 (chromosome), CR936258 (plasmid PL131), and CR936259 (plasmid PL23). Supplementary material is available online at www.genome.org.

Acknowledgements

We acknowledge Hermann Lederer, Markus Rampp, Reinhard Tisma for their help with the MIGENAS system (www.migenas.mpg.de); Burkhard Linke and Folker Meyer for successful cooperation regarding GenDB (Meyer et al. 2003); Günter Raddatz and Stephan Schuster for their expert advise on using genome assembly and annotation tools; Douglas Griffith and Martin Engelhard for careful manuscript reading and critical discussions; and Jan Wolfertz for continuous maintenance and improvements of the HaloLex system. We further would like to thank the reviewers for valuable advice on the manuscript.

CHAPTER 4

Characterisation of Halophilic Secretomes

As in *Halobacterium*, most of the secreted proteins that were predicted for several other haloarchaea exhibit a twin-arginine (Tat) motif required for protein translocation in the folded state. However, the extent to which haloarchaeal secretomes are anchored to the membrane after protein export has not been studied yet. A quarter of the halophilic secreted proteins but only around ten percent of non-halophilic secretomes exhibit a lipobox motif indicating processing by a signal peptidase II and linkage to an N-terminal lipid anchor (NLip) via a cysteine residue of the motif. In *N. pharaonis*, lipobox-containing proteins were found to be even more frequent, and lipid anchoring might prevent protein extraction under alkaline conditions.

Several proteins resemble the cell surface glycoprotein (Csg) of *Halobacterium salinarum* in the C-terminal region suggesting that these are retained at the membrane by a C-terminal lipid anchor (CLip) as described for the Csg. NLip and CLip proteins commonly exhibit peptide repeat patterns, which are located adjacent to the anchor region. These peptide repeats stretches often contain potential O-glycosylation sites and might enhance flexibility of anchored proteins. A subclass of predicted membrane proteins has a single hydrophobic domain at the extreme C-terminus. Here, it is postulated that this domain represents a C-terminal protein anchor (CAnch), which is found to similar extent in halophiles and non-halophiles.

Both, N- and C-terminal membrane anchoring modules occur in cell-envelope forming proteins but also in secreted proteins that are involved in interactions with integral membrane complexes. This suggests that protein interactions, e.g. of halocyanins with respiratory complexes and substrate-binding proteins with ABC transporters and transducers, are facilitated by membrane retention.

4.1 Introduction

Protein translocation from the cytoplasm to the extracellular space or to other cellular locations occurs in cells of all three domains of life. Proteins destined for secretion are translocated via the general secretory (Sec) pathway in which they are co-translationally (eukarya) or post-translationally (bacteria) targeted by the signal recognition particle and

subsequently exported by a membrane-spanning pore (Bolhuis 2002). Although components of the Sec system are present in all domains of life, composition of its targeting and translocation machinery differs significantly. Archaea lack a number of components required in eukarya and bacteria, thus, novel components and probably also distinct mechanisms in Sec-mediated protein transport are indicated in archaea (Bolhuis 2002; Pohlschroder et al. 2004).

While the Sec system enables the transport of yet unfolded proteins, a second specific translocation pathway, the twin-arginine (Tat) pathway, mediates the export of folded proteins such as cofactor-containing proteins which are required to be preassembled prior to their translocation across the membrane. Components of the Tat system as well as Tat substrates (characterized by a twin-arginine motif) were found in complete genomes from all domains of life but the composition of the Tat machinery was specific to phylogenetic taxa (Dilks et al. 2003).

Due to their dual membrane envelope, Gram-negative bacteria require further secretion systems in addition to the described translocase systems in order to transport proteins across their outer membrane (Figure 4.1). For this terminal protein secretion step, Gram-negative bacteria exploit the type II system, but type V systems (autotransporters, two-partner secretion systems) and the fimbrial usher porin are also known to translocate proteins across the outer membrane (Preston et al. 2005). Apart from two-step transport, secreted proteins can also cross both membranes in a single unified process (type I, III, IV systems). While type I systems comprise a subfamily of ABC transporters which are involved in protein secretion, type III and type IV systems are contact-dependent systems that involve the formation of pili to transport proteins from one cell to another. The type III secretion system is a specialised export system, which is found in Gram-negative pathogens such as *Pseudomonas aeruginosa* and which plays a central role in their host interactions. The VirB system of *Agrobacterium tumefaciens* represents the best-studied example of a type IV secretion system known to deliver proteins and protein-DNA complexes to eukaryotic cells. The pilus formed during type IV secretion should not be confused with the 'type IV pilus' whose pilins are translocated by a specialised apparatus (Com in *Bacillus subtilis*) homologous to the type II secretion system. For the transport of archaeal flagellins, involvement of a type II-like transport system (FlaIJ in *Halobacterium salinarum*) was suggested since *flaIJ* mutants were not able to form flagella (Patenge et al. 2001). In general it should be noted, that subsets of different secretion systems share common structural features. For example, members of the PulD family are involved in type II and type III secretion as well as in phage assembly (Preston et al. 2005). The spectrum of known protein transport systems is completed by mechanosensitive channels (MscL), Holins originating from bacteriophages, and vesicle-mediated transport, e.g. of the cytotoxin ClyAc.

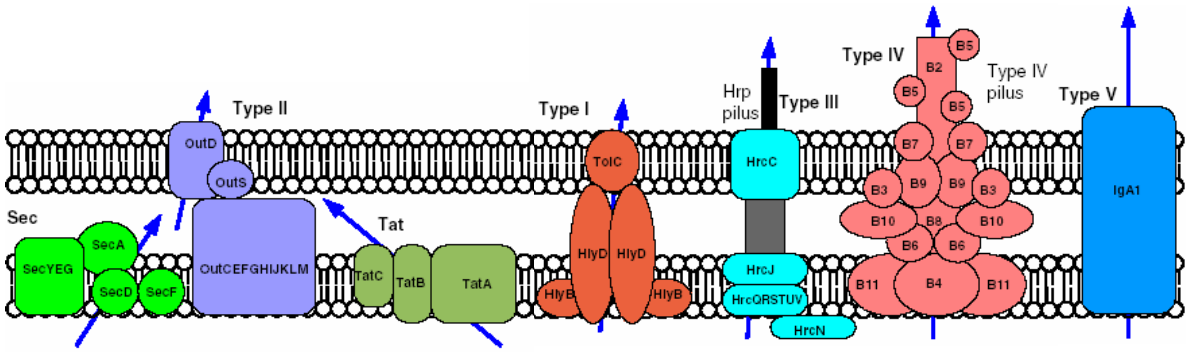


Figure 4.1. **Protein secretion systems in Gram-negative bacteria** (from Preston et al. 2005). The core components of each of the main systems are illustrated schematically.

The collective activity of a variety of secreted proteins has been associated with pathogenesis, and well-studied Gram-negative pathogens were found to exploit different secretion systems (e.g. type III secretion by *Pseudomonas* and type IV secretion by *Agrobacterium*) (Preston et al. 2005). In Gram-positive bacteria, the virulence factor ESAT-6 of *Mycobacterium tuberculosis* has been proposed to be secreted via a novel secretion system that contains YukA (Pallen et al. 2003). Due to these findings, the field of secretome analysis (or profiling) is currently emerging where the entire complement of secreted proteins in an organism is studied in order to gain a broader understanding of the mechanisms in bacteria-host interactions and pathogenesis.

Proteins destined for secretion contain targeting signals, mostly in form of N-terminal signal peptides, which are recognized by components of their protein translocation systems (Figure 4.2). Signal peptides of the Sec pathway are characterized by an N-terminal part (N-domain) with positively charged residues, which is followed by a hydrophobic stretch (H-domain). Tat-type signal peptides are similarly structured but show a more specific N-domain than Sec-type signal peptides, since it contains a specific twin-arginine motif surrounded by hydrophobic residues. The H-domain of Tat-type signal peptides is furthermore characterized by less hydrophobic residues compared to Sec-type signal peptides. While a selection of general prediction tools for Sec- and Tat-type signal peptides exists (listed in Preston et al. 2005), attempts to predict type I to type V signals are not successful yet, since the respective signals are less frequent and seem to be organism-specific. Targeting signals might also depend on the secondary and tertiary structure of folded proteins, in particular the signals for type II systems in Gram-negative bacteria. Due to these limitations, substrates for less common secretion systems are identified by homology to known substrates, gene neighbourhood with their transport system as well as by promoter analysis on the basis of common regulation factors, e.g. type III-specific sigma factor HrpL (Preston et al. 2005). However, homology-based predictions can be misleading since orthologs might be secreted by different translocase systems depending on the organism.

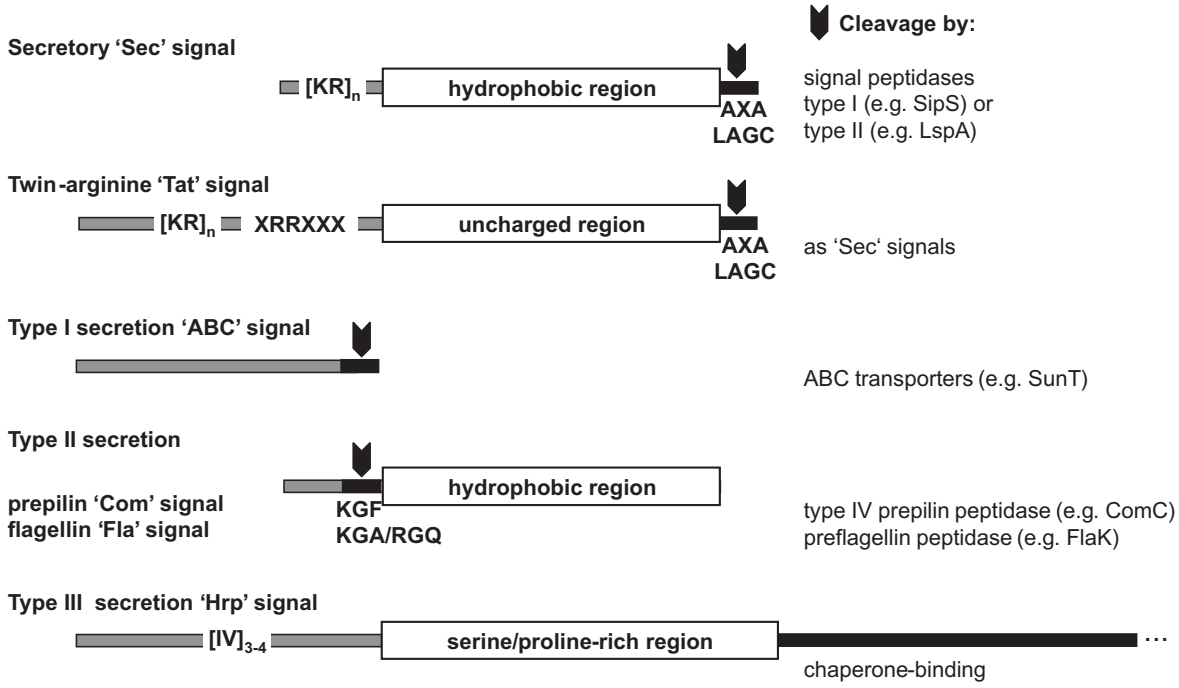


Figure 4.2: **Features of N-terminal signal peptides.** Signal peptides of the general secretion (Sec) pathway consist of an N-terminal domain (N-domain, grey) containing positively charged residues ($[KR]_n$, n = 0–5) and a hydrophobic domain (H-domain, white). Signal peptidases process the secreted proteins at their cleavage site (C-domain, black with arrow). Tat-type signal peptides are similar to Sec-type signal peptides, but show a specific twin-arginine motif within the N-domain and a less hydrophobic H-domain. Signal peptides of other secretion systems are highly diverse and not well characterized yet (Preston et al. 2005; Rose et al. 2002; Tjalsma et al 2000).

Chitinases, for example, reveal a twin-arginine motif in *Halobacterium* but not in *Thermococcaceae* (Rose et al. 2002). Motif and domain searches, e.g. for specific cleavage sites or autotransporter domains, result in more accurate signal predictions.

Proteins with N-terminal signal peptides are usually processed by signal peptidases following their export across the membrane. In case of cleavage by signal peptidase I that recognizes relatively unspecific cleavage sites (Figure 4.2), the secreted proteins are released in the extracytoplasmic (ExCyt) space. Signal peptidase II specifically processes secreted proteins that contain a lipobox motif (LAG↓C) in front of the cysteine residue. These proteins are subsequently retained at the membrane as an N-terminally anchored lipoprotein (NLip) through linkage between the cysteine residue and a lipid chain (thioester linkage in bacteria, thioether linkage in archaea) (Hayashi and Wu 1990; Mattar et al. 1994). Whereas homologs of signal peptidase I are present in archaeal genomes, signal peptidase II genes are yet to be identified. However, action of signal peptidase II has been described for a halocyanin of *Natronomonas pharaonis* where a prenyl-based lipid was attached to the processed protein (Mattar et al. 1994). Prepilin/flagellin signal peptides are recognized by type IV prepilin and preflagellin peptidases where cleavage occurs in between the N- and H-domain at the cytoplasmic site of membranes in contrast to signal peptidases of type I and II that remove complete signal sequences post protein translocation (Tjalsma et al. 2000; Thomas et al.

2001). Thus, processed pilins/flagellins are retained within the membrane by an N-terminal hydrophobic peptide stretch. Membrane attachment of secreted proteins may also occur via the C-terminus of protein chains. Cleavage at a specific C-terminal cell wall sorting signal (LPXT↓G) and simultaneous attachment to peptidoglycan by a membrane-bound transpeptidase (sortase) has been reported for *Staphylococcus aureus* and also suggested for *B. subtilis* (Pallen et al. 2003; Tjalsma et al. 2000). Furthermore, cell surface glycoproteins (Csg) of several halophiles have been shown to be attached to the membrane by a C-terminal lipid anchor (CLip) (Figure 4.3) (Kikuchi et al. 1999; Konrad and Eichler 2002). However, the mechanism of C-terminal lipid attachment in halophiles as well as the C-terminal lipobox pattern (C-lipobox) are yet unknown.

Although the utilization of different protein translocation systems by *H. salinarum* strain NRC-1 has been investigated previously (Bolhuis 2002; Rose et al. 2002), the extent of membrane anchoring has not been investigated yet. Here, numbers of secreted proteins that

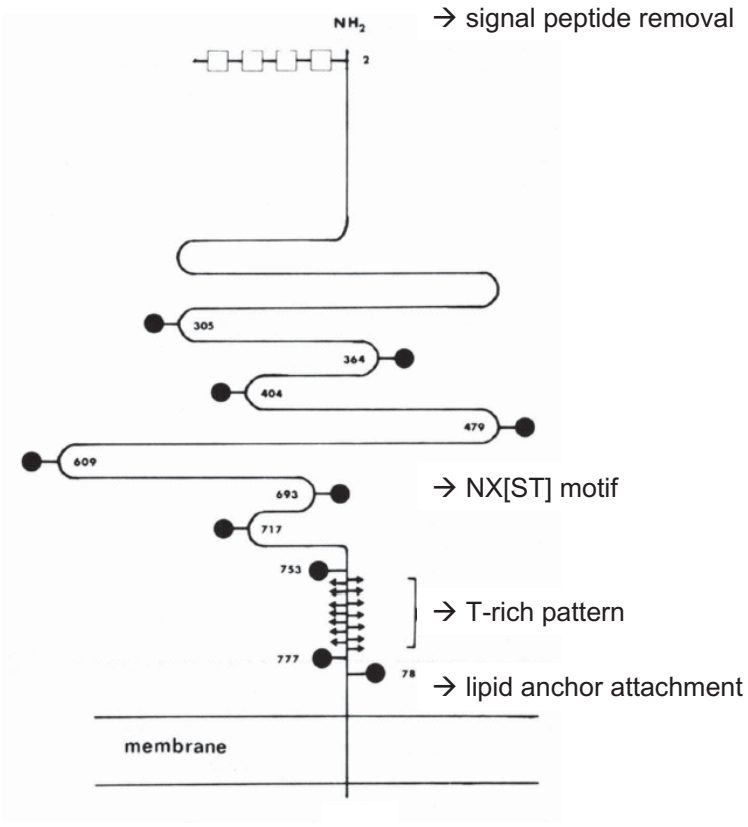


Figure 4.3: **Topological features of the *H. salinarum* cell surface glycoprotein (Csg)** (from Lechner and Sumper 1987). The halobacterial Csg precursor is characterized by a Sec-type N-terminal signal peptide, several N-glycosylation sites (circles), a C-terminal peptide repeat pattern containing O-glycosylation sites (arrows), and a hydrophobic stretch directly adjacent to the peptide repeat. Within this region of the C-terminal hydrophobic stretch, a lipid was shown to be attached to the Csg (Kikuchi et al. 1999). The *H. salinarum* Csg is furthermore characterized by a large repeated unit saccharide (squares). This and other N-linked sugar units are sulphated and introduce negative surface charges added to the acidic amino acid residues of the protein itself.

are likely attached to the membrane by N- or C-terminal anchors were estimated for halophilic and non-halophilic archaea and typical features of these proteins were described. Furthermore, available proteomics data were screened to validate suggested signal peptide processing in *H. salinarum* and *N. pharaonis*.

4.2 Secreted proteins

4.2.1 Utilization of Sec and Tat protein translocation pathways

All substrates exported through the Tat-, Sec-, and FlaIJ-mediated protein translocation pathways were predicted for 9 completely sequenced archaea as described in Methods. Probable substrates of the Sec system were found in all genomes but halophiles exhibit reduced numbers of Sec-type signal peptides compared to non-halophilic archaea (Table 4.1). Since the Sec-mediated protein translocation pathway is a generally applied pathway, Sec substrates comprise all types of protein functions. In haloarchaea, high fractions of proteins with yet unassigned functions (CHY, HY, and NOF classes, Supplemental Table 3.2) were observed amongst putative Sec substrates (*H. salinarum*: 59%, *N. pharaonis*: 66%).

The Tat system exports folded proteins which are required to be pre-assembled before protein translocation (Rose et al. 2002). A previous study over 84 bacterial and archaeal genomes identified proteins with Tat-type signal peptides in all except 11 organisms with numbers ranging from one to 145 substrates (Dilks et al. 2003). However, the Tat pathway was only widely used (no. of Tat substrates >2% of the theoretical proteome) for *Halobacterium* and the plant pathogen *Caulobacter* which exhibits many substrate-binding proteins amongst predicted Tat substrates. Here, high numbers of Tat substrates were found for all halophilic archaea comprising up to 4% of the theoretical proteome while non-halophilic archaea possess less than 1% proteins with Tat-type signal peptides. It becomes clear from this study that the Tat pathway is indeed predominantly used under halophilic conditions as suggested previously when *H. salinarum* str. NRC-1 was analysed (Rose et al. 2002).

In halophilic environments denaturation of proteins can be avoided by exposing acidic amino acids at the protein surface (Madigan et al. 2000). In the high salt cytoplasms of haloarchaea, proteins therefore reveal low isoelectric points, and this holds also true for their predicted extracellular proteins exhibited to the outer halophilic environment (average isoelectric points around 4.2 for *N. pharaonis* and *H. salinarum*). However, in order to avoid aggregation, secreted proteins are required to fold rapidly after protein translocation or need

Table 4.1: **Secreted proteins in halophilic and non-halophilic archaea** (separated by a horizontal line). Most signal-peptide-containing proteins either pass the Tat- or Sec-mediated translocation system, followed by N-terminal cleavage through signal peptidase I or a yet unassigned peptidase of type II. The FlaIJ system presumably transports flagellins and possibly further substrates, which are cleaved by a membrane-bound preflagellin peptidase. The predicted number of proteins exported via the respective protein translocation pathway is given followed by the percentage compared to the complete theoretical proteome.

	Tat substrates	**Sec substrates**	**Flg-like substrates**	Genes	Genome size [Mb]
Natronomonas pharaonis (NP)	111 (3.90%)	110 (3.87%)	12	2843	2.80
Halobacterium salinarum str. R1 (HS)	80 (2.84%)	111 (3.93%)	10	2821	2.72
Halobacterium salinarum str. NRC-1 (HN)	73 (2.78%)	113 (4.31%)	7	2622	2.61
Haloarcula marismortui (HM)	161 (3.80%)	203 (4.79%)	6	4240	4.37
Haloquadratum walsbyi (HQ)	69 (2.48%)	105 (3.78%)	1	2777	3.24
Methanosarcina mazei (MM)	10 (0.30%)	259 (7.68%)	3*	3371	4.15
Archaeoglobus fulgidus (AF)	17 (0.70%)	187 (7.78%)	2*	2420	2.21
Pyrococcus furiosus (PF)	6 (0.28%)	176 (8.28%)	2*	2125	1.95
Sulfolobus solfataricus (SS)	5 (0.17%)	179 (6.01%)	1*	2977	3.04

* - The number of flagellin precursors was retrieved from the published genome annotation.

to be exported already in the folded state via the Tat pathway. An analysis by Rose et al. indicated that *Halobacterium* frequently adopted twin-arginine motifs in proteins that were previously transported by the Sec system, e.g. in chitinases and halolysin, in order to enable protein translocation in the folded state (Rose et al. 2002). Thus, extracellular folding without the assistance of chaperones would have been circumvented. In future, it will be interesting to investigate, if not only haloarchaea widely utilize the Tat pathway for protein export, but also halophilic bacteria such as *Halobacillus halophilus*. Although these organisms exhibit normal salt levels within their cytoplasm, they probably also possess secreted proteins with acidic surfaces and Tat-type signal sequences as haloarchaea.

It should be noted that predicted numbers of Tat substrates slightly vary between the between the two highly related *Halobacterium* strains as well as in comparison to the genome-wide surveys that have been conducted previously for *H. salinarum* str. NRC-1 (Rose et al. 2002; Bolhuis 2002). These differences are due to varying gene start assignments (see Chapter 2) and result from the different applied prediction procedures for Tat motifs.

Typical Tat substrates were found in all analysed archaeal strains, amongst them cofactor-binding subunits of redox membrane complexes such as F_{420}-non-reducing hydrogenases, molybdopterin oxidoreductases as well as fumarate and nitrite reductases. In halophiles the sulphur- and copper-binding proteinaceous coenzymes thioredoxin and halocyanin which are also involved in redox reactions were found amongst predicted Tat substrates. Furthermore, *Halobacterium* uses the Tat-mediated protein translocation pathway for components of its

aerobic and anaerobic respiratory complexes, e.g. Rieske iron-sulphur protein, terminal oxidase assembly factor as well as subunit A of dimethylsulfoxide reductase (DmsA). The latter is known as a typical Tat substrate since it exhibits a twin-arginine motif across species (Dilks et al. 2003).

The number of redox components amongst Tat substrates usually exceeds the number of non-redox proteins such as alkaline phosphatase D and phospholipase C (Dilks et al. 2003). In haloarchaea, though, the high number of predicted Tat substrates is due to the export of many non-redox components via the Tat-mediated pathway. *Halobacterium* likely translocates several secretion enzymes (chitinases of the *chi* gene cluster, halolysin, alkaline phosphatase, a serine proteinase) via the Tat system. Other haloarchaea lack these enzymes (except alkaline phosphatase, see Supplemental Table 6.3), but like *H. salinarum* they contain substrate-binding proteins of ABC transporters that are probably exported through the Tat system. *H. salinarum* further encodes Tat-containing substrate-binding proteins (*basB, cosB, tmpC, potD*), which are shown or predicted to be involved in chemotaxis (Figure 3.6 in Chapter 3.2.8). The majority of predicted Tat-type signal peptides in haloarchaea further exhibits also an N-terminal lipid anchor (Chapter 4.3.1).

4.2.2 Proteins with flagellin-like cleavage sites

Signal peptides of bacterial pilins and archaeal flagellins are cleaved in between the N- and H-domain, thus, leaving a N-terminal hydrophobic peptide stretch that might function as a membrane retention signal as reported for prepilin-like signal peptides (Tjalsma et al. 2000). Cleavage sites for several *B. subtilis* Com proteins (KG↓F) and *Methanococcus voltae* flagellin B2 (KG↓A) have been experimentally validated (Tjalsma et al. 2000; Thomas et al. 2001), and the same putative cleavage sites were found in other archaeal flagellins (*Archaeoglobus fulgidus,* methanogens, Pyrococci*)*. However, putative cleavage sites differ for haloarchaeal (RG↓Q), *M. jannaschii* (RG↓A), and some *Pyroccus* (RG↓A) flagellin signal peptides.

Surprisingly, when searching for signal peptides with probable haloarchaeal RGQ cleavage sites, not only flagellin genes but also several other proteins with unknown functions were found in all 5 halophilic strains. As flagellins, all of these RGQ-motif proteins showed a hydrophobic stretch following the putative cleavage site (Supplemental Table 4.8). The RGQ-motif proteins with unknown functions are orthologous amongst different halophiles and their genes were found to be located adjacent to each other on the respective chromosomes (Figure 4.4). While one gene cluster was conserved between *H. salinarum*, *N. pharaonis*, and *Haloquadratum walsbyi*, and, a second less conserved gene cluster occurred only in *N. pharaonis* and *H. marismortui*. Interestingly, the RGQ-motif proteins of the *H. salinarum*

gene cluster reveal an unusual alkaline pI value in contrast to its orthologs in *H. walsbyi* and *N. pharaonis*.

The genes that encode proteins with flagellin-like signal peptides form putative transcription units with genes homologous to membrane proteins (COG2064) and ATPases (COG0630) involved in type II/IV secretion. FlaJ/FlaI also belong to these orthologous groups and were predicted to be involved in flagellin transport (Thomas et al. 2001). For *H. salinarum*, this is also indicated by experimental data showing that *flaIJ* genes are essential to form functional flagella (Patenge et al. 2001). It can be hypothesized that the putative membrane and ATPase components function as protein translocation systems and that the adjacent RGQ-motif genes are somehow involved in protein export as further subunits of these systems or as substrates. In case flagellins are cleaved at the RGQ motif as suggested, clustered RGQ-motif proteins might also be processed by membrane-bound preflagellin peptidase and retained within the membrane. This suggestion is emphasized by the finding that *H. walsbyi* neither forms flagella, nor contains flagellin genes, but it encodes a preflagellin peptidase homolog and clustered RGQ-motif proteins. Due to this observation it is unlikely, though, that RGQ-motif proteins in other motile haloarchaea are part of the flagella or motor machinery.

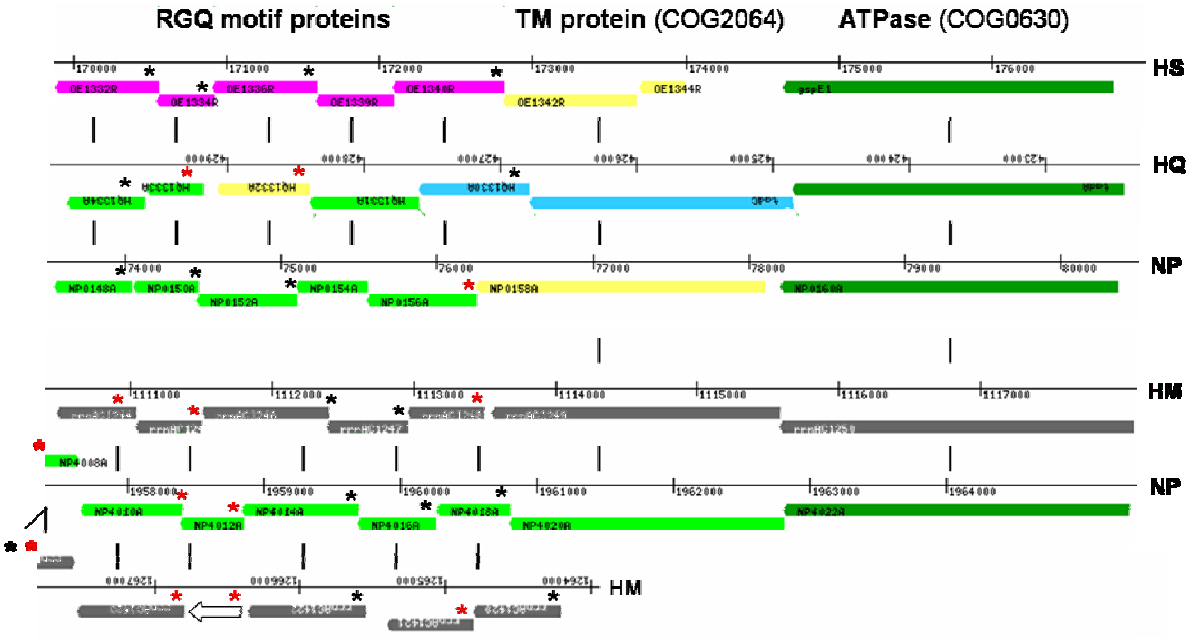


Figure 4.4: **Gene clusters encoding proteins with flagellin-like cleavage sites.** Haloarchaeal flagellins exhibit a predicted RGQ cleavage site which is also found in signal peptides of other proteins with unknown functions (black asterisk, variants of the motif are marked by a red asterisk). The clustered genes are orthologous amongst each other (vertical lines) and are located adjacent to a putative membrane protein and ATPase and a similar to FlaI/FlaJ, and components of typeII/IV secretion systems. Most of the encoded proteins have acidic pI values (green) common for haloarchaeal proteins but also neutral (blue) and alkaline (yellow) proteins are found in the gene clusters. Extremely alkaline proteins with pI values above 10 (pink) were found for the *H. salinarum* RGQ-motif proteins. Some RGQ motifs were only detected when shifting gene starts. The white arrow marks a yet unassigned gene found by tblastn. Species abbreviations are given in Table 4.1.

4.3 Membrane-anchored proteins

4.3.1 Lipobox-containing proteins

N-terminally anchored lipoproteins (NLip) were predicted using a newly developed program, for which rules of the PROSITE entry (PS00013, PDOC00013) for the 'prokaryotic membrane lipoprotein lipid attachment site' (lipobox) were adapted. The relatively unspecific PROSITE pattern allows amino acid variations at all lipobox motif positions except for the cysteine residue, where the lipid anchor is attached to after signal peptidase cleavage. Statistics of lipobox search results over several genomes gave actual lipobox motif variations, though, which occur within the archaeal domain of life (Supplemental Table 4.9). LAGC was identified to be the most frequent lipobox variation in halophilic archaea and *A. fulgidus* and was already validated for the blue copper protein halocyanin of *N. pharaonis* (Mattar et al. 1994). However, the LAGC motif does not occur frequently in other non-halophilic archaea. Whereas putative halophilic NLip proteins mostly contained XAGC variations (X = [LVTAFIS]) or an LSGC motif, non-halophilic NLip proteins frequently revealed XSGC variations (X = LVAFIG) of the lipobox. Due to these species-specific lipobox motif variations the generic PROSITE pattern could not be modified to reduce the number of false positive NLip proteins. Thus, only the frequent lipobox variations for the respective species were considered, and proteins with non-frequent lipobox variations were excluded as false positives. The crenarchaeon *Sulfolobus solfataricus* did not reveal any frequent lipobox motif, and it is likely that this species does not exhibit proteins with N-terminal lipid anchors at all. All halophilic species exhibit at least a third more predicted NLip proteins as non-halophilic archaea (Table 4.2). However, there are also deviations in respect to lipoprotein numbers

Table 4.2: **Proteins with a predicted N-terminal lipid anchor.** Signal peptides are specifically cleaved by a yet unknown signal peptidase II at the lipobox, and are then modified by a lipid-anchor. For species names abbreviated by a 2-letter code and further explanations see legend of Table 4.1.

	Lipoproteins	Tat : Sec Ratio	Tat-exported lipoproteins	Sec-exported lipoproteins	Lipoproteins with unknown export pathway
NP	91 (3.18%)	28 : 1	85	3	3
HS	49 (1.74%)	11 : 1	42	4	3
HN	49 (1.87%)	8 : 1	39	5	5
HM	116 (2.74%)	10 : 1	105	11	0
HW	49 (1.76%)	23 : 1	46	2	1
MM	23 (0.685)	1 : 4	4	15	4
AF	16 (0.66%)	1 : 1	8	6	2
PF	9 (0.42%)	1 : 2	3	5	1
SS	0 (0.00%)	0 : 0	0	0	0

within the halobacterial branch of archaea, with *N. pharaonis* and *Haloarcula marismortui* revealing most NLip proteins. For *N. pharaonis* it was speculated that its lipoproteins might prevent protein extraction under alkaline conditions (Chapter 3.2.6).

Most halophilic signal peptides of predicted NLip proteins contain not only a lipobox but also a twin-arginine motif indicating transport of NLip proteins via the Tat-mediated protein translocation pathway (Table 4.2). In *H. marismortui*, *H. walsbyi,* and *N. pharaonis* NLip proteins comprise 72% to 81% of all predicted Tat substrates. The fraction of lipoproteins amongst *Halobacterium* Tat substrates is slightly lower due to its additional extracellular enzymes that are absent in other halophiles. In contrast to haloarchaea, where the translocation of NLip proteins by the Tat system is prominent, there is no bias for either the Tat or the Sec translocation system in non-halophilic archaea. It is unclear, why Tat and lipobox motifs are more frequently combined in halophiles. Proper protein folding is probably a prerequisite for post-translational modifications, and secretion of already folded proteins likely facilitates the attachment of lipids and saccharides that further protect halophilic proteins in the hostile environment.

Function classes for the predicted set of NLip proteins in *H. salinarum, N. pharaonis* and *H. walsbyi* were analysed (Figure 4.5). All 3 species exhibit many transporter (TP) subunits with probable N-terminal lipid anchors, most of them are substrate-binding proteins of ABC transport systems. Enzymes (MET) such as a serine protease, glucose dehydrogenase, and L-aspartate oxidase as well as proteins involved in energy metabolism (halocyanin, nitrite

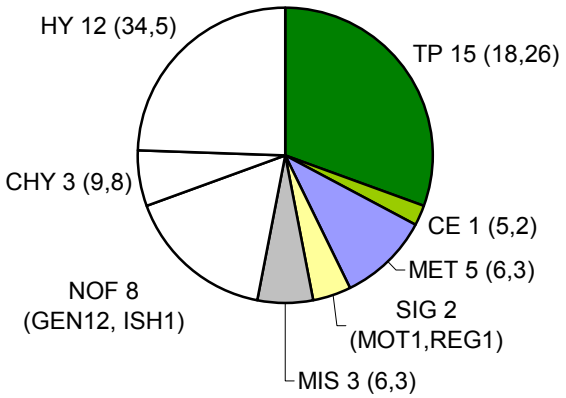


Figure 4.5: **Function class distribution of halophilic N-terminal lipoproteins.** *H. salinarum* (*N. pharaonis*, *H. walsbyi*) NLip proteins were categorized into function classes. Most lipoproteins are subunits of transporters (TP, dark green), followed by NLip proteins involved in metabolism (MET, blue) such as halocyanin and sugar metabolism enzymes. Several NLip proteins probably take part in cell envelope formation (CE, light green). However, for half of the predicted lipoproteins no specific functions could be assigned (white). The complete function class abbreviations are listed in Supplemental Table 3.2.

reductase, ETF dehydrogenase) were also found among NLip proteins. Furthermore, three of the four known *Halobacterium* substrate-binding proteins (*basB, cosB, potD, tmpC*) participating in signal transduction (SIG, MIS) reveal a Tat motif combined with a lipobox. In *N. pharaonis*, several probable glycoproteins involved in cell envelope formation (CE) are likely anchored by an N-terminally lipid. However, 30% of *H. walsbyi* and even 60% of the *N. pharaonis* NLip proteins have no assigned function yet.

Several predicted archaeal NLip proteins commonly show peptide repeat patterns following the cysteine residue that functions as the lipid anchor attachment site (Table 4.3). The observed peptide repeats differ in length (single, dipeptide, tripeptide, and tetrapeptide repeats) and number of repeating units (e.g. up to 12 times for Asn-Gly dipeptides). Only a limited set of amino acids is found within the peptide repeats; glycine, asparagine, and proline as well as amino acids with acidic (Asp, Glu) or alcohol (Ser, Thr) functions. Most of the threonine and some serine residues within or adjacent to peptide repeat motifs were predicted to function as glycosylation sites for O-linked saccharide units (Table 4.3). Repeats of threonine residues have indeed been shown to function as O-glycosylation sites for Gal-Glc disaccharides in S-layer glycoproteins of *Halobacterium* and *Haloferax* where the peptide repeat stretch was found adjacent to a C-terminal lipid anchor (Sumper 1993). A large proportion of NLip proteins have potential O- as well as N-glycosylation sites elsewhere within the protein sequence emphasizing the fact that many of the secreted proteins in archaea might be further modified by glycosylation following their export and signal peptidase cleavage.

The amino acid composition of observed peptide repeats varies among the investigated species. For *Natronomonas,* 8 lipoproteins with Asn-Gly repeats were found, which did not occur in any of the other archaea. *Halobacterium* NLip proteins often showed Thr repeats while Thr-Ala dipeptide repeats are present in *H. marismortui*. Ser-Gly dipeptide repeats and Gly repeats were found in *Haloquadratum,* and varied versions of these repeat motifs were also observed in *Haloarcula* and *Halobacterium*. The observed peptide repeats that occur adjacent to lipobox motifs might be important for the function of lipoproteins or might at least support the function of NLip proteins. The peptide repeats were found in subunits of transport systems as well as in halocyanins, and it was suggested for the *Natronomonas* Asn-Gly repeat regions that they might form flexible hinges which promote protein-protein interactions in membrane complexes (Chapter 3.2.6). Variation of peptide repeats between different haloarchaeal organisms might also point out a probable function as extracellular signatures (in combination with the attached saccharides).

From the fact that peptide repeat types seem to be rather species-specific it can be further assumed that peptide repeat modules spread rapidly through haloarchaea by continuous addition/loss of repeating units of an adopted repeat. This scenario is illustrated in

Natronomonas, where Asn-Gly repeats with one to 12 repeating units are found and a recent repeat modification was observed in the halocyanin N-terminus (8 units in *N. pharaonis* strain Gabara (NP3954A) and 7 units in strain SP1 (HCY_NATPH)).

Table 4.3: **A selection of lipoproteins that contain peptide repeat patterns.** Signal peptides of NLip proteins contain twin-arginine and lipobox motifs (both in bold) which are often followed by peptide repeat stretches (4x1, 3x2, 2x3-repeats in blue). These peptide repeat regions exhibit frequently predicted glycosylation sites (underlined) or are part of probable glycoproteins (O - O-glycosylation, N - N-glycosylation). The arrow indicates the cleavage/lipid attachment site. The representative peptide repeats are found in *N. pharaonis* (NP), *H. salinarum* (OE), H*. marismortui* (pNG, rrnAC), and *H. walsbyi* (HQ) transporter subunits (TP, ABC), halocyanin-like proteins (HCY), a glucosidase (GLC), aspartic acid-rich protein (MIS), predicted lipoproteins (LIPO). One of the proposed cell envelope (CE) proteins of *N. pharaonis* also exhibits N-terminally a lipobox motif and peptide repeats. Several peptide-containing proteins have no assigned function (HY).

Id	Lipobox region ↓ Peptide repeat region	O	N	Annotation
NP0544A	…DTSNLD**RR**SLLKGVGVAGIAG**LAGC**TGNGDGNGNGNGNGNGNGNGNGNGDDGA		X	TP
NP1774A	…DPTLD**RR**SYLSAVGAAGLAAG**AAGC**VGLDDGNGDGNGDGNDDDSDDDSDDDTRTI	X	X	ABC
NP2860A	MQPS**RR**TFVKSGIGVVGLGA**LAGC**AEEDPDPGAEDEETTPDEEEPDEETPEPET	X		MIS
NP3142A	…TQDTWT**RR**RMLAVAGAGAVVG**LAGC**ADESSEEPEEEPEEAPEEDPEEDPEEEPEA	X		LIPO
NP4216A	MPS**RR**ELLRLGGATLAAASAAG**IAGC**S…PAEDPDEEDPDEDTADETETEPTPDEDE			CE
NP4692A	MAPP**RR**RQILSIGGLGLAAA**IAGC**GDTAPDDDTPESVDDDETDATPEPDDDDEP	X		HCY
NP5000A	…DGRFTG**RR**AFIVGAAATGVAA**FAGC**TDNGNGDGNGTGNGTGNGTGNGDGNGGADS	X	X	ABC
NP6022A	…PPTAS**RR**RMLAAAGTALATFS**LAGC**SDESDDDDPEETEPEEPEPEPEPEPEPEPE	X	X	LIPO
NP6104A	…PSDRVS**RR**SVLATTGAAAAFG**LAGC**LGGGGGDGLSGTIDASGSNTVAPITSWAGE		X	ABC
OE1361F	MPS**RR**DVLRLGAGVLAAG**TAGC**TDTAPNRVAAAETAATTTRETTTDSRSESP	X	X	ENZ
OE2317R	MADFD**RR**EFLKLAGGTVGASL**VAGC**SSGGGGGDTTKVGTVYGTGGLGDGSFNDQA	X	X	ABC
OE3641F	MTS**RR**RFVAAVGSATAASLG**LAGC**VGDRETTTATEETTTTTEGTTTTTGGTTAT	X		ABC
OE4563F	MA**RR**LLAVAVVCLVV**LAGC**QGGMSGDATTDPTTTAPTTEQMTDAAGTTT	X	X	HY
pNG7023	MTNSN**RR**KFLKATGVGLLGG**LAGC**TRGGDSSGGDGSDGSSDGSDGSGGSDGGTS	X	X	ABC
pNG7340	…VEDSID**RR**RLLQALGAGGAIA**IAGC**SGDGGSGSDGGDGDSESGGGGSDGSTQSVQ		X	TP
rrnAC0510	MQRPST**RR**QFLTGTGVAALAI**TAGC**VSSGSSSEPAAETGTGTETATETATPTATP	X	X	HCY
rrnAC0830	…TN**RR**AFLKRTGAVTTVGLLGG**LAGC**STEQTGGDGGDGGDGGDGGMTGTGTESGGD	X	X	HY
rrnAC0915	…SNKRLT**RR**NALRIAGAAGAAS**LAGC**GGSDGGDGSSGDGSSGDGGDSTSGDGSSGD	X		HY
rrnAC1987	…TNSIG**RR**SFIRAAGASAALGL**FAGC**SGDGGGGGENPDGGGGIAETVTIGHLAPLN			ABC
rrnAC3132	…SVN**RR**QLLKSTGVAGVAGLTG**LAGC**SGGDGGDGGGDGGDGGDGGGDGGDDYPSLG	X	X	ABC
rrnAC3204	MDRN**RR**QFLGALTAVVTGT**IAGC**SGGDSGESSPTATATATQVSTRTATATPTD	X		HCY
HQ1383A	MSS**RR**RFLQGIGAVSTVGL**VAGC**TGSTNNSNNNTDSVDQTTATTTSDSVPFGD	X	X	ABC
HQ1471A	…GDTQVS**RR**KFLLTSGAIGAAG**LAGC**SSSEGNSSEESDGGDSSDSSDSSDSSGGSE	X	X	ABC
HQ1619A	…DSLH**RR**DVLKAAGASTVGIAG**LAGC**AGGGGGGDSGDGGGESGSDDTSSDSSSEEY			ABC
HQ2192A	…EKTRTS**RR**KFLAASGSLSAAA**LAGC**SGGGGGGGGGSGSGSGSDGRTLRQGYLLPL	X		ABC
Mm:NP_633963	MKKMLKIMAMLF**AAGC**AEQGGEETVEEAAEESAPVEETASAEVNASG	X	X	ABC
Pf:NP_578837	MKRMIMYLS**VSGC**ISEQTQTQTLESNSPTQTTTTTSPQITVTF	X	X	GLC
Pf:NP_579137	MKKGLLAILLVG**GSGC**IGGGTQTQTQTPTETGSPTQTTTPSGVTQA	X	X	ABC
Af:NP_069723	…VYFFWHLAEDMKAKVLVFIAL**FAGC**AGEEKTPTTTQTTTTETASEKKLSAAFVYV	X		HY

4.3.2 Proteins with a C-terminal membrane anchor

For the cell surface glycoprotein (Csg) of *H. salinarum* (Figure 4.3 in Chapter 4.1), lipid attachment to a C-terminal peptide of the Csg has been observed (Kikuchi et al. 1999). However, the C-terminal lipid attachment site (C-lipobox) could not been identified yet. The peptide, to which the lipid is attached to, does not contain a cysteine residue, and it was shown that a different type of lipid modification compared to the thioether linkage of archaeal NLip proteins must occur. It has been suggested by Kikuchi et al. that the serine/threonine residue (S_{827} in the *H. salinarum* Csg, Table 4.4A) that directly precedes the C-terminal hydrophobic domain might function as lipid attachment site for halophilic cell surface glycoproteins.

Table 4.4: **(A) Probable C-terminal lipid proteins** (CLips) were predicted by similarity with the cell surface glycoprotein (Csg) of *H. salinarum* (asterisk). For this protein, C-terminal lipid anchoring has been proven and a serine residue (box) was suggested as modification site (Kikuchi et al. 1999). However, the CLip candidates exhibit a conserved PGF motif (bold), which might also function as C-lipobox for lipid attachment. While one *Methanosarcina mazei* surface protein also contains a PGF motif (Supplemental Table 4.10), the predicted *H. marismortui* and several *H. walsbyi* Csg candidates lack this motif. C-terminal regions of probable CLip proteins resemble inverted signal peptides of NLip proteins, since they contain positively charged residues (bold), a hydrophobic stretch (grey), a probable C-lipobox (bold) as well as peptide repeats (blue). Most probable CLip proteins are likely to be N- or O-glycosylated and possible O-glycosylation sites within peptide repeat stretches are underlined.
(B) Probable C-terminal membrane anchor proteins (CAnch) with a C-terminal hydrophobic stretch (blue) are common in *S. solfataricus* while no putative NLip proteins were found in this species. The listed examples of putative *S. solfataricus* CAnch proteins are all involved in transport processes.

Id	Repeat region	Hydrophobic region	Charged region	O	N
A: Halophilic cell surface glycoproteins					
NP4620A	PEDTPEETPETPEETPEETPEDTPEEPDDQA	GFGAVVALIALIGAALLAT	RRRNALDN	X	X
NP4622A	EDTPEDTPEDTPEDTPEDTPEDTPEDPDDQA	GFGAVIALIALIGAALLAT	RRRNALDN	X	X
NP4734A	PEETPDDTPEETPEATPDDTPEETPEPDDQA	GFGAVIALIALLGAALLAT	RRRADRN	X	X
OE4755F	PTRTTTAATTTPETQSETTTTSRETGGPT	PGFTAVGALVAVVIVFAGVGL	RRRRE	X	X
OE4759F*	ETTTEMTTTQENTTENGSEGTSDGESGGS	IPGFGVGVALVAVLGAALLAL	RQN	X	X
HQ1197A	TQTVTSEPPTPTPDPTPTPETPTSTPSV	IPGFGLIVALVAIISLTMILRY	RRRQ	X	X
HQ1207A	TPEPTATEEPTPTATPEPTATSTPETGTGT	PGFGIVVALIALIAAALLAV	RRNN	X	X
HQ1346A	TSTSTSTSTPTPTPTRST…PTQTTETP	GFTAITAIAAVGIILIAASFRLH	RR	X	X
***M. mazei* surface layer protein B**					
NP_634319	GSNFGSEAALADAENIEDEPESGTVQKESVNTP	GFVAIYGLAGLLAVFLY	RRK		X
B: *S. solfataricus* substrate-binding proteins of ABC transporters					
NP_343949	TTTTTLQTTTTS…SVTSMTTTTSSSSSTL	IYAVIGIVIVIIIIVVAVVLL	RGRGRGGPGF	X	X
NP_343997	ATTSVTTTTSVTTTSISTTTVTVTSTSTI	PIIIAIVIVIIVIIAAVAILM	RRR	X	X
NP_344354	VSTTTSVSTSVSTTTATVTTTVTSSSNTT	LYAIIAVVVIIIVIIGVILGL	RRR	X	X
NP_344363	ASTSVTTTTSMSTTSVTSTVSTSSGLST	GVIAGIIVVIIVIIAVVAYVVV	RRR	X	X

When blast searching with the C-terminal part of the Csg sequence, several other candidates for C-terminal lipoproteins (CLip) were found in the halophilic genomes (Table 4.4A). These showed regional similarity to the cell surface glycoprotein, namely to a C-terminal hydrophobic domain and a preceding T-rich repeat pattern. Apart from the proposed serine/threonine lipid attachment site, another motif, PFG, was found to be conserved amongst putative CLip proteins. This motif is located at the start of the hydrophobic region and might also be a C-lipobox candidate for C-terminal lipid attachment.

The C-terminal regions of probable CLip proteins strongly resemble signal peptides of NLip proteins. The inverted signal peptide-like structure of their C-termini consists of a positively charged tail ('inverted N-domain'), preceded by a hydrophobic stretch ('inverted H-domain'), a putative C-lipobox ('PGF'), and a peptide repeat stretch (Table 4.4). These peptide repeats were found to be species-specific as already discussed for NLip proteins. The inverted signal peptide-like C-termini suggest that CLip proteins are cleaved prior to lipid attachment. C-terminal processing of the Csg and other probable CLip proteins by a peptidase would then result in a new free carboxyl end (e.g. of the glycine residue in the PGF motif), which might form an ester linkage with the modifying lipid. However, the hydrophobic peptide chain of CLip proteins might also remain unprocessed so that the *Halobacterium* S-layer protein and other CLip proteins would be attached to the membrane by a C-terminal lipid anchor as well as by a C-terminal transmembrane domain. Studies of the *H. salinarum* purple membrane indirectly indicate C-terminal peptidase processing of the Csg, though. The purple membrane of *H. salinarum*, which is located below the S-layer built of Csg molecules, consists only of bacteriorhodopsin and lipid molecules, but contains no peptides (Corcelli et al. 2002). If CLip proteins are indeed processed prior to their membrane attachment as indicated remains to be elucidated in future.

A more general search was performed for the 9 archaeal genomes in order to find proteins with C-terminal membrane anchors (CAnch) that were postulated as secreted proteins with hydrophobic stretches at their extreme C-terminus (Table 4.4B, Supplemental Table 4.10). As a result, CAnch proteins were found to similar extents in halophilic and non-halophilic euryarchaeota (except for *S. solfataricus*) comprising 0.4 to 0.9% of the theoretical proteome (Figure 4.6 in Chapter 4.5). Since all of the predicted CLip proteins were found amongst CAnch candidates, the observed C-terminal hydrophobic stretches might be a part of proposed inverted signal peptides as described above but they might also function as C-terminal transmembrane domains themselves.

In the predicted CAnch proteins, clustered positively charged amino acid residues (arginine and lysine) were frequently observed, which directly follow the C-terminal hydrophobic domains. Most CAnch proteins are probably translocated via the Sec system but few CAnch proteins such as some of halocyanins and *A. fulgidus* F420-nonreducing hydrogenase

(NCBI:NP_070210) are putative Tat substrates. In all haloarchaea, phospholipase C orthologs were amongst predicted CAnch proteins translocated by the Sec system. In contrast, phospholipase D has been described as a typical non-redox Tat substrate across species (Dilks et al. 2003).

As for NLip proteins, substrate-binding proteins and other ABC transporter subunits were found amongst CAnch proteins, especially in *S. solfataricus* (7 cases) (Table 4.4B). This crenarchaeote probably does not possess any N-terminal lipid anchors but instead exhibits C-terminal hydrophobic stretches more frequently than other archaea (1.3% of the theoretical proteome). Apart from transporter subunits, thermopsin precursors, proteases, and a *b*-type cytochrome have acquired C-terminal cell sorting signals in *S. solfataricus.*

4.4 Identification of secreted proteins by proteomics

The cytosolic (Tebbe et al. 2005) and membrane proteomes (Klein et al. 2005) of *H. salinarum* have recently been published, and currently 1117 proteins have been validated by various proteomics techniques. A large proportion of proteins of *N. pharaonis* (1170) have also been identified (F. Siedler, pers. comm.). The verified *H. salinarum* and *N. pharaonis* proteins were classified according to their proposed translocation system and cellular localisation (Table 4.5). For both, *N. pharaonis* and *H. salinarum*, a third of the probable integral membrane proteins and 40-60% of the probable membrane anchored proteins were found by proteomics. Furthermore, many extracellular proteins such as halocyanin F, halolysin, and alkaline phosphatase have been identified for *H. salinarum.*

In order to validate N-terminal peptides that are processed by signal peptidase cleavage in *H. salinarum* and *N. pharaonis*, normal MASCOT searches against standard protein sequence databases cannot be applied. To overcome this limitation, a special database was created and subsequently used for the MASCOT searches that contained the various

Table 4.5: **Identification of secreted proteins by proteomics.** Numbers of identified proteins with N-terminal signal sequences and anchoring signals are given sorted by their predicted translocation system and cellular localisation, respectively. The fraction of identified in comparison to complete set of predicted proteins for a category is given in parentheses. (Ex)Cyt - extracellular/cytoplasmic proteins, TMProt - integral membrane proteins, Anch - N- or C-terminally anchored proteins.

Species	Total identif.	Translocation system			Cellular localisation			
	proteins	Tat	Sec	Fla	Cyt	TMProt	Anch	ExCyt
N. pharaonis	1176*	43 (39%)	27 (25%)	1 (8%)	904 (44%)	188 (35%)	52 (42%)	21 (19%)
H. salinarum	1170*	50 (63%)	46 (41%)	5 (40%)	909 (42%)	161 (34%)	46 (67%)	55 (35%)

* - Numbers include 5 formerly deleted ORFs for which motif prediction tools have not been run.

potential processed N-terminal peptides. However, in spite of applying a different search procedure to check against available proteomics data of *N. pharaonis* and *H. salinarum*, only few N-terminal peptides that might have been cleaved by signal peptidase I, were identified by MS/MS (Table 4.6) and none by MS.

For *N. pharaonis*, two processed N-terminal peptides were identified, both with additional N-terminal modifications (pyroglutamate, N-acetylation). Six processed N-termini were identified in *H. salinarum*, 3 of them part of integral membrane proteins and the other three of flagellins. Unexpectedly, the identified N-termini of flagellins A, B, and X differ from the predicted ones, which would start with the glutamine residue of the RG↓Q motif proposed to be recognized and processed by preflagellin peptidase (Thomas et al. 2001). All three flagellins revealed cleavage sites behind the H-domain of the signal peptide, though, in contrast to isolated flagellins from methanogens cleaved before the hydrophobic stretch (Kalmokoff et al. 1992). This could mean that *H. salinarum* flagellins are further processed after preflagellin cleavage. However, at this stage unspecific proteolytic cleavage of the flagellins and other listed halophilic proteins cannot be excluded. The specific search for processed N-terminal peptides resulted only in few hits, although enquiries were run against a vast amount of proteomics data. This indicates that the applied search specifically detected

Table 4.6: **Identification of probable processed N-termini after signal peptidase cleavage** by MS/MS. Cleavage sites predicted by SignalP (CL_{pred}) are compared to the ones identified through MASCOT enquiries (CL_{exper}). Only peptides reaching MASCOT difference scores (DScore) above 20 were considered. Amino acids of identified peptides were underlined when found within the y-series. Further peptide modifications (Modific.) as well as the number and position of predicted transmembrane domains (TM) are given. For OE1866F, an internal (i) peptide fragment resulting from trypsin digestion was detected, which is located before the predicted processed N-terminus.

Id	CL_{pred}	CL_{exper}	DScore	Peptide	Modific.	TM	Function	Sample
OE1988R	27	27	70-38	SSGYQSA	-	2: 41,63; 84,106	cytochrome C oxidase	1dg0063_15 and other gels
OE2469F*/ OE2470F*	13	33	56	AGVLINTAGFLQSK	-	-	flagellin A	1dg0108_08
OE2397F*/ OE2398F*/ OE2399F*	13	33	38	AGVLINTAGTLQSK	-	-	flagellin B	1dg0116_23
OE3688F	33	28 29	36 16	GSVVGTLPLGSR SVVGTLPLGSR	- -	2: 42,61; 66,85	HY	ndgel0156_18 1dg0042_12
OE1866F*	41	45 i37	28 48	QAADGPEALSR IAHAQHVTQAADGP…	PyroQ -	3: 57,76; 80,97; 118,137;	HY	ndgel0156_12 ndgel0156_15_2 (and in Halolex)
OE2695F	8	37	22	MLQSR	Mox	-	flagellin X	1dg0078_14
NP2464A*	21	26	30-28	QMQDPLFVK	PyroQ	1: 507,526;	CHY	1dg0004_02
NP6012A	27	18	29	SLAVPR	NAc	-	CHY	1dg0001_09

* - Proteins which have been identified previously but without verification of the N-terminus.

modified peptides from highly abundant proteins that may have somehow been modified upon sample preparation. This is further supported by the fact that (i) flagellins A and B were processed at different sites than flagellin X and that (ii) alternative processed N-terminal peptides were found for OE3688F and OE1866F (Table 4.6).

Searches for N-termini of NLip proteins that have been processed by signal peptidase II were also performed, but no proteins with lipid-anchors could be identified when performing specific MASCOT searches. This may be explained by the applied proteomics techniques. Identification of hydrophobic peptides, especially of peptides with covalently attached lipids, is difficult by MALDI-TOF or LC/MS-MS. As a further complication, many processed NLip N-terminal peptides are likely glycosylated, and glycoproteins are known to escape detection by standard proteomics techniques. The validation of processed N-terminal peptides will require more elaborate proteomics approaches in future, which specifically cope with such modified peptides. As could be shown, the problem cannot be solved by advanced bioinformatics analysis, but requires application of advanced experimental techniques.

4.5 Conclusions

For haloarchaeal secretomes high numbers of proteins are predicted to be attached to the outer membrane by N-terminal membrane anchors compared with non-halophilic archaea. On the other hand, numbers of extracellular secreted proteins are reduced in halophiles (Figure 4.6, Supplemental Table 4.11). In the haloalkalophile *N. pharaonis*, similar extents of N-terminal lipoproteins and extracellular proteins occur indicating that secreted proteins are protected by lipid anchoring from alkaline extraction. *S. solfataricus* has no predicted NLip proteins, and it remains to be elucidated in future whether N-terminal lipoproteins occur in euryarchaeota but generally not in crenarchaeota.

Secreted proteins which are involved in protein-protein interactions at the membrane seem to frequently adopt membrane-retention signals in order to prevent their diffusion into the extracellular space. Membrane attachment ensures close contact of interaction partners, and, thus, might facilitate protein interactions. For instance, substrate-binding proteins of ABC transport systems frequently contain lipobox motifs, e.g. in *N. pharaonis* (Figure 4.7), *Methanosarcina mazei* (NP_633386), *Archaeoglobus fulgidus* (NP_069268), and *Pyroccus furiosus* (NP_579503). Further, substrate-binding proteins of *N. pharaonis* and *S. solfataricus* (Table 4.4B in the previous chapter) were found to exhibit predicted C-terminal anchor regions. Many protein sequences of halocyanins also showed various membrane-retention modules (lipobox, C-terminal hydrophobic region, transmembrane domains, Supplemental

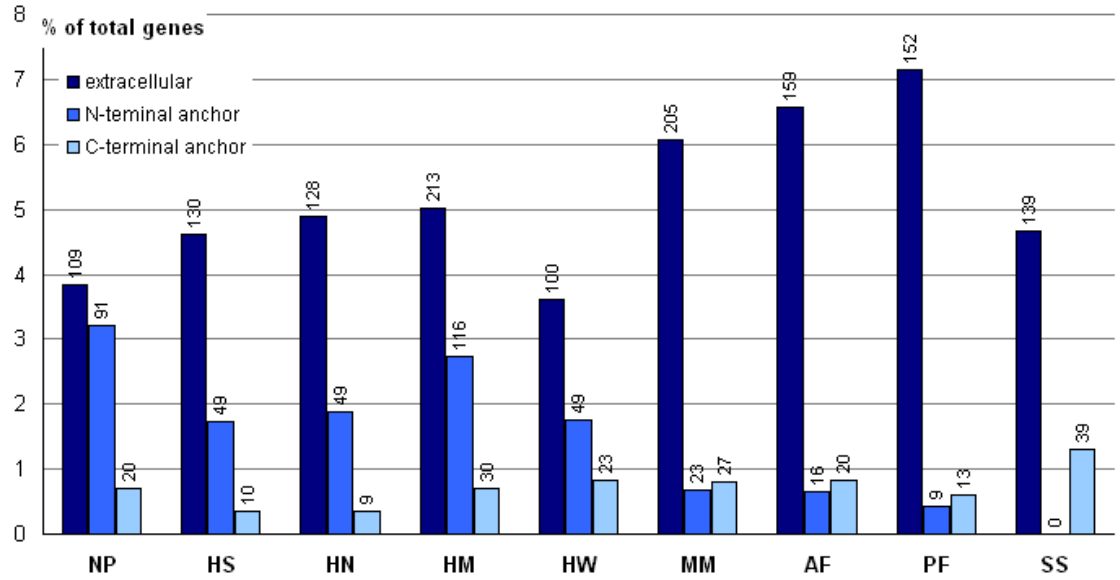


Figure 4.6: **Predicted extent of N- and C-terminal anchoring modules in secreted proteins** from 9 archaeal genomes. N-terminal anchor proteins comprise well-defined N-terminal lipoproteins and proteins with an N-terminal hydrophobic stretch as proposed for archaeal flagellins. C-terminal anchor proteins are defined as proteins with a signal peptide and a hydrophobic domain at their extreme C-terminus, and some of them also reveal features of C-terminal lipoproteins such as the Csg of *H. salinarum*. The actual number of proteins assigned to the different categories is given above the bars. For species abbreviations see Table 4.1.

Table 4.12), which might ensure interactions of the halocyanins with respiratory complexes. Different types of anchoring modules were also observed for glycoproteins probably involved in cell envelope formation in *N. pharaonis* (Chapter 3.2.7).

In future, it would be interesting to conduct an evolutionary study of how and when secreted proteins such as halocyanins have gained their varying membrane-retention modules. However, the adoption of cell sorting signals in secreted proteins might be difficult to study. Archaeal secretomes are exhibited to extremely different habitats, and, thus, require different types of cell envelope proteins and extracellular enzymes. There seem to be only few orthologous membrane proteins available for interspecies comparison but many species-specific secreted proteins with yet unknown functions in archaea.

Archaeal proteins with predicted N- and C-terminal lipid anchors reveal peptide repeat patterns that are located adjacent to the lipid anchors in the processed proteins. These peptide repeat patterns, which are found in proteins involved in cell envelope formation, transport, and metabolism, seem to commonly serve as O-glycosylation sites as shown for S-layer proteins in haloarchaea (Lechner and Sumper 1987). For Csg proteins in *Halobacterium* and *Haloferax*, it was shown that differences in glycosylations are important for the adaptation to specific halophilic environments (Sumper 1993). Therefore, N- and O-glycosidic modifications for halophilic secretomes should be predicted in future in order to establish complete glycoprotein inventories and to estimate the importance of glycosylation for cell surfaces under halophilic conditions.

```
RP: NP_947753     1    MTGTVLRGLHAAVLTGTLVLASGAALA   28
NP: NP2004A       1 MRNPTVSRRKLLASGAAAAGIGLAGCMGGNGG 32
                                    lipobox
RP:              29 AEKPIKLGVLEDQSGDFAVATIGKVHAIQLAADEINKAGGIMGRPLELVVYDTQSDNTRYQE 89..
EXP=8e-84,ID=39%    ++ P  +G+LED+SG+F +    K  A +LA +EIN  GGI+G  +E+V  D QSDN RYQE
NP:              33 SDGPT-VGILEDRSGNFQLNGTSKWQATRLAIEEINEDGGILGEEVEIVDPDPQSDNERYQE 93..

AF:           ..243 NPDVIIVSSGSGMGGGKDVVYEWVVSDDRLSGIKAVKEGRVYVVDADIINRPSYRLAEAIEVVAD 307
EXP=6e-24,ID=29%    NPDVI+ S    +   +   YE  V         AV+ G+V  VD++ +++P+ RL  A+E +AD
NP:           ..246 NPDVIVASDAEPL--PETAAYESTV---------AVEHGQVVTVDSNDVSQPAPRLVYALETIAD 299

AF: NP_071272  308 LIHK 311
NP: NP3814A    300 GLADAETADRPTDATVDDQPGVGIAPAAAGLVLAVVAGLLVRTLRRR 346
```

Figure 4.7: **Examples of substrate-binding proteins of ABC-type transporters** in *N. pharaonis* (NP) compared with orthologous subunits from *Rhodopseudomonas palustris* (RP) and *Archaeoglobus fulgidus* (AF) In the upper alignment the *N. pharaonis* protein exhibits a twin-arginine motif (bold) for protein translocation in the folded state and a lipobox (red) for N-terminal lipid attachment at the end of the H-domain (blue). In the second alignment the C-terminus of the *N. pharaonis* substrate-binding protein reveals a hydrophobic domain (blue) that is proposed to be involved in C-terminal membrane anchoring. The hydrophobic stretch is followed by positively charged residues (bold).

4.6 Methods

For the characterization of 5 halophilic and 4 non-halophilic secretomes, available prediction tools and newly developed motif recognition programs were applied. Primary prediction results from the different tools were combined to estimate the total number of secreted proteins and the number of proteins with membrane anchors (Figure 4.8).

4.6.1 Analysis of secreted proteins

Proteins exported via the Sec-mediated pathway were predicted using SIGNALP v3.0 (Bendtsen et al. 2004). However, substrates of other secretion systems containing a twin-arginine or flagellin cleavage motif (see below) were excluded. The insertion of integral membrane proteins is often dependent on the Sec system but Sec-independent signal peptide cleavage and membrane insertion also occurs (e.g. for bacteriorhodopsin) (Bolhuis 2002). Therefore, proteins with TMHMM predictions (Krogh et al. 2001) were generally excluded from the list of putative Sec substrates in case the TM domains are located behind position 70. This rule was applied, since both, the H-domain of signal peptides and TM domains, contain hydrophobic residues, so that SignalP and TMHMM predictions produce ambiguous results for the N-termini of proteins. Since signal peptide predictions were preferably considered in this study, TMHMM results has to be modified by excluding TM domains in the signal peptide region up to position 70.

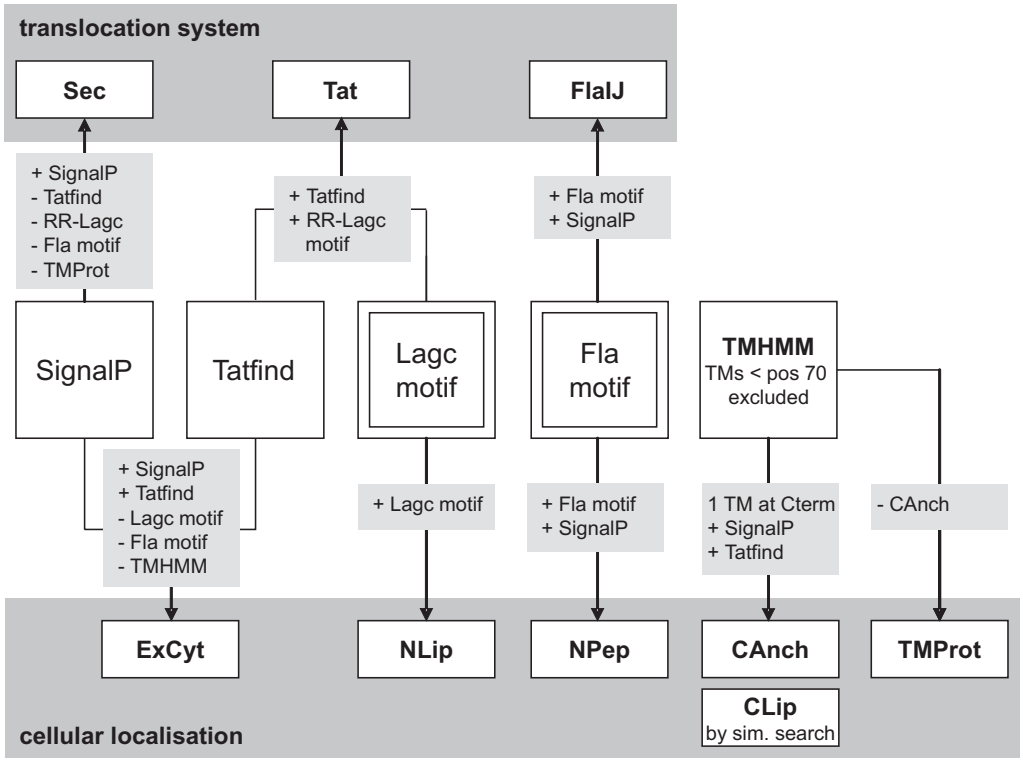


Figure 4.8: **Prediction of signal peptides and cell sorting signals.** Available (square boxes) and newly developed (double-lined boxes) programs were applied to predict Sec-like, Tat-like and Fla-like signal peptides of secreted proteins (upper part). These tools were also used to predict cell sorting signals, e.g. lipobox and flagellin-cleavage motif, as well as transmembrane (TM) domains (lower part). Arrows indicate which prediction results were used to retrieve final protein lists. Results from TMHMM were processed by excluding TM domains located in the N-terminal region of proteins up to position 70. A subcategory of TM proteins with one C-terminal TM domain was defined as C-terminal anchor proteins. The remaining TM-containing proteins (TMProt) are likely integrated into the membrane. Proteins with C-terminal lipids were found by similarity searches with the *H. salinarum* Csg. Lipid-anchor proteins were characterized using NetO(N)Glyc and a developed peptide repeat search routine.

Since the Tatfind program (Rose et al 2002) showed some false negatives, the complete number of substrates translocated via the Tat-mediated pathway was established by complementation with NLip proteins (see below) which additionally exhibit a twin-arginine motif. Some of the predicted Tat substrates revealed TM domain predictions but they were not excluded from the list of probable Tat substrates as they belonged in most cases to the category of CAnch proteins, e.g. *N. pharaonis* halocyanin 4 (NP0050A).

Signal peptides probable recognized by FlaIJ were predicted by the developed Fla_motif program (Perl) which checks the first 12 aa of proteins detected by SignalP for the preflagellin peptidase cleavage site (RGQ motif). Since this recognition site differs significantly amongst the domain of archaea (Thomas et al. 2001) only the numbers of haloarchaeal flagellin-like signal peptides containing a RGQ motif were determined. The number of flagellin precursers in non-halophilic species was retrieved from the published genome annotation. Proteins found by Fla_motif were also assigned to the category of anchored proteins with a hydrophobic peptide N-terminus.

4.6.2 Analysis of membrane-anchored proteins

N-terminal lipid anchor (NLip) proteins are characterized by their lipid attachment site (lipobox) which has been described in the PROSITE entry PS00013 (Hulo et al. 2004). Of the 3 given rules: (1) the sequence must start with Met, (2) the cysteine must be between positions 15 and 35 of the sequence in consideration, and (3) there must be at least one charged residue (Lys or Arg) in the first seven residues of the sequence, only rules (1) and (2) were applied within the Perl program Lagc_motif. Rule (2) was modified allowing the cysteine residue to be located up to position 50 in order to recognize lipobox motifs even in case of N-termini that have been incorrectly assigned. It should be noted that prediction of lipobox motifs and other N-terminal motifs strongly depends on the quality of start assignments and that false negatives occur in case of misassigned starts (e.g. *H. marismortui* rrnAC1377, Supplemental Table 4.12).

Since there is no data available describing a specific C-terminal lipid attachment site (C-lipobox), probable CLip candidates were found by similarity to the known CLip proteins (Kikuchi et al. 1999). An alignment of CLip candidates was used to propose a C-lipobox. Peptide repeat patterns in NLip and CLip proteins were initially determined for *N. pharaonis* by searching with different patterns against the raw set of 11874 distinct ORFs. Patterns were matched within sequence windows of 30 aa (50 aa for the [VP][TE][ED]T pattern) and the minimal number of matches was adjusted so that only valid genes were found (13 hits for [VP][TE][ED]T (2 repeating units), 8 hits for NG (4x), 7hits for EP (5x), 11 hits for PE (6x)). Most of the found peptide repeat patterns were located directly adjacent to predicted lipid-anchor regions in secreted proteins. In the other halophiles, peptide stretches of predicted N- or C-terminal lipid anchors were checked manually for peptide repeats. Glycosylation sites were predicted by applying NetO(N)Glyc (Julenius et al. 2005).

Proteins with a probable C-terminal membrane anchor (CAnch) were defined as proteins with an N-terminal signal peptide (type I or II), a single predicted TM domain close to their C-terminus, and a minimal length of 120aa. The allowed distance from the end of the hydrophobic domain as predicted by TMHMM to the C-terminus was set to 20 aa, since a preliminary statistics showed that single C-terminal TM domains mostly terminate less than 10 aa from the C-terminus (Table 4.4, Supporting Table 4.10). Numbers of integral membrane proteins were obtained by excluding CAnch proteins from the set of proteins with valid TMHMM predictions (Chapter 4.6.1).

The number of extracytoplasmic proteins (ExCyt) results from the remaining set of secreted proteins that exhibit no N- or C-terminal anchoring signals. Average pI values of ExCyt proteins were calculated with sequences behind the signal peptide region (behind position 70). Cytoplasmic proteins contain neither N-terminal signal sequences nor transmembrane domains.

4.6.3 Identification of secreted proteins by proteomics

Fractions of validated proteins for various translocation systems and cellular localisation categories were calculated by checking the total list of protein identifications in *H. salinarum* and *N. pharaonis* proteins derived by different proteomics techniques against the lists of secreted proteins and proteins with membrane-retention signals.

For validation of signal peptide processing through proteomics data, special databases were generated for each of the two haloarchaea. These contain the collection of processed N-termini as proposed by several prediction tools (Table 4.7). Since cleavage sites of signal peptidase I and preflagellin peptidase have not been experimentally validated yet for any of the two halophilic strains, theoretical peptides with varying cleavage sites were generated for database I. Ambiguous peptides which could be either an internal peptide derived by trypsin cleavage (peptide position -1 = [RK]) or an N-terminal peptide resulting from mRNA translation (peptide position 0 = [MV]) were marked and later excluded from the result list. Database I was checked against available MS/MS (*H. salinarum, N. pharaonis*) and MS (*H. salinarum*) data using the MASCOT Daemon enquiries and a Perl script, respectively. The latter matches masses of theoretical processed N-terminal peptides against masses of identified MALDI peptides. For MASCOT enquiries, peptide modifications by N-acetylation (NAc) and pyroglutamate (PyroGlu) were allowed added to the usual variable peptide modifications, which are due to applied experimental procedures (carbamidomethyl-cysteine and oxidized methionine-residues (Mox)).

Database II contains putative processed N-termini of NLip candidates, which always start with a cysteine residue. This has been shown to be modified by a diphytanylglyceryl anchor (lipanch_2020) via a thioether linkage (Mattar et al. 1994). Since C20-C25 lipids also occur in *N. pharaonis*, another potential lipid anchor was defined. The monoisotopic (and average) molecular weights of the 2 different lipid anchors including the modified cysteine residue were calculated using MolE v2.0; lipanch_2020: 737,67149 (738,2933) and lipanch_2025: 807,74969 (808,4277). Database II was also checked against available proteomics data.

Table 4.7: **Parameters used by the MASCOT Daemon (MS/MS) and by the developed mass checking program (MS).** Two databases with theoretical processed N-termini were generated with available proteomics data for each of the halophiles, *H. salinarum* and *N. pharaonis*. The first database contains peptide variations derived through cleavage by signal peptidase I and preflagellin peptidase and the second database stores putative N-termini derived through signal peptidase II processing.

Signal peptidase	Cleavage site	Type of cleavage	Prediction tool	Cleavage site variation	N-terminal modification
I	**AXA↓A**	unspecific	SignalP	±10 aa	NAc, PyroQ
II	**LAG↓C**	specific	Lagc_motif	±0 aa	lipanch, NAc
preflagellin	**RG↓Q**	specific	Fla_motif	±5 aa	NAc, PyroQ

4.7 Supplemental material

Supplemental Table 4.8: **Alignment of N-termini with flagellin-like cleavage sites.** The RGQ motif is the predicted cleavage site (vertical arrow) of haloarchaeal flagellins by similarity (Thomas et al. 2001) but was also found within signal peptides of the listed proteins which are found in conserved gene clusters (Figure 4.4).

Code	Sequence ↓
NP0150A	-------MPDRG QLSLSIVEAGVGVVFVLAVALGFALGVPAPDTETPQLD
rrnAC1946	--------MRRA QLPLSLVEVALGTVLILGVALGFALGTPAPDRQGPQLD
rrnAC1424	--MNVPRGQQRA QSAPLGLLLVLSLVIVGSGVVVSLGATALVDTEAGLDV
rrnAC1425	MGSAERGLGSRG QSAPLGLALVFAVMIVSTTAVVALGADAITSTQTQLDV
NP0148A	--------MNRG QLVLVAAAVIAIGLVPILFAYLQLGFHPDVDRTPAVAG
rrnAC1945	-------MTRRG QLVLVAATVVAVALVPILFASLQLGYHDDVRATADYDD
NP4016A	---------MRG QAHTVEAFVAAVLIVGGLVFATQATAVTPLSASTSNQH
rrnAC1421	---------MRA QAHTLEAIVSGMLLLASLVFALQMTAVTPLSASTSSQH
OE1334R	----MPERGRRG QASLPAVEAAIGVFVILAVAATFTVGVPGDGGHTRTAQ
OE1332R	-------MTQRG QFVLAGAAVAALALASVAVAYLQFGYAPSVATPRPTPS
NP4008A	-------MDDRA VSNTVGIVLILGMTIAAVSALVFVGGAVLEDTRADTEQ
NP4010A	-------MTDRA VSDVVGYVLVFSLIVATIGIVTTVGFATLDDRQSAEQL
rrnAC1423	-------MADRA VSEVLSFALVFSLIVASIILVSVSGLGALQNARDAEQM
NP0156A	-------MSRRA QSNVVGVALLLGIAVVSMGALTATVGMLVDSNAAAADA
OE1340R	--------MTRG QSAVVGVAVLVAATVVAVAALTASVGTVVTEHAAAADS
rrnAC1247	---------MRG QAYTLEGVLAAIVVVTATVYGLSAVDTGPFQTGAQQRT
NP0152A	--------MSRG QANLPALAIALLVLTTVAVLAVTIADASVRGAERNAAD
rrnAC1947	...VPANSLRA QTSLAALGIALVLLTVVTSLGIAIADTAIAGADRTPDE
OE1336R	--------MQRG QANLVALVVGVLLVGAAVTLAVGAGADAFARADRSPAE
NP4012A	------MTDDRG VSVTVGYVLNLAIAAILVSALLMAGGSLIESQTEQVTH
NP4014A	-----MTDRDRG QLLLVAGIVLAVLFVALALLVNTAIYTDNVATRDGDAA
rrnAC1246	-----MVANERG QLLLIGGVAIAIVVFSTILFAHSLAVTDGITTTGSADT
rrnAC1248	-------MDSRA QTPQDFAVGVSVLLVTIIGVLAFVQGSAVGVYESPDVQ
NP4018A	---MISVPTDRG QTTLDFAIGTSVFLLATIFVVAYVPTMFDPFAGGGGSK
rrnAC1420	...IEEHRRG QTTLDFAIGMSLFLSVLIFIFLFIPGLLSPFSAGVQEE

Supplemental Table 4.10: **A selection of putative C-terminal membrane anchor proteins.** Predicted translocation systems and functions for these proteins are indicated. For further explanations see legend of Table 4.4.

Id	Hydrophobic region	Charged region	SEC/ TAT	Function
NP0050A	ETLIGGPEEEDDDEPEDLPPAFEMDSEVFAMFGVAGLLALLSPLGVLYLM	RRQNGGEPE	TAT	halocyanin 4
NP0608A	TVTPETTAAPDEATPTPGTTDDGSLPMPESQSGFGVVAAVLALLAGVVSA	RRR	SEC	5'-nucleotidase
NP1876A	ETRSCNEFEDNPDAVRLETEDDEAVEDLAGFGVLVALIAIGLALVRFSRLG	RSKNA	TAT	CHY
NP3814A	LETIADGLADAETADRPTDATVDDQPGVGIAPAAAGLVLAVVAGLLVRTL	RRR	SEC	ABC
OE1879R	VNENGQAPSGGGGPVERSPHEMGVPIQAHYVGIATILMMILSMVYTFFVL	KYGESRNA…	TAT	halocyanin E
OE3017R	NGTATTVDLTAETDTTTTPPADTDAGTGLPVPGFGIGVAVLAVFGAALLAS	RR	SEC	sugar hydrolase
HQ1596A	TVDDKEELNTVSTETNVNQSVSTGINLDNTMKNLAGATLVIIFIIAVMWS	RR	SEC	cell surface protein
HQ1193A	VDGIITGSFIDTELGRVVDNTVKSSIQTRLMNPVAAGFILILSIVSLLLY	RRR	SEC	CSG
rrnAC0764	ASESTAESAGTANAAGTTADSGDRSTTSANGPLPIALAIAALAAVAALAAR	TSRQP	SEC	CSG
rrnAC1372	PAETAAAQQTAEPQQAGVSGGSSDGSSPTSAAAFMMAAALLIALLAVAYAQ	RKR	SEC	CytC peroxidase
Mm:NP_633625	NSESDYVEGYEMQKETPVTEEESGGPSFSGSDIFGILFVLAAVGGIYLGF	RKKKV	SEC	Cob synth. protein
Af: NP_070210	APCIACSEPGWPDKFSPFYAELPSIPSFLGLNPTTLGAGIFGATAVGIGV	HAVRRKLR…	TAT	F_{420} hydrogenase
Af:NP_070807	SGDSAYDTMLISNAARIAYAGGYAECSRDSVWIYVLSVLCVVEALIIAVIK	VRL	SEC	ABC
Pf:NP_578102	EVLIQNSKSVIGALKEEKKVVVKEGNEKVQYIAISLLTGLIVGVSIGVVI	RKCPVF	SEC	ABC
Ss:NP_343433	GRFIGTYNLTLGGTIVVNKPIVEKLQLSINNLLLEITAIIIVIVIIMLIL	RKRR	SEC	Thermopsin
Ss:NP_342629	QASQGMEQAFIQALVQNGLMSSLSPLPIIPQYMLLLISVILTPMVVVITP	RKRW	SEC	ABC
Ss:NP_343936	QELINIQSSRKGSAITNYIPQMILVITIIFYAITKDIIITLLMVILAFSI	SLSGRKNT	SEC	ABC
Ss:NP_343441	ENNITLIAKDLWGKTAVKTLIVNSGYNYVGIGIIAGIILIIVIVVILVIS	KRK	SEC	Protease
Ss:NP_343565	INSTVGNNVNYITTIGNNHAKSSYPSLDSGSILTIGIVLDIITIIALILI	KRRKKFI	SEC	Peptidase
Ss:NP_343890	NLTFNLLNNYHLIIVQDHLKALQGSVNLLTVIAIISLIIAIIAVALLFVF	TRRR	SEC	Serine protease
Ss:NP_344124	STTTSTISSTSVTTTTSITTVTSIPSTTIYVTIVGVVIAIIALIILYVVF	RR	SEC	Cytochrome b

Supplemental Table 4.9: **Frequent lipobox variations observed in archaeal proteins**. Bold numbers mark lipobox variations that occur at least three times in the given species (for species abbreviations see Table 4.1).

Species	**XAGC** **L**AGC	**V**AGC	**T**AGC	**A**AGC	**F**AGC	**I**AGC	**S**AGC	**LXGC** L**S**GC	L**L**GC	L**T**GC	**XSGC** **V**SGC	**A**SGC	**F**SGC	**I**SGC	**G**SGC	**SGSC**	**Frequent lipobox motifs**	**All lipobox motifs**
HM	**63**	**13**	**9**	0	**9**	**7**	**6**	**9**	0	1	0	1	0	0	1	0	**116**	154
NP	**39**	**16**	**11**	**8**	**8**	**5**	2	**4**	0	0	0	0	0	1	0	0	**91**	115
HQ	**25**	**3**	**5**	2	**3**	**5**	1	**4**	0	**4**	0	0	1	0	0	0	**49**	83
HN	**25**	**6**	**7**	**3**	1	**5**	0	**3**	0	2	1	0	0	0	0	0	**49**	76
HS	**23**	**7**	**7**	**3**	2	**6**	0	**3**	0	2	1	0	0	0	0	0	**49**	75
AF	**4**	0	0	0	1	1	0	**5**	**4**	1	0	0	**3**	0	0	0	**16**	42
PF	0	2	2	0	1	0	1	0	1	0	**3**	**6**	1	1	1	0	**9**	30
MM	0	0	2	2	0	0	0	1	0	0	**3**	**4**	**4**	**4**	**3**	**5***	**23**	67
SS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	7

* - Proteins with the motif variation SGSC observed in clustered genes of the *M. mazei* genome are likely false positives NLip proteins.

Supplemental Table 4.11: **Predicted cellular localization for complete predicted proteomes of** selected archaeal genomes. For species abbreviations see Table 4.1.

	Extracellular		Anchored		Membrane		Cytoplasmic		Total
	#	%	#	%	#	%	#	%	
NP	109	3.83	123	4.33	534	18.78	2077	73.06	2843
HS	130	4.61	69	2.45	474	16.80	2148	76.14	2821
HN	128	4.88	65	2.48	444	16.93	1985	75.71	2622
HM	213	5.02	152	3.58	832	19.62	3043	71.77	4240
HW	100	3.60	73	2.63	489	17.61	2115	76.16	2777
MM	205	6.08	53	1.57	573	17.00	2542	75.41	3373
AF	159	6.57	38	1.57	369	15.25	1856	76.69	2422
PF	152	7.15	24	1.13	348	16.38	1602	75.39	2126
SS	139	4.67	40	1.34	511	17.16	2288	76.86	2978

Supplemental Table 4.12: **Cell sorting motifs in halocyanins.** All halocyanins are likely translocated by the Tat system (RR- twin-arginine motif), and most probably exhibit an N-terminal lipid anchor (LAGC - lipobox) or a C-terminal membrane anchor (CANC – C-terminal hydrophobic stretch). One of the *Halobacterium* and one of the *Haloarcula* halocyanins have 2 predicted transmembrane (TM) domains indicating integration into the cell membrane.

Motifs	Id	Title
RR-CANC	NP0050A	halocyanin 4
RR-LAGC	NP0938A	halocyanin 3
RR-LAGC	NP1600A	halocyanin-like protein 2
RR-LAGC	NP3232A	halocyanin-like protein 3
RR-LAGC	NP3954A	halocyanin 1
RR-LAGC	NP4692A	halocyanin-like protein 1
RR-LAGC	NP4744A	halocyanin 2
RR-LAGC	OE1391R	halocyanin hcpG
RR	OE1859R	halocyanin hcpF
RR-CANC	OE1879R	halocyanin hcpE
RR-LAGC	OE2157F	halocyanin hcpH
RR-LAGC	OE2171F	halocyanin hcpC
RR	OE2704F	halocyanin hcpD
RR-LAGC	OE3320F	halocyanin hcpA
RR-TM	OE4073R	halocyanin hcpB
RR-LAGC	pNG6063	halocyanin precursor-like
RR-LAGC	rrnAC0508	halocyanin precursor-like protein
RR-LAGC	rrnAC0510	halocyanin precursor-like
RR-CANC	rrnAC0735	halocyanin precursor-like
RR-TM	rrnAC1152	halocyanin precursor-like
*RR-LAGC**	rrnAC1377	halocyanin precursor-like
RR-LAGC	rrnAC3204	halocyanin precursor-like
RR-CANC	rrnAC3329	halocyanin precursor-like
RR	rrnB0097	halocyanin precursor-like
RR-CANC	HQ1548A	halocyanin hcpE
RR-CANC	HQ1557A	halocyanin hcpF
RR-LAGC	HQ2357A	halocyanin hcpH

* The N-terminal motif of this *Haloarcula* halocyanin was only detected when the start was reassigned to MTETDD.

CHAPTER 5

Metabolic Pathway Reconstruction for *Halobacterium salinarum*

A metabolic database, Pathnet, was developed which integrates general pathway information based on KEGG and organism-specific data derived from genome projects as well as from experiments described in the literature. The stored pathway data can then be accessed and combined with experimental data resulting from different 'omics'-techniques through coloured KEGG maps. The metabolism of *Halobacterium* salinarum was reconstructed by creating 514 expert-curated reaction entries in Pathnet that are used as input for future metabolic models for this halophile. As the quality of these models strongly depends on the accuracy of function annotation for genes, EC number assignments of enzyme genes were rigorously assessed in the metabolic reconstruction process. The established metabolic subsystems for *Halobacterium* were reviewed through comparison of genomic and biochemical data. Thereby, non-orthologous gene displacements that are frequently occurring in the archaeal domain of life as well as remaining pathway gaps were discussed. Identified enzyme genes of *H. salinarum* were consistent with the previously observed degradation of glycerol and amino acids. For *de novo* biosynthesis of nucleotides and prenyl-based lipids, standard biosynthesis pathways were predicted. However, the lack of pentose-phosphate pathway genes in most archaea raises questions about the origins of major biosynthesis precursors such as ribose 5-phosphate and erythrose 4-phosphate. The biosynthesis of 14 amino acids and 8 coenzyme classes were proposed, and heme and cobamide synthesis was analysed in detail.

5.1 Introduction

A metabolic pathway consists of a set of metabolic reactions, in which chemical compounds are converted amongst each other. The participating compounds are commonly divided into main and side substrates (often coenzymes). Under physiological conditions, enzymes are required for most biochemical reactions in order to catalyze the conversion of metabolic

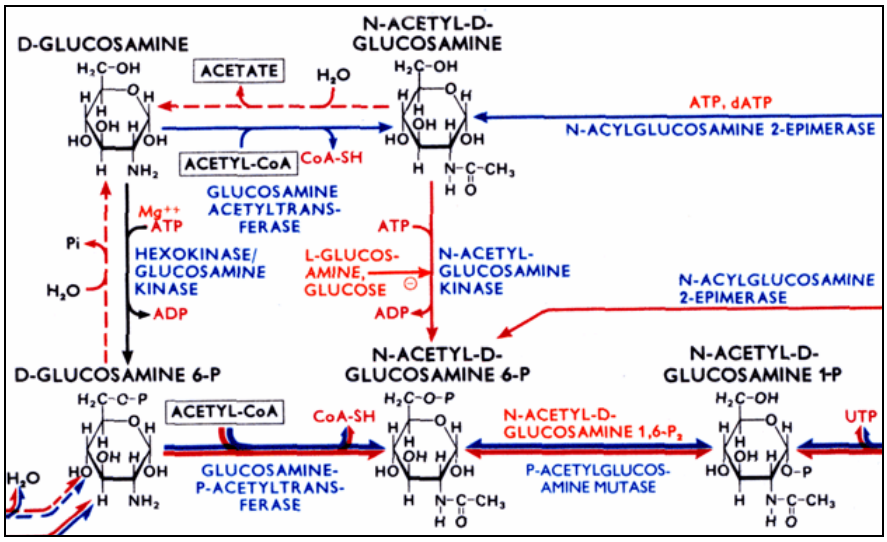


***Figure 5.1*: Metabolic pathway represented by the Roche Applied Science 'Biochemical Pathways' wall chart.** Main compounds (given by a chemical structure) of the amino sugar biosynthesis pathway are connected by directed arrows that represent a metabolic reaction. Side substrates and coenzymes are linked by arcuate arrows. Enzymes (in blue capital letters) are assigned to each metabolic reaction. Compounds known to influence enzyme activities are indicated by minus/plus signs depending on the regulatory effect. Colours of reaction arrows define the type of organism, in which the reaction occurs (black - general pathway, red - prokaryotes, blue - animals, green - plants/yeast). Further legend explanations can be found in (Michal 1999).

compounds. Enzymes consist of one or more protein chains encoded by genes on the organisms' genome. The protein sequences resemble each other in different species, a fact which is used to predict the enzyme equipment of a newly sequenced organism by similarity-based function transfer. The activity of an enzyme in a metabolic reaction is regulated by lower molecular compounds, often by the product of the pathway the enzyme is involved in (feedback inhibition). Data on metabolic pathways can be represented on metabolic maps (Figure 5.1).

Metabolic databases such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa et al. 2004) and MetaCyc (Krieger et al. 2004) provide collections of metabolic reactions and pathways which were observed experimentally in any organism (reference pathways). Furthermore organism-specific metabolic pathways can be retrieved that have been predicted automatically from complete genome sequences. For the enzymes involved in a metabolic pathway, genes are identified within the published genome sequence by similarity search. The genes are then internal re-annotated and assigned to orthologous groups (KO database) within the KEGG system (Kanehisa et al. 2004). PathoLogic, an automatic annotation program used to generate computationally-derived Biocyc databases, considers not only data from similarity searches, but also other evidences, i.e., is the gene part of an operon or are there functionally-related genes nearby in the genome, in order to assign genes and to fill pathway gaps (Green and Karp 2004).

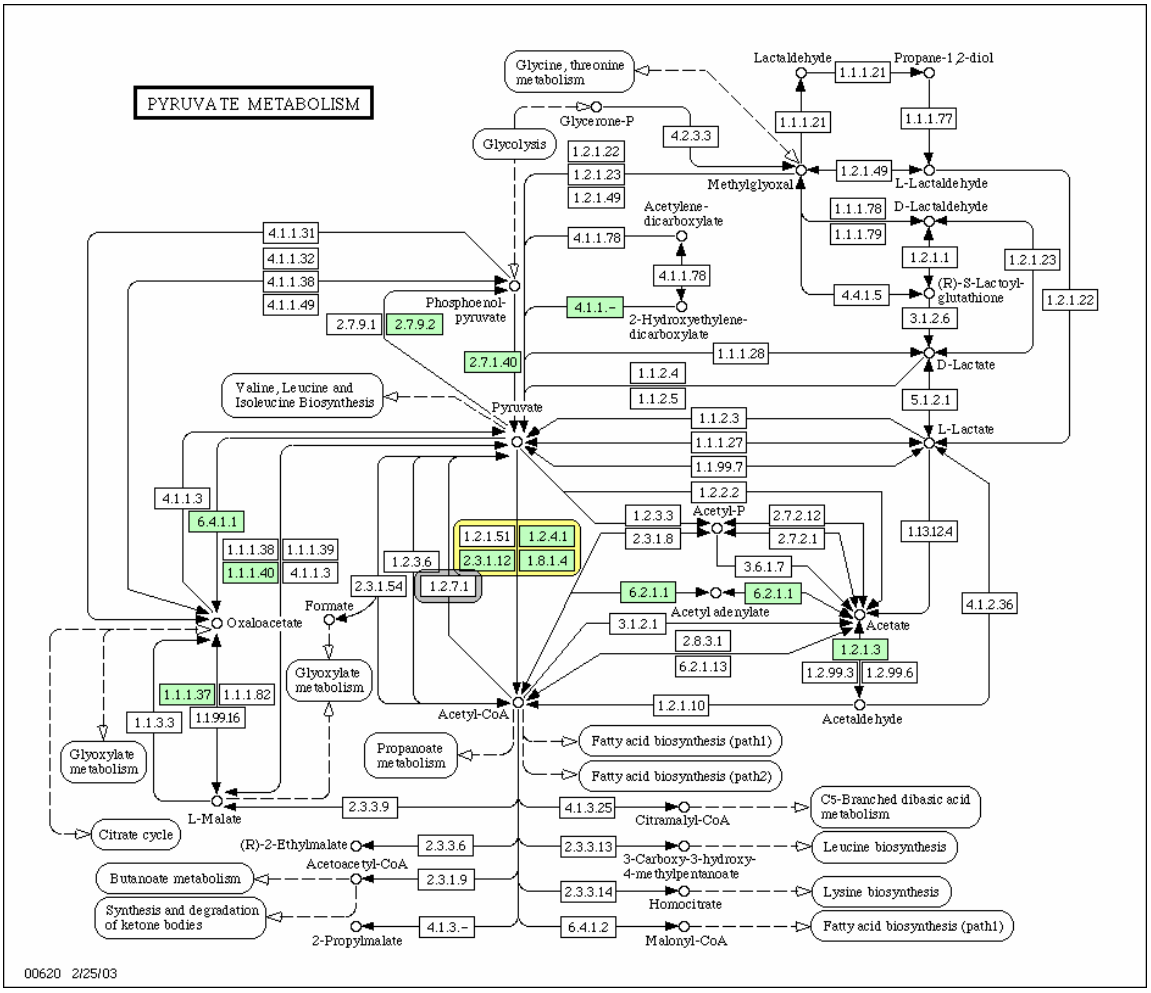


Figure 5.2: **KEGG map for the pyruvate metabolism in *Halobacterium salinarum* strain NRC-1.** Metabolic reactions for which halobacterial genes were assigned by an automatic prediction routine are coloured in green. Experimental studies already verified that oxidative decarboxylation of pyruvate is catalyzed by the ferredoxin-dependent oxidoreductase complex (EC 1.2.7.1, Por, grey box) in *Halobacterium* (Kerscher and Oesterhelt 1981). However, although *por* genes are present in the genome, the respective reaction remains unassigned in KEGG, since *por* genes were grouped into K00175, an orthologous group of another similar enzyme. Instead of Por, the canonical lipoamide-dependent pyruvate dehydrogenase complex (EC 1.2.4.1, EC 2.3.1.12, and EC 1.8.1.4, yellow box) was assigned for the conversion of pyruvate to acetyl-CoA in *Halobacterium*. However, an enzyme misassignment is likely, since it is indicated that the encoded lipoamide-dependent complex only takes part in an analogous reaction of the branched-chain amino acid pathway (Heath et al. 2004).

Enzyme assignment by automatic pathway reconstruction procedures is prone to mispredictions, though, which result in errors within organism-specific metabolic maps (Figure 5.2). Expert curation improves pathway predictions by assessment of automatic assignments and by inclusion of published experimental data available for in the given species. For *Escherichia coli* this concept of customizing automatic prediction results was realised in the intensively curated EcoCyc database (Keseler et al. 2005).

Metabolic databases comprise mainly qualitative pathway data, establishing the enzyme set of metabolic pathways that potentially exist in an organism. For a quantitative description of metabolic pathways, experimental data on enzyme kinetics, enzyme regulation, and pathway fluxes needs to be collected but is only available for few model organisms.
In this project, the metabolism of *Halobacterium salinarum* was reconstructed and stored in a metabolic database, Pathnet. This database integrates computationally-derived enzyme assignments, curated genome data, and experimental information on *H. salinarum* extracted from the literature. Halobacterial metabolic reactions and pathways were represented by metabolic maps and linked to experimental data from genome-wide transcriptomics and proteomics approaches. The established *H. salinarum* metabolism was further reviewed in this chapter.

5.2 Enzyme assignment

5.2.1 Enzyme classification

Enzymes are required as biocatalysts for most metabolic reactions. They consist of one or more protein chains which are encoded by genes and are classified by Enzyme Commission (EC) numbers. These are given by the Biochemical Nomenclature Committees (IUPAC-IUBMB) which catalogue published enzyme activities that have been observed for one or a set of reactions. The enzyme activities are assigned to a hierarchical classification system which groups them, according to the type of catalysed reaction, to one of six enzyme classes and further three subclasses, e.g. EC 4.1.2.13.
It has to be kept in mind that the EC classification is based on enzyme activities determined by enzyme assays in crude cell extracts or fractions of purified enzyme and not on the actual enzymes themselves. These are identified by purification and partial protein sequencing, which is often followed by subsequent cloning of the enzyme gene to retrieve its complete sequence. The characteristics of the isolated enzymes do not always match completely the enzyme activity described beforehand for EC classification. In most of these instances, the isolated enzyme reveals a broader specifity towards substrates and coenzymes so that it has to be described by several EC numbers. For example, prenyl transferases that are required for chain elongation of isoprenes often operate on variable prenyl chain lengths, thus, covering a set of previously defined enzyme activities (EC 2.5.1.1, EC 2.5.1.10, and EC 2.5.1.29). This enzyme classification problem is further pronounced by the fact that enzymes of different organisms often catalyze similar but not identical reactions due to differing substrate and coenzyme specifities. In case of the prenyl transferases, this is

expressed by different maximal prenyl chain lengths that are recognized by isolated enzymes of different organisms. The Enzyme Commission partly takes this versatility of enzymes into account, e.g. by the definition of five subtypes of glyceraldehyde dehydrogenases with different coenzyme specifities. The task of covering subtypes of all enzymes that evolved specifically in different organisms by individual EC numbers cannot easily be accomplished, though. Apart from this, enzyme subtypes with different (co)substrate specifities are usually highly similar on sequence level. Thus, it is not practical to designate different enzyme subtypes by individual EC numbers, since the exact specifity of an enzyme and with it its EC number cannot be concluded from the enzyme sequence anymore but can only be determined experimentally.

Beside apparent multifunctional enzymes due to broader coenzyme and substrate specifities in isolated enzymes compared to the reference enzyme activity (described by the EC-No.), there are also true multifunctional enzymes, where different enzyme activities are located in multiple catalytic centres on different parts of the protein chain (multidomain enzymes). This might for example be advantageous in order to coordinate cellular concentrations of enzymes that are involved in the same pathway. Within the multidomain enzymes, channelling of substrates from one catalytic centre to the next often occurs in order to prevents diffusion of the substrate.

It should further be noted that the same enzyme activities can be catalyzed by enzymes with no similarity on sequence level. These cases of non-orthologous enzymes are the result of convergent evolution and are often covered by the same EC number due to analogous enzyme activities. Another problem of the EC classification system is the incomplete coverage of known enzyme activities, so that enzyme activities have to be described by ambiguous preliminary EC numbers such as ‘EC 2.5.1.-‘ used for para-hydroxybenzoate-polyprenyltransferase (ubiquinone biosynthesis) but also for protoheme IX farnesyl-transferase (cytochrome synthesis). The effect of the ambiguity of partial EC numbers on the annotation of pathway databases has recently be discussed by Green and Karp (2005). On the other end of the scale, no or only one enzyme sequence was reported yet for the majority of described enzymes activities (2332 EC-No. without a single sequence entry, 338 EC-No. with one sequence in the Swiss-Prot database).

Finally, it has also to be considered for metabolic reconstruction that EC numbers might also be changed in the course of time when there is evidence for a different reaction type as previously assumed. For example, citrate (si)-synthase was reclassified from EC 4.1.3.7 to EC 2.3.3.1 and so were further 9% (395 of 4309 ENZYME database entries) of the EC numbers. Since genome annotation is usually not further updated after publication, EC number annotation has to be updated prior to the metabolic reconstruction process in order to avoid inconsistencies.

5.2.2 Enzyme assignment by similarity-based function transfer

Since enzymes are encoded by genes, EC numbers define not only an observed enzymatic activity for a metabolic reaction but can also be applied to describe functions of genes when annotating genomes. Based on the assumption that genes with similar sequences encode proteins with similar functions, function descriptions and, in the case of enzymes, also EC numbers can be transferred in large scale from known to newly sequenced genes. For this similarity-based function transfer, translated protein sequences of new genes are usually searched against public databases such as Swiss-Prot by applying blastp, and the function of the best blast hit is automatically transferred upon the query sequence. However, with the arrival of more and more new genomes and, thus, fewer proportions of described proteins with experimental verified functions available in the public databases, it is less likely to transfer an experimentally verified function. Commonly, several function transfer steps took place between the new and the originally verified gene which cannot be traced back. In worst case, there is no sequence similarity between the new gene and the verified gene after several transfer steps so that there is no functional link anymore between the new sequence and its assigned EC number so that the function assignment is questionable. However, the

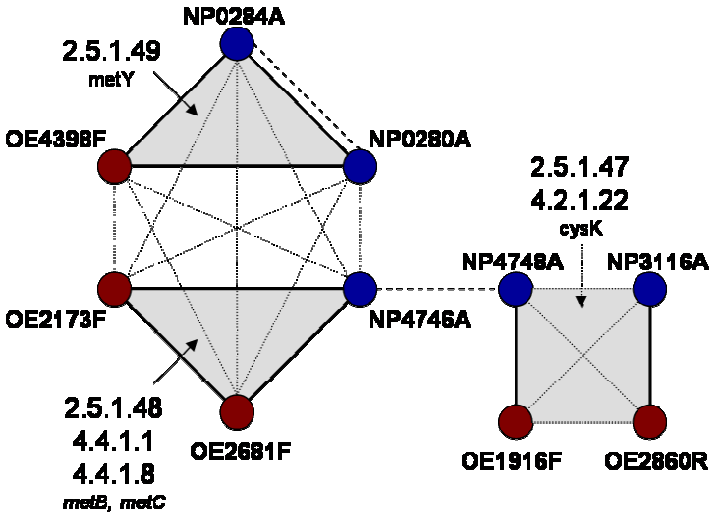


Figure 5.3: **Clustering and enzyme assignment of halophilic thiol-lyases/-synthases** (dots) from *H. salinarum* (red) and *N. pharaonis* (blue). Genes are connected according to their sequence similarity level (solid: E-value ~e^{-100}, dotted: E-value ~e^{-50}), and form two orthologous clusters, each with two subgroups of closely related genes (interconnected by solid lines). Gene pairs found adjacent in the halophilic genomes are marked by a dashed line. For the left orthologous cluster, different enzyme functions (EC-No. pointing to grey areas) could be determined for the two subclusters by the developed automatic enzyme assignment routine (Chapter 5.2.3) but not for highly similar genes (OE2173F, OE2681F). When assigning enzyme functions for genes of the right orthologous cluster, even different metabolic functions of distantly related genes (OE1916F, OE2860R) cannot be elucidated by similarity-based function transfer. Thus, the specific enzymatic functions of the different halophilic thiol-lyases cannot unambiguously be predicted and can only be determined by experimental studies.

quality of a function assignment for an annotated sequence cannot be assessed afterwards, since the source of the assignment is not given. In this project, enzyme functions were therefore determined based on sequence similarity to genes of experimentally verified enzymes whenever possible.

As mentioned above, enzyme subtypes that are described by different EC numbers (e.g. the different subtypes of glyceraldehyde dehydrogenases) might be highly similar on sequence level. This holds also true for many enzymes that apply similar reaction mechanisms. Thus, the exact enzymatic function cannot be resolved unambiguously by similarity search-based EC number transfer (Figure 5.3). However, usually only one EC number is considered in EC number transfer techniques implicating wrongly that the exact coenzyme or substrate specifity for the new gene can be predicted. Often, there is one of the alternate enzyme subtypes preferred over the others and this arbitrarily selected EC number is transferred from one newly sequenced genome to the next. Thus, it is not surprising that EC numbers for some subtypes are covered by many sequences in the databases and others only by few experimentally verified sequences. An example is the succinate-CoA ligase of which an ADP-forming (EC 6.2.1.5) and a GDP-forming (EC 6.2.1.4) subtype is known. Enzyme sequences for both subtypes belong to the same orthologous cluster (COG0074), though, and most succinate-CoA ligases show activity for ADP and GDP (BRENDA website). Although the curated Swiss-Prot database currently contains ADP- and GDP-forming enzyme sequences in the proportion of 3:1, there are 131 sequences for the ADP-forming succinate-CoA ligase (EC 6.2.1.5) but only 10 sequences for the GDP-forming succinate ligase (EC 6.2.1.4) in the uncurated TrEMBL database containing mainly sequences from complete genome projects annotated by similarity-based function transfer.

The limitations of large scale annotation that is based on sequence similarity was analysed in more detail. When automatically assigning EC numbers to *H. salinarum* genes by blast search, the number of cases for which more than one EC number can be assigned to a gene was determined (Figure 5.4). It was found that the fraction of ambiguous EC number assignments is reduced with stricter E-value cutoffs but remains at 26% even for likely assignments (E-value better than e^{-50}). Further, analysis of these ambiguous cases showed that alternate EC numbers vary mainly in the last position of the EC number in case strict E-value cutoffs were applied. These reflect cases of highly similar enzyme subtypes with different coenzyme or substrate specifity as the mentioned glyceraldehyde dehydrogenases and prenyl transferases, respectively. However, when reducing E-value cutoffs for EC number assignments, many ambiguous assignments occur where candidate EC numbers vary already in EC number positions one to three. This means that specific EC-No. cannot be assigned reliably anymore, and highlights the problem of transferring EC numbers by

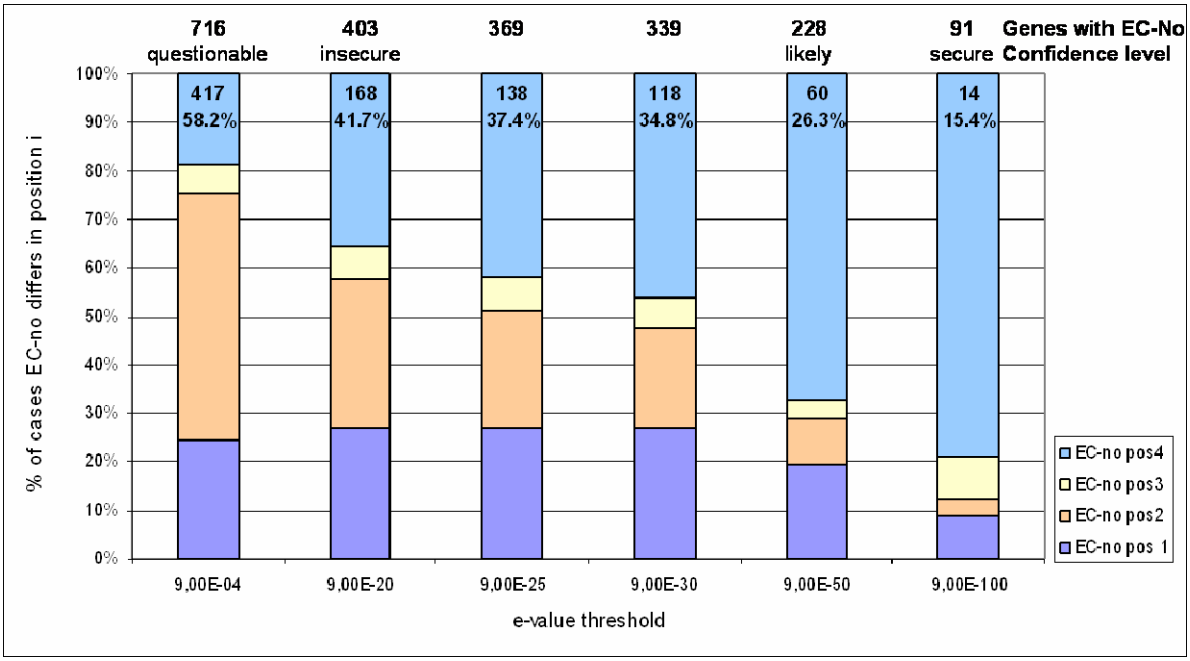


Figure 5.4: **Ambiguous EC number assignments for *H. salinarum* genes by blast searches.** The total number of genes with EC-No. assignments for the given E-value cutoff is indicated above columns while the number and fraction of genes with ambiguous assignments is noted within the columns. The ambiguous EC-No. assignments were further analysed by comparison of alternate EC numbers. With stricter E-value cutoffs EC numbers vary mainly in the 4. position of the EC numbers, which means that coenzyme and substrate specifity cannot be resolved by similarity search anymore. For moderate E-value cutoffs, EC numbers differ commonly in the 1.-3. EC-No. position, so that EC number misassignments and apparent pathway gaps occur frequently in case of choosing one of the alternate EC numbers for sequence annotation.

automatic means. Since the E-value cutoff for function assignments in published genomes is often around e^{-20} or even worse, exact function and enzyme assignments of genes has to be questioned. This is an important observation, since metabolic reconstruction is based on these enzyme assignments. When choosing an EC number for a new gene, a pathway step is subsequently designated as present while the reaction for the alternate EC number that was not chosen is considered as absent (pathway gap). In case EC number assignments are not reliable anymore, pathway reconstruction is error-prone as illustrated for the pyruvate metabolism of *Halobacterium* (Figure 5.2 in Chapter 5.1).

In conclusion, function annotation of newly genomes by assigning one function to a gene is inadequate for metabolic reconstruction, since enzymes with broad substrate specifity require several EC number assignments. Furthermore, function transfer techniques are limited in resolving exact EC numbers for enzyme subtypes or even for enzymes that are involved in completely different pathways, e.g. the different 2-oxoacid dehydrogenase complexes. In these cases, alternate EC numbers have to be considered in order to avoid EC number misassignments.

5.2.3 Development of an enzyme assignment routine

For metabolic reconstruction of *H. salinarum*, a new EC number assignment routine was developed. In contrast to the usual EC number prediction procedures, *Halobacterium* genes were not searched against public databases that contain enzyme sequences but searches were performed *vice versa*. For each EC number, a set of enzyme sequences was selected from the Swiss-Prot database and searched against the *H. salinarum* database (Figure 5.12 in Methods). Thus, all EC numbers with available enzyme sequences were considered equally and enzyme subtypes that are most abundant in public databases were not preferred.

In order to improve the reliability of EC number assignments, several similarity-based search strategies were applied. First, the Swiss-Prot sequence set for each EC number was searched against the *H. salinarum* database by blastp (BLAST search) so that the EC number could be assigned to the best hits in *H. salinarum* (see Methods for details). Secondly, the selection of Swiss-Prot sequences was used to create a HMMER profile for each EC number, which was searched with against *H. salinarum* (HMMER search). As described in Methods, two further similarity searches (COG search, PFAM search) were performed and E-values of all four searches combined to a total score (Figure 5.13 in Methods). Thus, for each EC number with known enzyme sequences (46% of the 4309 EC-Numbers) candidate genes could potentially be identified in the *H. salinarum* genome. Based on the total score of all similarity searches, 511 enzyme functions (EC numbers) were assigned to 359 *H. salinarum* genes for an E-value cutoff of e^{-20} when allowing alternate EC number annotations. Confidence levels for enzyme assignments were not solely based on the total score, though, but depend amongst other criteria also on the analysis whether an EC number is the best hit for a gene and *vice versa* (EO flag in Table 5.1).

The automatically-retrieved enzyme assignments were used for metabolic reconstruction of *H. salinarum* (Chapter 5.4) but were thereby manually assessed. On the one hand, this is necessary in order to check sets of alternate EC numbers predicted for a gene. When scores for the best and second best EC numbers differ only slightly, e.g. for similar subtypes of an enzyme, the EC-No. ambiguity was taken into account by assigning more than one EC number (Table 5.1A). On the other hand, an EC number might also be assigned to several genes in *H. salinarum* (Table 5.1B-D), e.g. frequently when subunits of a probable enzyme complex exist, in case of paralogous genes whose different functions cannot be resolved by similarity search (Figure 5.3 in the previous section), and sometimes for cases of non-orthologous enzymes.

Finally, cases of multidomain proteins have to be assessed carefully, since enzyme domains fused in one organism might occur separately or with differing domain order in another organism (domain shuffling, Figure 5.9 in Chapter 5.6.2). When applying usual function

Table 5.1: **Selected results of the developed enzyme assignment routine.** For each EC number - gene pair, it was marked by an EO flag whether this was the best assignment in respect to the enzyme (E-), the gene (-O) or both (EO). Scores were determined by E-value transformation, e.g. score 20 for e^{-20}. For a gene there might be alternate EC numbers with high scores which cannot be resolved similarity search **(A)**. On the other hand, the same enzyme function might be assigned to several genes in case of subunits of complexes **(B)**, paralogs **(C)**, or non-orthologous enzymes **(D)**. The EO flag indicating the best EC-No. - Gene-Id assignment is helpful to distinguish between reliable (EO) and unreliable (--) enzyme assignments of less conserved sequences **(E)**.

	EC-No.	Gene-Id	Confidence level	Best EC-OE combination	Total score	BLAST score	COG score	HMMER score	PFAM score
A:	*1.1.1.42*	OE3634F	secure	EO	121.1	129.0	115.0	114.8	125.6
	1.1.1.41	OE3634F	insecure	E-	49.9	49.0	-2.5	27.5	125.6
	1.1.1.85	OE3634F	insecure	E-	46.2	39.7	-2.5	22.2	125.6
B:	*4.1.1.21*	OE1951F	likely	EO	39.6	63.7	69.7	-2.5	27.6
	4.1.1.21	OE1952F	insecure	-O	21.2	26.2	26.3	-2.5	35.0
C:	*2.1.1.131*	OE3216F	likely	EO	39.6	49.5	-2.5	84.3	27.1
	2.1.1.131	OE3214F	insecure	-O	23.9	29.7	-2.5	40.1	28.1
D:	*5.3.3.2*	OE6213R*	secure	EO	61.9	91.5	-2.5	154.7	4.0
	5.3.3.2	OE7093R*	secure	-O	61.9	91.5	-2.5	154.7	4.0
	5.3.3.2	OE3560F	possible	-O	10.1	17.4	16.0	-2.5	9.5
E:	*2.7.4.16*	OE3818F	insecure	EO	13.1	26.7	-2.5	25.1	3.0
	2.7.7.9	OE1014R	insecure	--	16.0	27.3	-2.5	-2.5	41.6

* - identical genes
EC-No. - EC title (total/archaeal sequences in Swiss-Prot): EC 1.1.1.42 - Isocitrate dehydrogenase (NADP+) (42/2);
EC 1.1.1.41 - Isocitrate dehydrogenase (NAD+) (17/0); EC 1.1.1.85 - 3-isopropylmalate dehydrogenase (135/5);
EC 4.1.1.21 - Phosphoribosylaminoimidazole carboxylase (43/8); EC 2.1.1.131 - Precorrin-3B C(17)-methyltransferase (2/0);
EC 5.3.3.2 - Isopentenyl-diphosphate delta-isomerase (80/18); EC 2.7.4.16 - Thiamine-phosphate kinase (12/5);
EC 2.7.7.9 - UTP--glucose-1-phosphate uridylyltransferase (36/1).

transfer routines based on sequence similarity, functions of multidomain proteins are commonly automatically transferred to genes regardless if they exhibit the same or different domain topology. This further adds to EC number misannotations in public databases.

5.2.4. Enzyme classification in the KEGG database

The Kyoto Encyclopedia of Genes and Genomes (KEGG) contains a selection of databases connected to metabolism, most importantly a collection of over 6300 metabolic reactions. In contrast, the ENZYME collection denotes only around 3900 classified enzymes, most of them assigned to only one metabolic reaction. However, for circa 5400 and 400 of the KEGG reactions, one or more EC numbers are denoted, respectively. This shows that KEGG applies EC numbers in a broader context compared to the original enzyme classification in order to cover a greater part of the metabolic network with EC numbers. However, the basis that leads to the assignment that a certain enzyme participates also in an additional reaction is unclear, and it is also unknown how consistent the assignment procedure was carried out.
KEGG further increases the coverage of the metabolic network with EC numbers by using partial EC numbers (e.g. EC 3.1.3.-) for reactions which are yet to be described by the

Enzyme Commission and by using EC numbers in several pathway contexts. For example, a reaction in the pyridoxal-phosphate biosynthesis pathway starting from erythrose 4-phosphate is covered by an enzyme that has been originally assigned to an analogous reaction which is part of the serine biosynthesis pathway (EC 2.6.1.52, KEGG-Map00750). Furthermore, highly similar enzyme subtypes for different coenzyme and substrate specifities seem to be grouped which is recommendable due to their similarity on sequence level and the resulting limitations of similarity-based function transfer (Chapter 5.2.2). For example, reaction R01061 ($NADP^+$-specific conversion of glyceraldehydes 3-phospate) is linked to the glyceraldehyde dehydrogenase subtypes, EC 1.2.1.12, EC 1.2.1.13, and EC 1.2.1.59, although EC 1.2.1.12 was designated to be only NAD^+-specific. Also the different prenyl transferases for different prenyl chain length (EC 2.5.1.1, EC 2.5.1.10, EC 2.5.1.29) were assigned together to the respective reactions.

In conclusion, KEGG has an extensive collection of metabolic reactions, and EC numbers have been assigned to 86% of these reactions. However, it is not clear which rules were applied to establish this high coverage of the KEGG reaction network with EC numbers and enzyme genes. Probable inconsistencies in assignments might therefore result in 'noisy' metabolic reconstructions and models.

5.3 Database structure and implementation of a metabolic database

In order to develop a metabolic database for *H. salinarum* (Pathnet), an appropriate database model was established. For this, the main objects of a metabolic pathway/network, reactions, compounds, enzymes, and corresponding genes as well as their relationships amongst each other, were analysed.

A metabolic reaction is defined by a chemical equation which sets participating chemical compounds in a context, e.g. '2 trans,trans-farnesyl diphosphate <=> pyrophosphate + presqualene diphosphate + H+' (KEGG reaction: R00702). To obtain a main equation ('trans,trans-farnesyl diphosphate <=> presqualene'), side substrates (diphosphate, H+) are omitted. Furthermore, the main direction of the reaction can be indicated by arrows (<=, =>, <=>). In the Pathnet database, a reaction is represented by several entries, one reference entry in the reactions table containing the total and the main equation of the chemical reaction, and organism-specific entries in the *org_reactions* table (Figure 5.5, entry examples for some Pathnet tables are given in Table 5.2). For an organism-specific reaction, it can be established with certain confidence whether the reference reaction exists in the given organism.

The vast collection of KEGG reactions was used to fill the reference *reactions* table, in which total equations and main equations were defined (for details see Supplemental Table 5.8). Since chemical compounds may have several synonymous names (e.g. farnesyl-PP, farnesyl-diphosphate), KEGG equations include unique compound identifiers rather than compound names. The KEGG compound collection was stored in the table *compounds*, and is used to generate user-readable chemical equations. Around 70% of the KEGG reactions are linked to static KEGG maps that represent metabolic pathways/topics. Here, reactions were also linked to defined 'textbook' pathways (Supplemental Table 5.7) (Michal 1999) that were grouped to five metabolic subsystems (Table 5.3 in the next subchapter).

Since KEGG-based *reactions* required corrections and modifications due to data inconsistencies, several columns of the *reactions* table were duplicated, one storing the original KEGG data (columns prefix: 'orig'), the other the modified data. Thus, KEGG reactions and compounds collections can easily be updated in the future without interfering

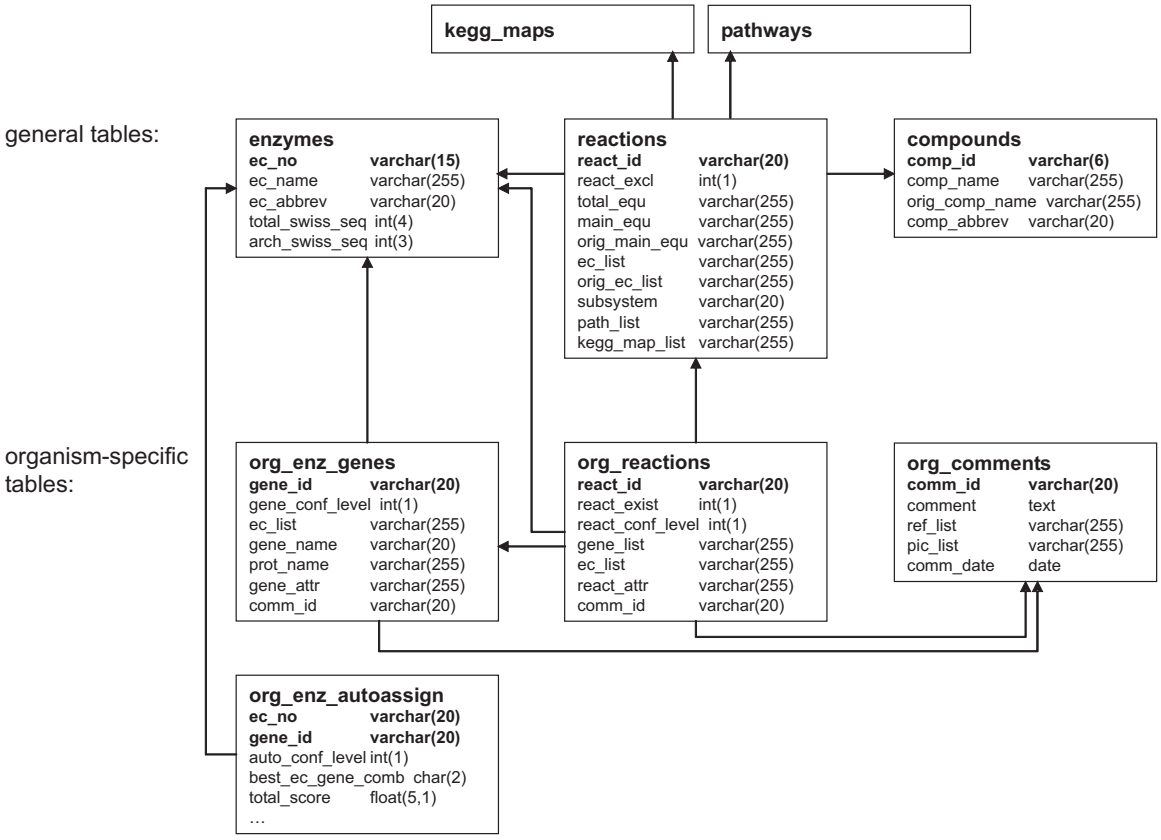


Figure 5.5: **Structure of the metabolic database Pathnet.** The tables *enzymes*, *reactions* and *compounds* contain general information connected to pathways that was obtained from KEGG ligand and Swiss-Prot databases (for details see Supplemental Table 5.8). Entries of *org_enz_genes* and *org_reactions* tables store manual pathway reconstruction data for a given organism, whereas the *org_enz_autoassign* table contains automatic enzyme assignments for genes (Chapter 5.2.3). The table *org_comments* stores additional texts connected to varying tables and links to literature references. Pathway lists (*kegg_maps*, *pathways*) are currently stored externally. Entry examples for some Pathnet tables are given in Table 5.2.

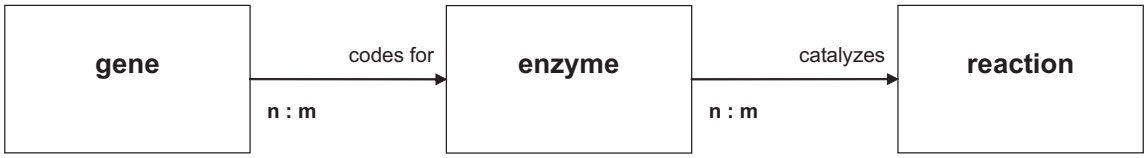


Figure 5.6: **Relationships of central pathway objects.** Metabolic reactions of a pathway are catalyzed by enzymes. These consist of protein chains encoded by genes on the genome. As discussed in the text, the relations between the three pathway objects are ambiguous in all directions.

with curated data of the *reactions* entries. Curation is especially required for KEGG main equation definitions, which are incomplete (only available for circa three quarters of the KEGG reactions) and error-prone. This is a severe problem, since metabolic models might be based on a reaction set excluding side substrates. Another problem arises from the fact that the direction of a reaction might differ from species to species. In these cases, a set of main equations should be defined for one reaction. In future, it should be estimated whether a significant amount of reactions show different directions in varying organisms. As a consequence, main equations might then be included in the organism-specific *org_reactions* tables. In the current version of Pathnet, modified main equations and their directions are included in the general *reactions* table and relate to *H. salinarum* and other haloarchaea.

Metabolic reactions are usually catalyzed by enzymes, which are encoded by genes in any organism. For the general *enzymes* table of Pathnet, a collection of reference enzymes defined by EC numbers was retrieved from the ENZYME database of Expasy. As described above, EC numbers that define enzymatic reactions are also commonly used as identifiers of enzymes or enzyme genes when annotating genomes. Thus, EC numbers establish a connection between the set of genes found in a genome and the set of chemical reactions. For each reference enzyme, the number of total and archaeal protein entries in the Swiss-Prot database was ascertained by selecting entries containing the according EC-No. in their protein name description (Supplemental Table 5.7, Table 5.2).

For the Pathnet database model, the relationships between the central objects reactions, enzymes, and genes were analysed and limitations in the availability of metabolic data for certain pathways were considered in order to obtain a serviceable database. All relationships between reactions, enzymes, and genes were found to be ambiguous (n:m relations) (Figure 5.6), so that relations between reactions and genes are often complicated making metabolic reconstruction a rather difficult task. A metabolic reaction is catalyzed by one enzyme defined by an EC number. However, KEGG applies EC numbers more flexibly, so that over 400 KEGG reactions are linked to several EC numbers (Chapter 5.2.4). Vice versa, an enzyme might take part in different reactions in case of broader substrate specifity. Again, it should be noted that KEGG might use an EC number for more reactions as originally defined by the Enzyme Commission.

An enzyme consists of one or, in case of enzyme complexes, of multiple protein chains which are encoded by genes (Table 5.1 in Chapter 5.2.3). This ambiguity between enzymes and genes is furthermore caused by paralogous genes to which the same enzymatic function can be assigned, since their differing specific function cannot be resolved by similarity search but only by experiments. For example, two paralogous genes potentially encoding the archaeal type of fructose-bisphosphate aldolase (EC 4.1.2.13, OE2019F, OE1472F) have been observed in *Halobacterium*. Added to this, non-orthologous genes might occur, which encode enzymes with the same activity. In *H. salinarum*, genes for both, the type I isopentenyl-biphosphate (IPP) isomerase (EC 5.3.3.2) and the type II IPP isomerase were found (Table 5.1D in Chapter 5.2.3). When assigning genes to enzymes the situation might also be ambiguous. Although an enzyme gene usually codes for one particular enzyme activity, some enzymes such as prenyl transferases can be described by more than one EC number due to limitations of the enzyme classification system (Chapter 5.2.1). Furthermore, enzymes can fuse to multidomain proteins with several catalytic centres.

Taking the ambiguity between the three central metabolic objects into account, a database model with three single organism-specific tables for *reactions*, *enzymes*, and their *genes* (defined by unique primary keys) are required. These tables would be connected by two linking tables with *gene-enzyme* and *reaction-enzyme* key combinations. However, such a model cannot be established yet, since enzyme activities are not known for all metabolic reactions and not all pathways are completely understood yet. Especially in the biosynthesis of coenzymes, pathway steps or even coenzyme precursors are not known. In case of sparse knowledge of a pathway, possible reactions that might bridge pathway gaps need to be postulated in order to interconnect all reactions of the metabolic network. A significant amount of observed enzyme activities was also not classified yet by a unique EC-No. (the Enzyme Commission requires publications with fully characterized enzyme activities). Finally, there are also spontaneous reactions in cells which do not require a catalyst, e.g. the pyrroline ring formation of L-glutamate 5-semialdehyde (proline synthesis, R03314).

Due to limited data for certain parts of the metabolic network, reactions and enzymes cannot always be interlinked unambiguously, which is necessary for a functional database. Therefore, the Pathnet database structure contains only two organism-specific tables *org_reactions* and *org_enz_genes*, each of them connected to the reference *enzymes* table by EC-No. lists if possible. For organism-specific *org_reactions* and *org_enz_genes* entries, an attribute list can be given indicating features of the entry (Supplemental Table 5.7, Table 5.2). Metabolic reactions and enzyme-encoding genes for a given organism are directly linked by a gene list given in the *org_reactions* table. As stated above, *org_reactions* establishes whether a reference reaction exists in an organism with certain confidence, and

so confidence level was assigned for enzyme genes entries stating the reliability of the EC number annotation.

In both tables, *org_reactions* and *org_enz_genes*, the attribute 'exper' marks entries for which experimental data were extracted from the literature (details are then specified in the *org_comments* table for the organism). Metabolic reactions can be experimentally validated through NMR or other labeling studies, but also through enzyme activity tests. For genes, mutagenesis experiments might be available, or protein-sequencing of isolated enzymes might verify its predicted function. Prokaryotic genes are often part of transcription units, which might have also previously been studied. A list of literature references, mainly in form of PubMed identifiers is given for *org_comments* entries.

Table 5.2: **Metabolic annotation for a selected enzyme (EC 1.1.1.6) in *H. salinarum*.** Tables A-C contain data that contributed to the metabolic reconstruction. Tables D and E present the resulting reconstruction data stored in the organism-specific *reactions*, *enzyme genes* and *comments* tables, which can be inserted and updated via a web-based form (Müller 2005).

A) General reactions

react_id	R01034
ec_list	1.1.1.6;
react_excl	2
total_equ	Glycerol + NAD+ <=> Glycerone + NADH
main_equ	Glycerol <=> Glycerone
subsystem	C
path_list	glycerol_met;

B) Hasal basicdata

gene_id	OE5160F
gene_name	gldA1
prot_name	glycerol dehydrogenase (EC 1.1.1.6)
function class	CIM
comment	-

C) Hasal enzyme autoassign

gene_id	ec_no	auto_conf_level	best_ec_gene_comb	total_score	blast_score	cog_score	hmmer_score	pfam_score
OE1602F	1.1.1.6	3 (possible)	--	27.8	12.0	96.4	-2.5	5.1
OE4036R	1.1.1.6	1 (unlikely)	--	0.7	-0.3	-2.5	-2.5	7.9
OE5160F	1.1.1.6	6 (secure)	EO	92.7	90.7	75.4	154.8	49.7

D) Hasal reactions

react_id	R01034
react_exist	1 (yes)
react_conf_level	6 (secure)
gene_list	OE5160F;
ec_list	1.1.1.6;
react_attr	exper; manual; few_seq; gap_arch;
comm._id	L00068
comment	pos_enz_activ; high glycerol dh activ. found in H. salinarum, but no EA in H. mediterranei and H. vallismortis; H. salinarum only arch. with this enz.;
ref_list	PMID:3255682;
pic_list	-

E) Hasal enzyme-encoding genes

gene_id	OE5160F
gene_conf_level	6
ec_list	1.1.1.6;
gene_name	gldA1
prot_name	glycerol dehydrogenase (EC 1.1.1.6)
gene_attr	exper; manual; autom;
comm_id	L00004
comment	struct_x-ray (PhD thesis, S. Offermann, 2003); EC 1.1.1.6 more likely than EC 1.1.1.1/ 1.1.1.2/ 1.1.1.261 by sim. search;
ref_list	-
pic_list	-

For *H. salinarum*, an additional table was created (*org_enz_autoassign*) which stores *enzyme-gene* combinations which have been automatically assigned as described in Chapter 5.2.3. These computationally-derived enzyme assignments contributed to the creation of manually-assessed *org_reactions* and *org_enz_genes* entries.
In conclusion, the metabolic database Pathnet distinguishes between reference and organism-specific data for metabolic pathways (Figure 5.5). Using the pool of reference reactions and enzymes retrieved from external databases as well as genome and enzyme assignment data, organism-specific reactions and enzyme-encoding genes can be created and linked to literature data. Thus, a curated metabolic network of the organism (*H. salinarum*) required for pathway mapping of experimental data and for metabolic modelling can be stored.

5.4 Reconstructing metabolic pathways of *Halobacterium salinarum*

The metabolism of *H. salinarum* was reconstructed using a web interface that integrates genome annotation data from the Hasal database (Halolex) with enzyme assignment and metabolic data from the Pathnet database (see Methods). For reconstructing the metabolism of *H. salinarum*, the focus was set on central dissimilatory pathways such as glycolysis and citrate cycle, and assimilatory pathways for the synthesis of main metabolic compounds such as proteinogenic amino acids, nucleotides, lipids and coenzymes. Reconstructed halobacterial pathways are reviewed in Chapter 5.6.
Table 5.3 and Supplemental Table 5.7 give a statistical overview of the created Pathnet entries for each of the defined metabolic subsystems and pathways. In total, 514 *org_reactions* entries were created for *H. salinarum* (as well as 161 *org_reactions* entries for *Natronomonas pharaonis* (see Chapter 6.2)), for which it was assessed whether the reaction exists or does not exist in the halophile. Over 300 of the assessed reactions were set to exist in *H. salinarum*, and for 86% of these existing reactions, genes could be identified within the complete genome sequence. In case a reaction exists but lacks genetic evidence, the decision was reached based on positive experimental data, mainly enzyme activity tests, reported in the literature. However, due to repeated renaming of halobacterial strains within the last decades, these experiments might not always be performed with the sequenced *H. salinarum* str. R1. The pathway database for *Halobacterium* contains 236 *enzyme genes* and lists 405 EC Numbers. Further, 226 *comments* were created which describe the created *org_reaction* and *org_enz_gene* entries in more detail. All available literature data regarding the metabolism of *H. salinarum* were included in the Pathnet database. These comprise

Table 5.3: **Statistics of the reconstructed metabolism for *H. salinarum*.** An overall entry statistics of *H. salinarum*-specific Pathnet tables and statistics for individual subsystems is given. The number (#) of *reactions* (R), *enzyme genes* (GEN), and *comments* (CMT) entries is listed. *H. salinarum reactions* were distinguished into present (exist) and absent (!exist) reactions. The number of experimentally validated *reactions* with no genetic evidence (!gene) is indicated. Further, numbers of listed enzymes (ENZ) and literature references (LIT) are given. The created Pathnet entries for *H. salinarum* are discussed in Chapter 5.6. Subsystems: C - central intermediary metabolism, P – pyrine/pyrimidine synthesis and metabolism, L – lipid synthesis, A – amino acid synthesis and metabolism, V – vitamin/coenzyme synthesis. An analogous table with a statistical overview of Pathnet entries for individual pathways is presented in Supplemental Table 5.7.

Subsystem	**# R_{exist}**	**# $R_{!exist}$**	**# $R_{!gene}$**	**# GEN**	**# ENZ**	**# CMT**	**# LIT**
C	64	48	9	52	110	110	20
P	48	17	1	34	42	13	2
L	29	0	1	22	23	15	6
A	79	77	1	71	127	60	15
V	100	51	31	61	109	32	0
all reactions	**320**	**194**	43	**236**	405	**226**	39

information from pathway experiments (labeling studies) and functional studies of enzymes, *genes* and transcription units. Thus, the Pathnet database is a valuable resource for *H. salinarum* literature data.

Pathway reconstruction data for *H. salinarum* can be utilized for the interpretation of proteomics and transcriptomics data. The Pathnet database links *enzyme genes* to metabolic *reactions*, *enzymes* and pathways. Therefore, identified/regulated *genes* from 'omics'-techniques can be mapped to metabolic pathways. By representing 'omics'-data on metabolic maps (Chapter 5.5), the data analysis was assisted, and, especially for larger gene sets, an overview regarding the regulation of the *H. salinarum* metabolism was gained.

Pathnet was also developed to enable access of *H. salinarum reaction* data for future metabolic modelling. Since all organism-specific reactions are linked to general KEGG-based reactions entries with chemical equations and compounds, the generation of input formats for modelling software should be straight forward. The definition of smaller metabolic units namely pathways, KEGG maps and metabolic subsystems, will be of use to extract a reaction subset for 'bottom-up' modelling approaches as performed by Metatool (Pfeiffer et al. 1999). Further, confidence levels and comments will be useful for evaluating single parts of metabolic models, and subsequently aid the modification of the metabolic reconstruction and models for *H. salinarum*.

5.5 Graphical representation of metabolic pathways

Reconstructed metabolic pathways of *H. salinarum* are presented through static pictures (e.g. Figures 5.8-5.11) available on the Halolex website. Furthermore, metabolic pathway graphs can be generated on demand by a graphical program as described in Methods. The so created pathway graphics consist of KEGG reference maps with customized colouring of enzyme objects (Table 5.6 in Methods). Colours of each enzyme box or label for a chosen KEGG map depend on the reconstruction data selected from the Pathnet database, which are shown via links on the created coloured KEGG map.

The developed graphical program proved to be useful in several aspects; for monitoring of the reconstruction process, for representation of reconstruction data for a pathway, and for mapping genome-wide experimental data from 'omics'-techniques to pathway data. While reconstructing a metabolic pathway in *H. salinarum* via the developed web interface, an overview of existing reactions and pathway 'gaps' is gained by KEGG map colouring (option: reconstruction data). Reactions for which experimental results are available are also indicated, which will support experimental design of future metabolic studies as well as the modification of metabolic models, since direct access to the *H. salinarum* literature is given.

The graphical program was additionally extended in order to map genes with EC number annotations from complete genome sequences (option: genome annotation). It has applied to the genome annotation of *H. salinarum*, *N. pharaonis*, and recently sequenced *Haloquadratum walsbyi*, but can potentially be used for any sequenced organism. However, the significance of the resulting coloured KEGG map strongly depends on the quality of the enzyme annotation of genes.

The developed graphical program permits flexible integration of any type of biological data into metabolic maps. Within this project, the graphical program was adapted in order to map 'omics' data to metabolic pathways. Proteins identified through proteomics as well as regulated genes derived from transcriptomics and quantitative proteomics experiments were linked to pathway data (options: (regulated) proteomics, transcriptomics) (Figure 5.7). From the set of identified or regulated genes, *enzyme genes* were selected which are linked to certain metabolic *reactions* within the Pathnet database. *Enzyme genes* found in 'omics' studies were then colour-marked in the generated KEGG map. In case two environmental conditions were compared in the experiment, up- (green) and down- (red) regulations are indicated, thus, conclusions can be drawn whether a metabolic reaction/pathway is regulated upon change in growth conditions.

Coloured KEGG maps combining metabolic reconstruction and regulation data can also reveal data inconsistencies. This is evident for linear pathways where only one gene is regulated in a different direction, or for subunits of a complex which are regulated differently

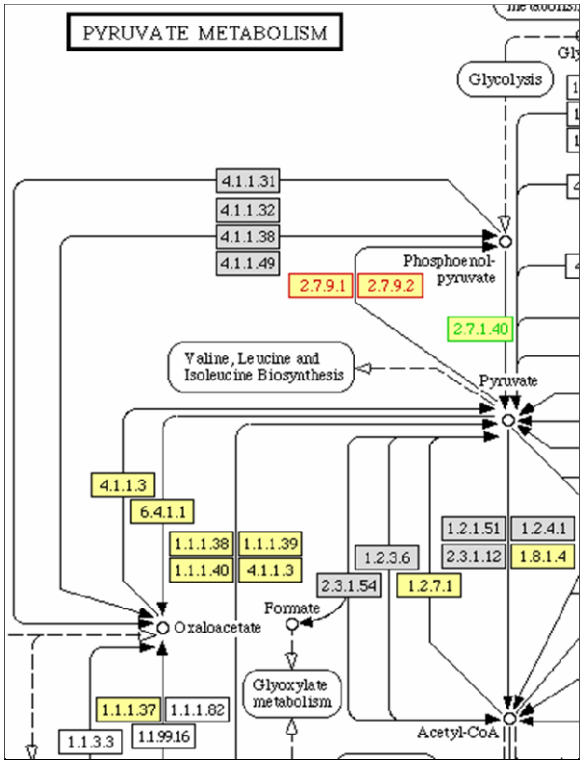


Figure 5.7: **Pyruvate metabolism map with linked regulation data for *H. salinarum*** (for the complete KEGG map see Figure 5.2). Regulation data were derived by quantitative proteomics using the ICPL technique (A. Tebbe, pers. comm.). Protein levels of *H. salinarum* cell cultures grown in rich medium were compared with levels observed in synthetic medium. Among regulated proteins pyruvate kinase (EC 2.7.1.40) and pyruvate, water dikinase (EC 2.7.9.2) were found. The two enzymes catalyze reverse reactions involved in the glycolysis and gluconeogenesis of *H. salinarum* (Figure 5.8), respectively. Results show that the catabolism is more active when cells grow in synthetic medium which is expressed in upregulation of the pyruvate kinase (green label, OE1495R, regulation factor: 2.09) producing pyruvate that is fed into the citrate cycle. On the other hand, pyruvate, water dikinase is downregulated (red label, OE1500R, regulation factor: -3.16) in order to prevent a 'futile cycle'. The coloured KEGG map was derived by the developed graphical program (option: 'regulated proteomics').

Background colours of EC-boxes: yellow - *reaction* present in *H. salinarum*, grey - *reaction* absent in *H. salinarum*, white - *reaction* not considered. The complete colouring scheme of the developed graphical program is listed in Table 5.6 in Methods.

(Table 5.4A). However, apparent inconsistencies of regulation data can arise from metabolic reconstruction ambiguities. Especially for paralogous genes, e.g. thiol-lyase (Figure 5.3 in Chapter 5.2.2) or aldehyde deydrogenase genes (Table 5.4B), functional differences of the paralogs cannot be resolved by similarity search, and, thus, the complete paralog set has to be linked to the same metabolic reaction. As a result of the necessary ambiguous gene assignment to a reaction, regulation data appears to be inconsistent, in case different regulation occurs for the paralogs. However, detecting apparent inconsistencies of regulation data might assist to find distinct metabolic functions of paralogous genes. In case all steps of a pathway are regulated in one direction, it can be assumed that the paralog, which is regulated in the same direction, catalyzes the remaining pathway step.

5.6 The metabolism of *Halobacterium salinarum*

Results of the metabolic reconstruction for this species were summarized to give an overview about the metabolic capabilities of *H. salinarum* strain R1. Details and manually-drawn pathway maps for the described metabolic pathways and subsystems are available via the Halolex website ('Pathway tools' link).

Table 5.4: **Example of data inconsistency represented in coloured KEGG maps.** Regulation data for *enzyme genes* of *H. salinarum* were derived by transcriptomics experiments with DNA arrays (J. Twellmeyer, pers. comm.). Presented data was selected from the Halolex transcriptomics database (fs version) for experiment I. Transcription levels for cultures grown in the dark were compared with phototrophic cultures of *H. salinarum*.
(A) A gene cluster was found in the *H. salinarum* genome that encodes four subunits of the probable succinate dehydrogenase complex. However, regulation of the four *sdh* genes differs.
(B) The two halobacterial paralogous genes of fructose-bisphosphate aldolase are regulated differently. In contrast to the annotated gene functions, regulation results indicate that the encoded enzymes do not fulfil same functions in *H. salinarum.* Further experiments might clarify which of the two candidate genes encodes a functional fructose-bisphosphate aldolase.

Gene	Protein name	Regulation	Ratio	P-value	Rank
A: EC 1.3.99.1 Succinate dehydrogenase					
OE2865R	chain A (flavoprotein)	*unreg*	1.011	0.99	2269
OE2866R	chain B (iron-sulfur protein)	*up*	1.132	0.74	604
OE2867R	chain D (membrane anchor protein)	*unreg*	0.918	0.62	323
OE2868R	chain C (cytochrome b-556)	*down*	0.794	0.64	388
B: EC 4.1.2.13 Fructose-bisphosphate aldolase					
OE2019F	isoenzyme 1	*up*	1.237	0.35	101
OE1472F	isoenzyme 2	*unreg*	1.009	0.99	2372

5.6.1 Central intermediary metabolism

Carbohydrate-utilizing halophiles such as *Haloferax mediterranei* and *Haloarcula vallismortis* have been reported to catabolize fructose (Frc), mannitol, and sucrose via a modified Embden-Meyerhof (EM) pathway involving ketohexokinase (EC 2.7.1.3) and 1-phospho-fructokinase (EC 2.7.1.56) (Altekar and Rangaswamy 1992). These halophiles are further capable of glucose (Glc) degradation via the semi-phosphorylated Entner-Douderoff (ED) pathway (Danson and Hough 1992). Reconstruction of all three glycolytic routes, the EM, ED and the pentose phosphate (PP) pathway, showed that *H. salinarum* is likely lacking the ability to catabolize glucose (Figure 5.8). Degradation of other sugars is also unlikely in spite of the presence two sugar kinase homologs (OE4535F, OE3606R) in the *H. salinarum* genome.

Genes for 6- and 1-phosphofructokinase (Pfk), key enzymes of the hexose part of the classic (Glc) and modified (Frc) Embden-Meyerhof pathway, are absent in the *Halobacterium* genome. New archaeal types of 6-Pfk depending on ADP (*Thermococcus, Pyrococcus*) or PP(i) (*Thermoproteus*) as co-substrate instead of ATP have not been found either, which is consistent with the fact that 6-Pfk activity was not detected in halobacterial cell extracts (Rawal et al. 1988). However, the triose part of the EM pathway leading to pyruvate is functional as all required genes were found. Labeling (Ghosh and Sonawat 1998) and enzyme activity studies (Rawal et al. 1988; D'Souza and Altekar 1998) have shown a complete reverse EM pathway (gluconeogenesis) for *H. salinarum* which is required for the synthesis of hexoses from intermediate compounds. Labelled glucose was found to be

incorporated into different sugar moieties (Glc, Man, Gal) of halobacterial glycolipids (Weik et al. 1998). Furthermore, saccharide units were found to be attached to halobacterial surface proteins such as the S-layer protein and flagellins (Sumper 1987). Sugar moieties of lipids and proteins are likely synthesized via nucleotide sugars, and in accordance with this several nucleotide sugar enzymes such as UDP-glucose 4-epimerase (EC 5.1.3.2) and UDP-glucose 6-dehydrogenase (EC 1.1.1.22) were found in the *H. salinarum* genome.

The found gene set confirms experimental results regarding anaplerotic reactions which are catalyzed by malic enzyme (EC 1.1.1.38/29/40, OE3308F) and pyruvate, water dikinase (EC 2.7.9.1, OE1500R), but not phosphoenolpyruvate carboxykinase (EC 4.1.1.32/38/49) and oxaloacetate decarboxylase (EC 4.1.1.3) (Bhaumik and Sonawat 1994; Ghosh and Sonawat 1998).

Variants of the classic Entner-Douderoff pathway are typical glycolytic pathways in archaea (Danson and Hough 1992). In the non-phosphorylated ED pathway that has been described in *Sulfolobus* and *Thermoplasma,* glucose is directly catabolized to gluconate and subsequently to 2-dehydro-3-deoxygluconate without substrate phosphorylation. The derived hexose is then cleaved to pyruvate and glyceraldehydes. In the semi-phosphorylated pathway found in several halophilic archaea, glucose is also first converted to 2-dehydro-3-deoxygluconate which is then phosphorylated, though, prior to the split into pyruvate and glyceraldehyde 3-phosphate (GAP). For glucose-grown cells of *H. salinarum*, the first conversion from glucose to gluconate by glucose 1-dehydrogenase (EC 1.1.1.47, only feeble homolog OE1669F found) was experimentally proven, but indirect enzyme activity tests of the following reactions excluded a functional semi-phosphorylated ED pathway for this halophile (Sonawat et al. 1990). A 2-dehydro-3-deoxygluconokinase (EC 2.7.1.45) homolog (*kdgK,* OE1266R) was found in *H. salinarum*, but an 2-dehydro-3-deoxyphosphogluconate aldolase (EC 4.1.2.14) gene (*kdgA*) is missing in spite of searching with sequences of other haloarchaea (e.g. PIR: T44791, see Supplemental Table 6.3). Thus, *Halobacterium* might have lost the *kdgA* gene and with it the ability for glucose degradation probably due to the availability of other carbon substrates such as glycerol and amino acids in its environment.

Glucose degradation might also occur via the oxidative pentose phosphate pathway, where phosphorylated glucose is oxidized to gluconate 6-phosphate (Gluc6P, C6) and then converted to ribulose 6-phosphate (Ribul5P, C5) by oxidative decarboxylation. No functional oxidative PP pathway was described for any archaea yet, and consistent with this homologs for all required enzymes are absent in archaeal genomes. Although glucose 6-phosphate dehydrogenase (EC 1.1.1.49) activity had been proven (Aitken and Brown 1969), the corresponding gene is also missing in *H. salinarum*. However, in contrast to non-halophilic archaea, haloarchaeal genomes possess a phosphogluconate dehydrogenase (EC 1.1.1.44, 6PGD) gene (OE4581F) which is lacking the C-terminal 6PGD domain, though.

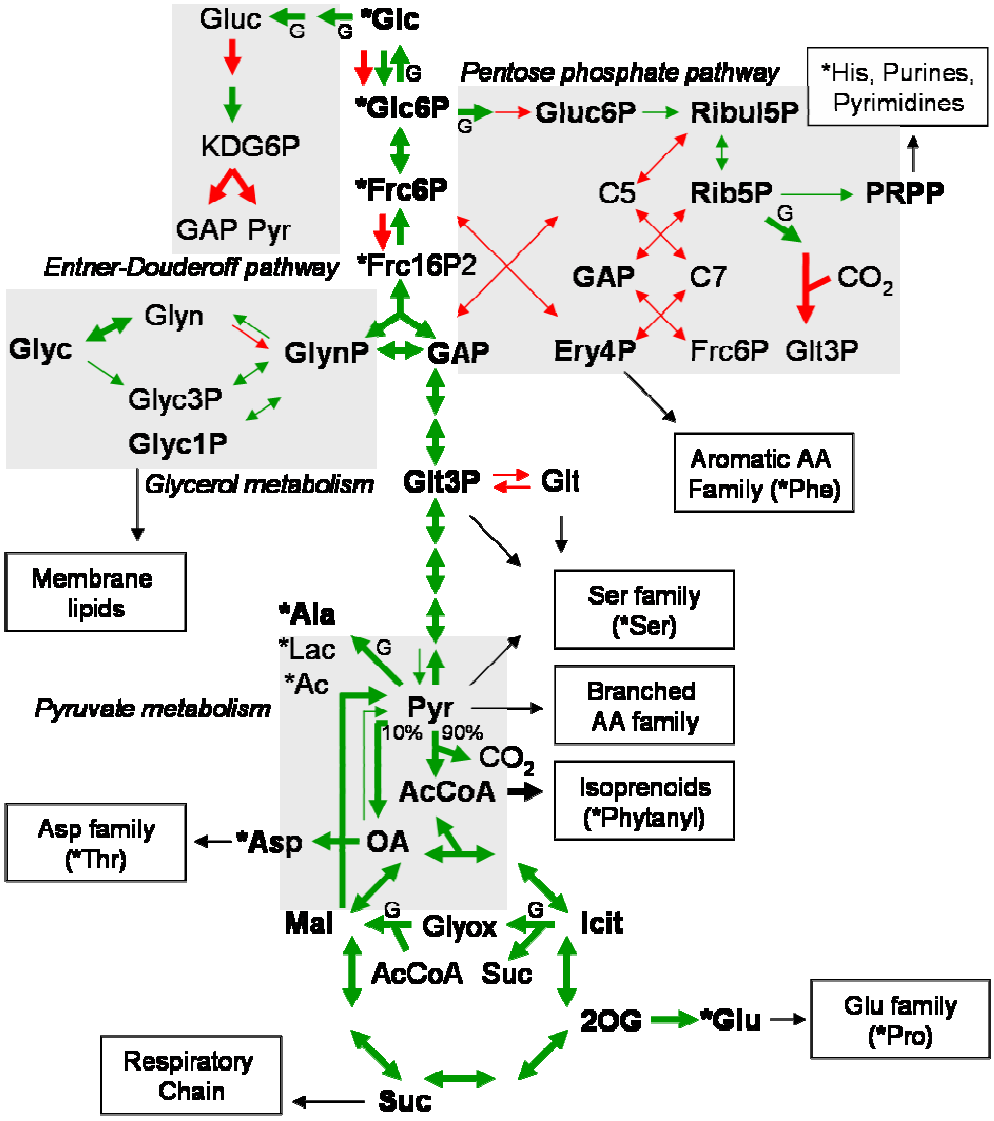


Figure 5.8: **The central intermediary metabolism of *H. salinarum*.** All 3 glycolytic pathways, the Embden-Meyerhof pathway (vertical reactions), semi-phosphorylated Entner-Douderoff pathway and the pentose phosphate pathway, are incomplete (green: reaction present, red: reaction absent). Experimentally verified reactions investigated by NMR experiments or enzyme activity tests are marked by bold arrows. However, for some of the verified reactions no genetic evidences (G) exist meaning that no homologs of the known enzymes were found in the genome sequence of *H. salinarum*. Glucose (Glc) synthesis takes place via the validated reverse EM pathway and starts with anaplerotic reactions catalyzed by malic enzymes and pyruvate, water dikinase. Pyruvate (Pyr) is mainly fed into a functional tricarboxylic acid cycle (circle) by pyruvate-ferredoxin oxidoreductase (90% of the flux), but also by pyruvate carboxylase (10% of the flux). Alanine (Ala), lactate (Lac), acetate, and aspartate signals have been detected from labelled substrates. Starting nodes for biosynthetic pathways leading to amino acid, prenyl-based lipids and nucleotides (boxes) are indicated. Compounds identified through labeling studies were marked by asterisks (Ekiel et al. 1986; Sonawat et al. 1990; Bhaumik and Sonawat 1994; Ghosh and Sonawat 1998).

Compounds: Ery4P – erythrose-4P, Frc – fructose, GAP – glyceraldehyde-3P, Gluc – gluconate, Glt – glycerate, Glyn – glycerone, Glyc – glycerol, Glyox – glyoxylate, Icit – isocitrate, KDG6P – 2-dehydro-3-deoxyphosphogluconate, Mal – malate, OA – oxalacetate, 2-OG – 2-oxoglutarate, Rib5P – ribose-5P, Ribul5P – ribulose-5P, Suc – succinate

The non-oxidative part of the PP pathway converts ribulose 5-phosphate (Ribul5P, C5) back to a hexose, fructose 6-phosphate (Frc6P, C6), and GAP (C3) in a complex five-step pathway. The complete enzyme set of this pathway is missing in most archaea except for ribose 5-phosphate (Rib5P) isomerase (EC 5.3.1.6, OE4185F) which is also involved in the first steps of the *de novo* synthesis of nucleotides and histidine. Since non-oxidative PP pathway enzymes are absent in most archaea, an alternative archaeal pathway for the synthesis of Rib5P as well as erythrose 4-phosphate (Ery4P, C4), a precursor of aromatic amino acids, is required. For methanogenic bacteria an alternative triose carboxylation pathway was proposed in which Ery4P is synthesized by carboxylation of glyceron-P or another triose (Choquet et al. 1994).

The third, reductive branch of the PP pathway, which is part of the Calvin cycle in plants, has already been investigated in halophiles (Rawal et al. 1988). Ribulose-bisphosphate carboxylase (EC 4.1.1.39, RUBISCO) was shown to be active in *H. mediterranei*, but not in *H. salinarum*. In consistence with this, no RUBISCO homolog was found in *H. salinarum* but in *N. pharaonis* (Supplemental Table 6.3).

In the natural environment of *Halobacterium*, glycerol is a highly abundant carbon source produced by the halotolerant green algae *Dunaliella salina*. Plasmid-encoded glycerol dehydrogenase (EC 1.1.1.6, OE5160F) can convert glycerol (Glyc) to glycerone (Glyn), but no glycerone kinase (EC 2.7.1.29) homolog required for glycerone phosphorylation could be detected by homology search (Figure 6.9 in Chapter 6.4.1). Thus, the subsequent step for the conversion of glycerone remains enigmatic. However, *H. salinarum* likely uses another glycerol degradation pathway, since it is the only archaea that exhibits a glycerol kinase (EC 2.7.1.30) homolog (OE3762R) so that it might be capable to phosphorylate glycerol directly to *sn*-glycerol-3P (Glyc3P). Glyc3P might then be converted to glycerone-P by glycerol 3-phosphate dehydrogenase (EC 1.1.99.5), whose genes (OE3763F to OE3765F) were found to be clustered in the genome. The derived glycerone-P, a triose intermediate of the lower EM pathway, can then be further catabolized to pyruvate. Additionally, *Halobacterium* potentially synthesizes *sn*-glycerol-1P required which is for archaeal lipid synthesis and which is produced as in other archaea from glycerone-P by glycerol-1-phosphate dehydrogenase (EC 1.1.1.261).

By 13C NMR spectroscopy it was shown that *H. salinarum* possesses an active tricarboxylic acid cycle which is mainly fed by pyruvate-ferredoxin oxidoreductase (EC 1.2.7.1, OE2622R, OE2623R) (90% of the flux), but also by pyruvate carboxylase (EC 6.4.1.1, OE3177F) (remaining 10% of the flux, Figure 5.8) (Ghosh and Sonawat 1998; Bhaumik and Sonawat 1994). Genes and enzyme activities of all citrate cycle enzymes were found (Aitken and Brown 1969; Wulff et al. 1972; Kerscher and Oesterhelt 1981; Gradin et al. 1985). In contrast to this, no genes for the two glyoxylate cycle enzymes (EC 4.1.3.1, EC 4.1.3.2) were

identified although enzyme activities have been demonstrated (Aitken and Brown 1969). In *N. pharaonis* and *H. walsbyi* genomes, however, isocitrate lyase (EC 4.1.3.1) genes are encoded. Pyruvate and 2-oxoglutarate (2-OG) are not converted by oxoacid dehydrogenase complexes, but by 2-oxoacid-ferredoxin oxidoreductases encoded by *por* and *kor* genes, respectively (Kerscher and Oesterhelt 1981). However, as several other archaea (Jolley et al. 2000) the *Halobacterium* genome revealed a gene cluster with all components of an oxoacid dehydrogenase complex with unknown function (OE4113F to OE4116F). The E1 component encoded within the *Thermoplasma acidophilum* gene cluster has recently been shown to accept branched-chain 2-oxoacids (Heath et al. 2004). Therefore, a function of the *Halobacterium* cluster in branched-chain amino acid degradation is most likely. Under anaerobic conditions, pyruvate is primarly converted to alanine presumably by an aspartate transaminase (Chapter 6.2.2), but also to lactate and acetate (Bhaumik and Sonawat 1994; Ghosh and Sonawat 1998). In spite of proven lactate dehydrogenase activity in *H. salinarum* cell extracts (Bhaumik and Sonawat 1994), no clear lactate dehydrogenase homolog was found.

5.6.2 Biosynthesis of nucleotides, lipids, and amino acids

The complete gene set for the *de novo* synthesis of IMP from Rib5P and UMP from carbamoylphosphate and Rib5P was found in *H. salinarum*. Furthermore all genes for the subsequent synthesis of purines and pyrimidines required for DNA and RNA synthesis are present in the genome. Halophiles reveal an interesting domain fusion pattern of purine synthesis enzymes, which differs from all other organisms (Figure 5.9). Genes required for step 3 and 9 of the purine synthesis are fused (OE2292F) instead of genes for step 9 and 10. The IMP cyclohydrolase domain (EC 3.5.4.10, step 10) common in most organisms is completely missing in halophiles and some other archaea (e.g. *Methanococcus jannaschii, Archaeoglobus fulgidus, Pyrococcus furiosus*). Instead, halophiles and some methanogenes possess an non-orthologous gene (OE4329F, COG3363), which encodes an archaeal type of IMP cyclohydrolase (Graupner et al. 2002). However, other archaea lack both of the known IMP cyclohydrolase types as well as further purine biosynthesis genes (Figure 5.9), so that other non-orthologous enzymes are likely to be discovered in future.

Membrane lipids of archaea consist of glycerol diether lipids with prenyl side chains instead of diacylglycerol esters. Membranes of *H. salinarum* reveal core lipids with two phytanyl side chains (C20), but also further prenyl-based compounds such as squalenes (C30), phytoenes (C40), menaquinones (C40) and a dolichol (C60) (Oesterhelt 1976; Lechner et al. 1985; Kushwaha et al. 1976) (Figure 5.10). Furthermore, several carotenoids preferentially bacterioruberins (C50) occur in the cell membrane (Oren 2002, pp. 179-183). The purple

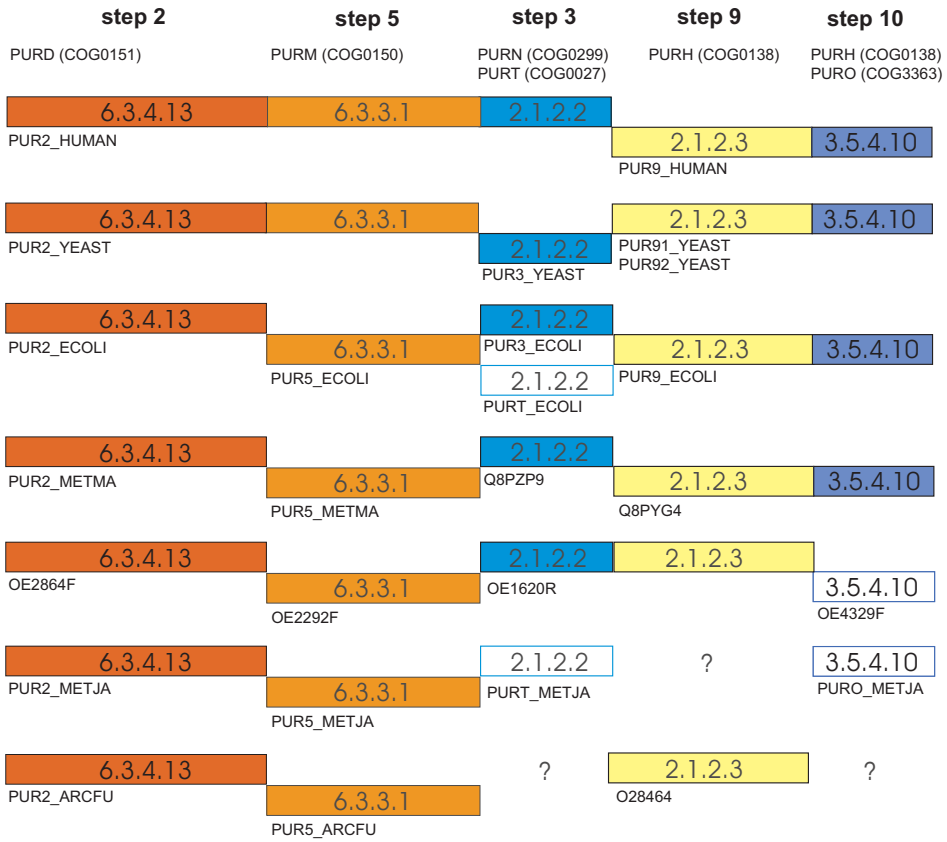


Figure 5.9: **Domain rearrangement of enzymes involved in the *de novo* synthesis of purines.** The purine synthesis pathway comprises 10 steps from PRPP to IMP. Fusions of enzyme genes are shown by linked boxes. Fused genes for step 3 and step 9 enzymes were only found in haloarchaea. Non-orthologous gene displacement occurs for step 3 and step 10 enzymes (unfilled boxes). However, further non-orthologous sequences (question marks) need yet to be discovered for archaeal purine synthesis pathways, e.g. for *Archaeoglobus fulgidus* and *Methanococcus. jannaschii*.

membrane of *Halobacterium* consists of bacteriorhodopsin that contains the photoactive retinal (C20), which is derived from β-carotene (Stoeckenius et al. 1979).

Prenyl side chains of lipids and other isoprenoids are polycondensated from activated C5 units (isopentenyl-diphosphate, IPP), which are synthesized via the mevalonate pathway in archaea. However, previous comparison of the mevalonate pathway amongst several archaea revealed gaps for three pathway steps (Smit and Mushegian 2000). Homologs of bacterial phosphomevalonate kinase (EC 2.7.4.2, P-Mvk, COG1577 (COG3890 for yeast and *Sulfolobus*)), diphospho-mevalonate kinase (EC 4.1.1.33, Dmv, COG3407), and IPP isomerase (EC 5.3.3.2, Idi, COG1443) were found to be absent in most archaea. In the meantime, the archaeal type of IPP isomerase (COG1304) has been identified and validated (Barkley et al. 2004; Yamashita et al. 2004), but the other two enzymes are still missing. However, *Halobacterium* and other haloarchaea possess bacterial diphospho-mevalonate kinase (OE1893F) and IPP isomerase (OE3560F) homologs, so that only one pathway gap for the phosphomevalonate kinase remains to be filled in the mevalonate pathway of *Halobacterium*. A gene (OE2647F, COG1608) that is located in the neighbourhood to the

mevalonate kinase (EC 2.7.1.36) gene (OE2645F, COG1577) in many archaea is the most likely candidate of the missing archaeal type of P-Mvk. Interestingly, the *Halobacterium* genome reveals not only the bacterial IPP isomerase variant but also the archaeal type of the enzyme (OE6213R, OE7093R). In future investigations, it will be interesting to find out why *Halobacterium* 'acquired' some bacterial-type and some archaeal-type mevalonate pathway enzymes. One of the two encoded IPP isomerase types might probably possess a higher substrate specifity/turnover in order to cover higher isoprenoid demands caused by bacterioruberin and retinal synthesis in halophiles.

A functional mevalonate pathway was already verified for *H. salinarum* by labeling studies (Ekiel et al. 1986). In contrast to *Methanospirillum hungatei*, unusual lipid labeling patterns were observed for *Halobacterium*. Interpreted results indicated that mevalonate (C6) is not synthesized from three activated acetate precursors but by two acetyl-CoA molecules and an

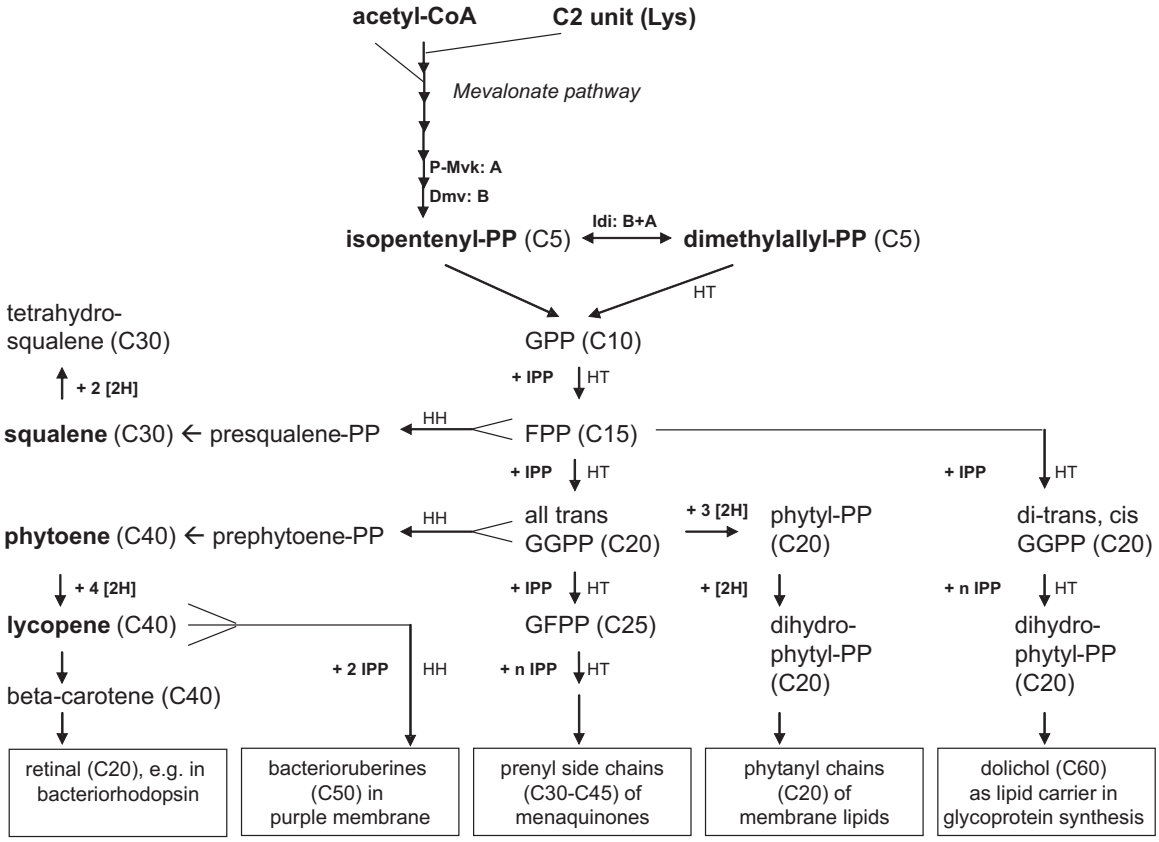


Figure 5.10: **Overview over the biosynthesis of isoprenoids in *H. salinarum*.** C5 prenyl units are synthesized via the mevalonate pathway starting from two acetyl-CoA molecules and a yet unknown C2 unit arising from amino acid degradation. For three pathway steps, non-orthologous gene displacement occurs in archaea, and it is indicated if *H. salinarum* encodes the bacterial (B) or archaeal (A) types of the respective enzymes (P-Mvk - phosphomevalonate kinase, Dmv - diphospho-mevalonate kinase, Idi - IPP isomerase). *Cis*- and *trans*-prenyl chains are derived through head-tail (HT) condensation steps with isopentenyl-diphosphate (IPP). C15 and C20 prenyl chains are modified by head-head condensations (HH, branched arrows) and desaturase reactions ([2H]). Major isoprenoid-derived membrane components identified in *H. salinarum* are shown in boxes.

unknown C2-unit not derived from acetate but from degraded amino acids (lysine and others). Thus, different types of two-carbon fragments, probably acetyl-CoA and acetyl-X, might be converted simultaneously by operation of a mixed thiolase. Although this unusual synthesis step has been proposed, a gene (OE3884F) encoding a typical 3-ketoac(et)yl-CoA thiolase (EC 2.3.1.9/16) has been found in the *H. salinarum* genome. In future, the substrate specifity of the encoded thiolase should be determined. Furthermore, the pathway for the synthesis of an alternative C-2 fragment, which is not generated from acetate, needs to be identified to clarify observed halobacterial lipid labeling patterns.

The C5 isoprenoid precursors (IPP) that are synthesized via the mevalonate pathway are polycondensated to *trans*- and *cis*-polyprenyl chains in head-to-tail fashion by (E)- and (Z)-prenyltransferases (EC 2.5.1.-), respectively (Figure 5.10). The chain-specifity of the respective homologs in *H. salinarum* (E: OE2650F, OE4010F, Z: OE3503F, OE3505R (probably non-functional)) can only be determined experimentally, but observed compounds such as phytanyl chains of membrane lipids, menaquinone side chain, dolichol indicate long-chain specifities (up to C60). Potential enzymes for the synthesis of squalenes and phytoenes by head-to-head polycondensation have also been identified in the *H. salinarum* genome (OE3093R, OE3376F).

For carotene biosynthesis, phytoene is reduced to lycopene by phytoene deaturase (EC 1.14.99.-), for which two potentially candidate genes (OE3381R, OE3468R) have been found. Lycopene is the branching point for the synthesis of bacterioruberins (C50) and β-carotene (C40) (Oesterhelt 1976). The reactions leading from lycopene to bacterioruberins have not been elucidated in detail yet but the lycopene cyclase (OE3983R) converting lycopene to β-carotene has been verified by mutagenesis studies in *H. salinarum* str. NRC-1 (Peck et al. 2002). Beta-carotene is cleaved by β,β-carotene 15,15'-dioxygenase to retinal, which is subsequently incorporated in retinal proteins. A homolog of β,β-carotene 15,15'-dioxygenase is present in *H. walsbyi* and *H. marismortui* but was not found in the genomes of *H. salinarum* and *N. pharaonis* yet. Thus, *Halobacterium* and *Natronomonas* must possess a non-orthologous enzyme responsible for oxidative cleavage of β-cartotene, and *brp* and *blh* have been shown to play a role in regulation or synthesis of retinal (Peck et al. 2001).

Amino acids such as glutamate and aspartate are presumably important carbon substrates for *H. salinarum*, which can be fed into the TCA cycle and respiratory chain to derive ATP. From its gene equipment, it can also be predicted that *Halobacterium* degrades glutamate to mesaconate via enzymes of the β-methylaspartate pathway (glutamate fermentation) that are encoded within the *mam* gene cluster (OE4204F-OE4207F). Mesaconate might further be converted to citramalate and subsequently to pyruvate and acetate as in thermophilic anaerobic bacteria (Plugge et al. 2001), but no enzyme sequences are yet available in public

databases to check this assumption by sequence comparison. It was already shown, though, that under anaerobic conditions arginine is converted to ornithine via the arginine deiminase pathway in *H. salinarum* (Hartmann et al. 1980; Ruepp and Soppa 1996). In the last step of this fermentative pathway, carbamoyl-P is used for substrate phosphorylation.

In contrast to the versatility of amino acid degradation pathways, *H. salinarum* is predicted to have reduced capabilities for the biosynthesis of amino acids compared to other halophiles such as *Haloarcula hispanica* (Hochuli et al. 1999) and *N. pharaonis* (Chapter 3.2.3). Most amino acids (10 or 15) were added in the two available synthetic growth media for *Halobacterium* (Oesterhelt and Krippahl 1973; Grey and Fitt 1976), but up to know it remains unknown which of these supplemented amino acids can be considered as essential amino acids and which are only supporting growth. Through metabolic pathway reconstruction, comparison with *N. pharaonis*, and labeling data retrieved from literature (Figure 5.8 in Chapter 5.6.1), the following set of 6 essential amino acids is proposed for *Halobacterium*, arginine, lysine, methionine, and all three branched amino acids (for details see Chapter 6.2). The proposed set of essential amino acids fits well to the set of amino acid that can be sensed by *H. salinarum* (Oren 2002, pp. 127-128) except for lysine which is not an attractant signal. Although all enzymes for cysteine biosynthesis have been found in the *Halobacterium* genome, the halophile was also shown to sense this amino acid by BasT (Kokoeva et al. 2002). The proposed set of essential amino acids derived through metabolic reconstruction might be experimentally verified by a series of growth experiments omitting potentially synthesized amino acids from the synthetic medium. However, since standardized growth on the available relatively ‘rich’ synthetic medium is already difficult to establish difficulties might arise by using this approach. The synthetic capability of *H. salinarum* for some amino acids (His, Phe, Ala, Ser, Asp, Glu, Pro, see Figure 5.8 in Chapter 5.6.1) has already been verified by labeling studies e.g. with labelled glycerol (Ekiel et al. 1986; Bhaumik and Sonawat 1994; Ghosh and Sonawat 1998). The high amount of unlabelled amino acids in the medium reduces labeling signals, though, and Ala, Phe, His, and Ser had to be omitted from the growth medium in order to detect their respective signals by NMR.

5.6.3 Coenzyme biosynthesis

Coenzymes are required for many metabolic reactions to support the catalyzing enzyme, and need either to be taken up as vitamin precursors (common in mammals) or to be synthesized *de novo* (common in prokaryotes). Here, several coenzyme synthesis pathways were assessed for *H. salinarum* and results are summarized in Table 5.6. Most of the considered pathways, however, have not been fully characterized yet, and the level of knowledge differs significantly. The precursor of biotin still needs to be established, and most coenzyme synthesis pathways contain unknown intermediary steps. On the other side, not all genes

probably involved in coenzyme biosynthesis (e.g. *thiC, thiI, moaA, moaE*) could be assigned to specific metabolic reactions yet. Considering the stated limitations, the *de novo* synthesis of menaquinone, coenzyme A (CoA), flavins, tetrahydrofolate (THF), molybdopterin, cobamide, and nicotinamide derivates (NAD^+/$NADP^+$) can be predicted for *H. salinarum*. Thus, *Halobacterium* has a reduced capability to synthesize coenzymes compared with *N. pharaonis* which shows *bio* and *thi* gene clusters for biotin and thiamine synthesis. Furthermore gene equipments for folate (C1) biosynthesis and metabolism differ significantly between various haloarchaeal strains (Chapter 6.4.3). In the synthetic medium for *H. salinarum* (Oesterhelt and Krippahl 1973), only thiamine, folate, and biotin are supplemented as vitamins.

Table 5.5: **Overview over coenzyme biosynthesis pathways in *H. salinarum*.** *De novo* synthesis pathways starting from the given precursors were assessed whether they are absent or present. However, some coenzyme synthesis pathways are not completely understood yet. It is stated if the coenzyme synthesis genes occur clustered in the genome (NA – not applicable; parentheses refer to the situation in *N. pharaonis*).

Genes	Coenzyme	Synthesis	Precursors	Unknown steps	Clustered genes	Comment
Quinones						
men	menaquinone	**yes**	chorismate, prenyl side chain	no	yes	
-	phylloquinone	no	chorismate, prenyl side chain	yes	NA	similar to men synthesis
ubi/coq	ubiquinone	no	chorismate, prenyl side chain	yes	NA	3 different pathways
pqq	pyrroloquilonine quinone	no	chorismate, prenyl side chain	yes	NA	
Vitamin B group						
thi	thiamine-PP	no	AIR, pyruvate, GAP	yes	NA (yes)	genes involved in thi metabolism found
pan	coenzyme A	**yes**	Val, β-Ala (Asp)	no	no	
pdx	pyridoxal-5P	no	E4P	yes	NA	
GTP-derived coenzymes						
rib	flavin nucleotides	**yes**	GTP, Ribu5P	no	no	enzymes for first 2 steps missing
fol	tetrahydrofolate	**yes**	GTP	yes	no	enzyme for first step missing, only reduced folate synthesis
moa/moe/ mob	molybdopterin	**yes**	GTP	yes	yes	
Porphyrines						
hem	hemes/siroheme	**yes**	Glu or Gly, Suc-CoA	no	yes	pathway up to uroporphyrinogen III/sirohydrochlorin complete
cob	cobamide	**yes**	precorrin 2, ATP, Thr, riboflavin	yes	yes	2 pathways: an- or aerobic
Other coenzymes						
nad	nicotinamide nucleotides	**yes**	Asp (Trp), ATP	no	yes	
bio	biotin	no	pimeloyl-CoA (malonyl-CoA?)	yes	NA (yes)	

In future, it should be determined which coenzymes are utilized for metabolic reactions in *Halobacterium* and whether these coenzymes are synthesized *de novo*. Usage of some coenzymes has already reported in literature through investigation of coenzymes requirements for halophilic enzymes, e.g. two of the glutamate dehydrogenases (NAD^+: OE1270F, $NADP^+$: OE1943F) (Hayden et al. 2002) and pyruvate-ferredoxin oxidoreductase (ferredoxin, CoA: OE1710R, OE1711R) (Kerscher and Oesterhelt 1981). Furthermore, a flavin compound was detected in the crystallized dodecin (Bieger et al. 2003). The *Halobacterium* genome contains several copies of proteinaceous coenzymes, e.g. ferredoxin, thioredoxin, halocyanin and Fe-S proteins, that are likely involved in a variety of halobacterial redox reactions. The usage of predicted essential coenzymes such as pyridoxal-5P within the metabolism of *Halobacterium* should also be assessed.

Recently a pathway for reduced folate synthesis via an alternative dihydrofolate reductase (Prd domain) has been proposed in halophilic archaea (Levin et al. 2004). Gene equipment suggests that *Halobacterium* likely employs only this pathway (Prd domain in OE1615R) for the *de novo* synthesis of folate and not the standard synthesis pathway via FolA (absent in *H. salinarum*, for details see Chapter 6.4.3). Thus, *Halobacterium* might require folate for growth although genes for THF synthesis are generally present in its genome. *H. salinarum* str. R1 likely lacks folate synthesis completely, since one of the folate synthesis genes (OE1570F) is disrupted by an ISH element in this strain. It should be noted that gene equipment for folate biosynthesis and metabolism differs significantly between the four halophilic strains and that these differences might account for differences in folate-dependent reactions of other pathways (Chapter 6.4.3).

Several compounds which are presumably involved in the halobacterial respiratory chain (Chapter 6.3.1), i.e. menaquinones and hemes, have been identified in *Halobacterium* (Oesterhelt 1976; Sreeramulu et al. 1998). In consistence with this, *men* and *hem* gene clusters for their biosynthesis were detected in the *H. salinarum* genome. However, only heme synthesis genes for pathway steps from glutamate to uroporphyrinogen III (UroIII) have been found. Genes encoding enzymes which catalyze subsequent modifications of the porphyrin system such as the genes for porphyrinogen decarboxylase (*hemE*) and several oxidases (*hemF, hemG, hemY*) are absent except for a heme prenyltransferase homolog (*cyoE*) (Figure 5.11). Thus, *H. salinarum* must employ yet unknown alternative enzymes for side chain modifications of UroIII.

Halobacterium likely synthesizes cobamide (coenzyme B12) from UroIII, a pathway that was extensively studied in *Salmonella typhimurium* (*cbi/cob* genes, anaerobic pathway) and *Paracoccus denitrificans* (*cob* genes, aerobic pathway) (Roth et al. 1996). The *Halobacterium* gene set for cobalamine synthesis resembles the sets of both of these model organisms, except for a gene duplication of *cbiH (*also named *cobJ)* (OE3214F, OE3216F).

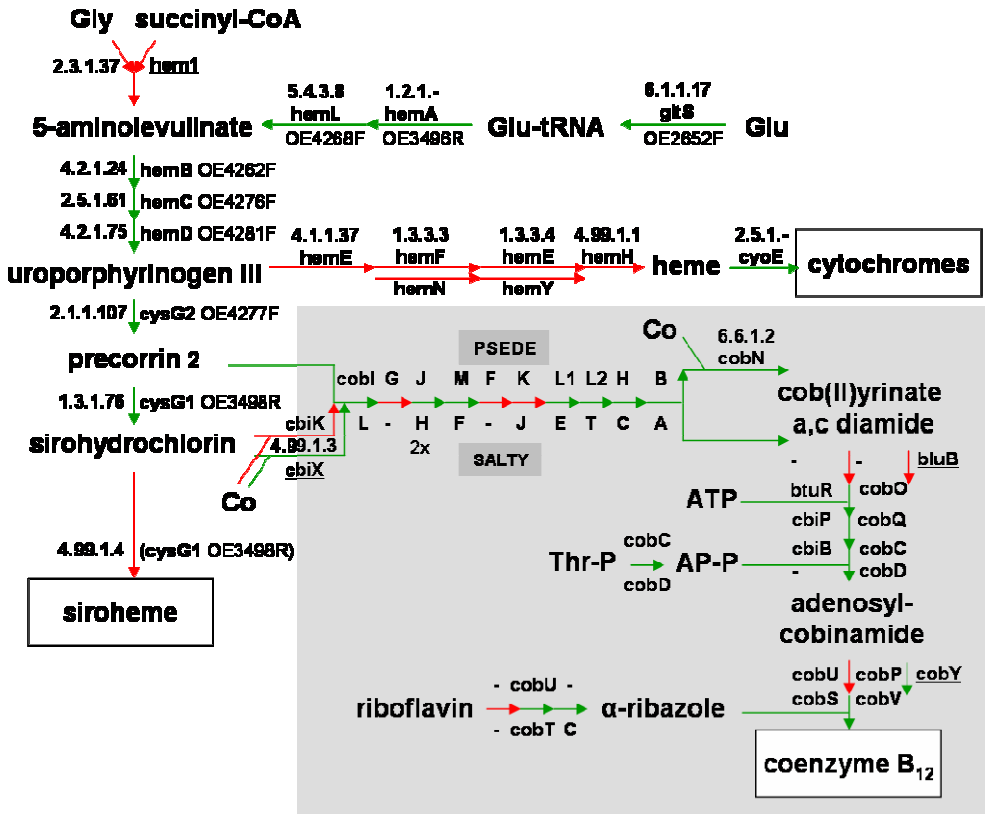


***Figure 5.11: De novo* synthesis of porphyrins and cobamide (coenzyme B_{12}) in *H. salinarum*.** Aminolevulinate precursors are derived via the glutamate pathway, and 8 molecules are condensed to uroporphyrinogen III (UroIII) via porphobilinogen. UroIII is presumably further metabolized to precorrin 2 and sirohydrochlorin (green arrows, genes present, OE-Id given). Genes for heme synthesis were not detected yet except for a heme prenyltransferase (red arrows, genes absent). Cobamide synthesis (grey box) was mainly studied in *Paracoccus denitrificans* (aerobic pathway, gene names below/right of arrows) and *Salmonella typhimurium* (anaerobic pathway, gene names above/left of arrows). The two pathway variants differ in their cobalt integration step, occurring early in *S. typhimurium* (*cbiK*) and *B. megaterium (cbiX)*, but late in *P. denitrificans (cobN)* (KEGG-Map00860). Since the *H. salinarum* genome reveals *cbiX* and *cobN* homologs, it cannot be concluded, which type of pathway is used to synthesize cobamide. Minus signs mark reactions for which no genes were assigned for *P. denitrificans* and *S. typhimurium* reactions yet. Underlined genes indicate genes from other species. AP - (R)-1-aminopropan-2-ol.

Two types of cobalt chelatases were found, CbiX (EC 4.99.1.3, OE3221F) and CobN (EC 6.6.1.2, OE3230F) required for cobalt integration within the anaerobic and aerobic pathway, respectively (Figure 5.11). Thus, it cannot be predicted whether *H. salinarum* synthesizes cobamide predominantly via the oxygen-dependent or -independent.

Halobacterium employs also a salvage pathway of the stable coenzyme B12 precurser dicyanocobinamide (Cbi) in order to generate cobamide (Woodson and Escalante-Semerena 2004). This archaeal Cbi salvage pathway differs from the conserved Cbi salvage pathway of bacteria in the type of pathway intermediates and enzyme genes. While archaea employ adenosylcobinamide amidohydrolase (EC 3.5.1.90) (*cbiZ*, OE3261F) as key enzyme of the salvage pathway, bacteria use adenosylcobinamide kinase (EC 2.7.1.156) encoded by *cobU*

for Cbi salvage. CobU is a bifunctional enzyme (EC 2.7.1.156/2.7.7.62) which is also required for the bacterial *de novo* synthesis pathway of coenzyme B12. Archaea lack a *cobU* ortholog, which is replaced by the non-orthologous *cobY* gene (OE3257F, Figure 5.11) (Woodson and Escalante-Semerena 2004).

5.7 Conclusions

Correct enzyme assignments are an important basis for metabolic reconstruction and subsequent modelling. Enzyme functions designated by EC numbers are usually assigned to new genes by transferring function annotations from similar genes. As discussed, this procedure is problematic since function assignments spread through public databases without knowing the original source of the functional information. Furthermore, only one EC number (usually the best blast hit) is assigned to a gene, although in many cases one specific EC number cannot be determined. This is due to the fact that enzyme subtypes with differing coenzyme and substrate specifity (4. EC-No. position varies) are highly similar on sequence level. Especially for moderate E-value cutoffs (e^{-20}), the concept of choosing one out of several EC number candidates was found to be highly unreliable, since alternate EC numbers commonly belong to differing enzyme subclasses (1.-3. EC-No. position differs) and metabolic pathways. Therefore, a new enzyme assignment routine was developed that takes the limited predictability of exact EC numbers into account by permitting the assignment of several EC numbers per gene. Reliability of EC assignments was further increased by integrating results from four similarity-based searches and by indicating the best EC number annotation for a gene. For metabolic reconstruction, the automatically retrieved enzyme assignments were further manually assessed in order to avoid mispredictions, e.g. in the case of multidomain proteins whose distinct enzyme activities need to be checked separately.

The developed metabolic database, Pathnet, which contains reference and organism-specific pathway data, is a valuable resource for computationally- and experimentally-derived metabolic information on *H. salinarum*. Metabolic data from Pathnet can be accessed through a web-interface and represented via customized KEGG maps. Further, experimental data from genome-wide ‘omics’-approaches can be linked to Pathnet entries in order to map identified proteins and regulated genes onto halobacterial pathways. Based on the reconstructed metabolic reactions, metabolic models will be created in future, which will comprise defined pathways, subsystems, or the complete metabolic network of *H. salinarum*.

Thereby, reaction confidence levels and comments as well as cross-linked literature data will be useful to assess created models.

The central intermediary metabolism of *Halobacterium* and most other archaea is characterized by a lack of pentose-phosphate pathway enzymes raising the question how precursors for the *de novo* synthesis of nucleotides, aromatic amino acids, and several coenzymes are synthesized. Most archaeal pathways reveal gaps for which no enzyme orthologs can be detected in the genome. Several of these archaeal pathway gaps (e.g. in the mevalonate pathway and for the *de novo* synthesis of purines) could be filled by the discovery of non-orthologous enzymes in recent years. However, remaining pathway gaps make the assessment of archaeal metabolic pathways difficult and potentially error-prone, and demand constant update of metabolic reconstruction data to recent findings. A future in depth-analysis of abundant non-orthologous gene displacements and convergent evolution in archaeal metabolic networks might give interesting insights into the evolution of pathways.

Several of the predicted synthetic capabilities of *H. salinarum* could be validated by experimental data from literature. However, enzyme acitivity tests are often opposed to genomic findings e.g. for glyoxylate cycle enzymes, and it is unclear whether this is due to studies in different *Halobacterium* strains (multiple renaming occurred within the last decades) or due to yet unknown non-orthologous enzyme genes. Therefore, activity tests should be repeated for the sequenced *H. salinarum* strain R1 in order to resolve data inconsistencies. Furthermore, *in silico* predictions for the synthesis of amino acids and coenzymes should to be verified in future.

5.8 Methods

5.8.1 Enzyme assignment

Accurate function prediction for the complete gene set of an organism is a prerequisite for metabolic reconstruction. Therefore, not only a standard blast search against a public database such as nr (NCBI) or Swiss-Prot (Expasy) was performed but four different similarity searches routines were run (Figure 5.11) in order to gain precise enzyme assignments for *H. salinarum* genes.

For the first similarity search method by blast (BLAST search), enzyme sequences were selected from the Swiss-Prot database beforehand by searching for enzyme classification numbers (EC-No.) within the description field of Swiss-Prot entries. The Swiss-Prot database was chosen as a reference database for functional annotation, since it is an extensive protein database with manually curated annotations and crosslinks to many databases.

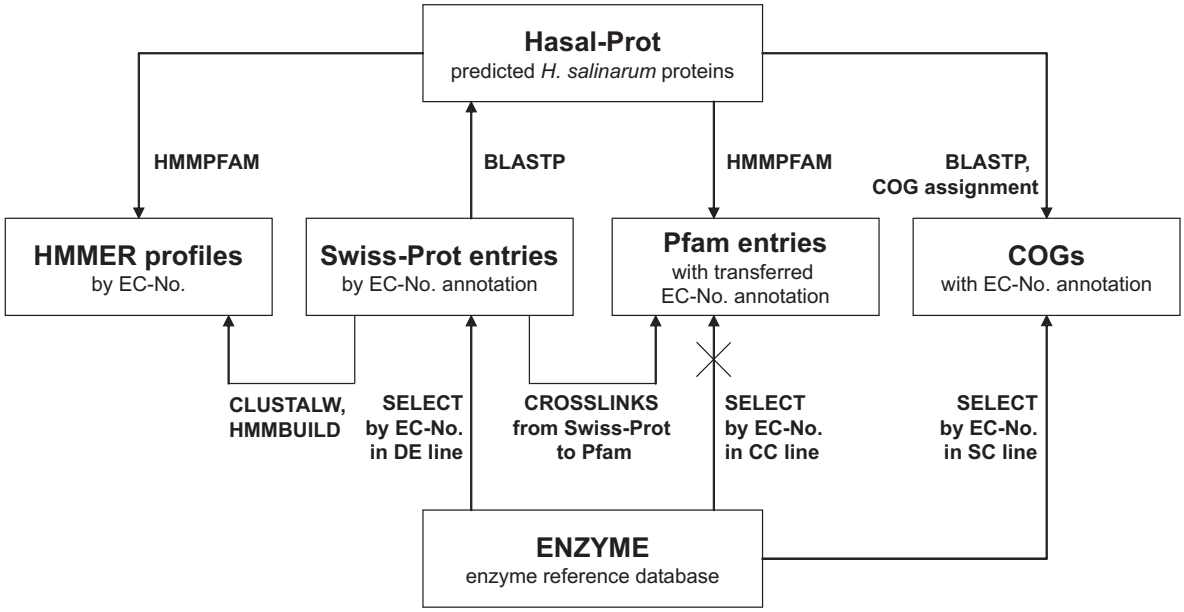


Figure 5.12: **Enzyme assignment via various similarity search methods** (BLAST, HMMER, PFAM, and COG search). In order to assign enzyme functions to the predicted protein set of *H. salinarum*, enzyme entries were first selected from public databases in flatfile format (Swiss-Prot, Pfam) or from info webpages (COG). For each of the enzyme sequence sets from Swiss-Prot, HMMER profiles have been created. Since only few Pfam entries contained EC-No. annotations, EC numbers have been transferred from Swiss-Prot by analysing Swiss-Prot crosslinks to Pfam. In a second step, similarity searches were run using the different databases with EC-No. indices. Thus, enzyme classification numbers were assigned to the search hits in *H. salinarum* by sequence-based function transfer. DE – description, CC – comment, SC – systematic classification.

The selected sequences for each EC-No. were used to search against the set of predicted *H. salinarum* genes by blastp, so that genes could be assigned to each of the classified enzymes. Usually EC numbers are assigned by blast searches that are performed *vice versa* by searching with the new genes against public databases and then transferring EC numbers of best blast hits. However, by pre-selecting enzyme sequences for each EC number, all EC numbers with available enzyme sequences are considered equally and not only the ones that are highly prominent in public databases (see Chapter 5.2.2).

The selected Swiss-Prot entries for an EC number were also used to create multiple sequence alignments by ClustalW in case at least two sequences were available. From these multiple sequence alignments for each EC-no., Hidden Markov Models were subsequently built which then contain conserved sequence information of several enzyme sequences from all three domains of life. From the collection of HMMER profiles for each EC number, an enzyme profile database was created which was searched against with the *H. salinarum* gene set (HMMER search). By applying a profile-based search method, the selectivity of the similarity searches is increased, in contrast to searching with each of the single enzyme sequences (BLAST search). However, the HMMER search is limited to enzyme profiles that contain exclusively similar sequences. For enzyme heterocomplexes, multidomain enzymes as well as for non-orthologous enzymes, this prerequisite is not

fulfilled, and, thus, only for 40% of the EC numbers (with two or more sequences) a usable profile could be created.

Full-length sequence similarity search by blast often detects only sequences in the 'twilight zone' (e.g. if E-values are worse than e^{-10}), so that function and EC number annotations based on these results are not reliable anymore. Sequences might reveal a domain or motif with conserved amino acid residues, though, e.g. the catalytic centre that is characteristic for an enzyme. Thus, previously questionable functional annotations can be fortified by domain searches. Within this project, the well annotated Pfam domain database was chosen to search against (PFAM search) using a search tool available at the Smart website. However, the subsequent assignment of enzyme functions to *H. salinarum* genes via found Pfam domains was problematic, since Pfam entry descriptions rarely contained EC number annotations. Enzyme classifications were therefore taken from Swiss-Prot entries following the analysis of Swiss-Prot crosslinks to the Pfam database. Assuming that crosslinks to Pfam by the Swiss-Prot database are consisted for all Swiss-Prot entries of one EC number, typical motifs of classified enzymes were retrieved and EC numbers could so subsequently matched to *H. salinarum* genes that contain the Pfam motif.

Finally, each of the *H. salinarum* genes was assigned to a cluster of orthologous groups (COG) (Tatusov et al. 1997), with a minimal E-value of e^{-05} (COG search). Enzyme annotations for COGs were extracted from the systematic classification field for a COG entry as given on the COG website. However, the coverage of EC number annotations within COG entries is relatively low (ca. 600 EC numbers were found in 4873 COGs) and the accuracy of these annotations is furthermore unclear. Thus, enzyme assignments of *H. salinarum* genes via the COG search should not be considered in case of high deviations to the other three searches.

In order to summarise the results of the four different sequence-based searches, the expectation values (E-values) resulting from the single searches were transformed into scores:

$$Score = - ln (E\text{-}value) / ln (10)$$

The search scores were then plotted against their frequencies, and the score distributions for each single search method were fitted to Loess polynomial functions using R statistics software (Figure 5.13). In case the four distributions result in similar graphs, the distributions can be standardised, so that the search scores of the single search methods could be standardised as well in order to compare them amongst each other. However, the attempted standardization of search scores could not be realized since distributions for the different search methods differ. While graphs for BLAST and PFAM search results are similar with

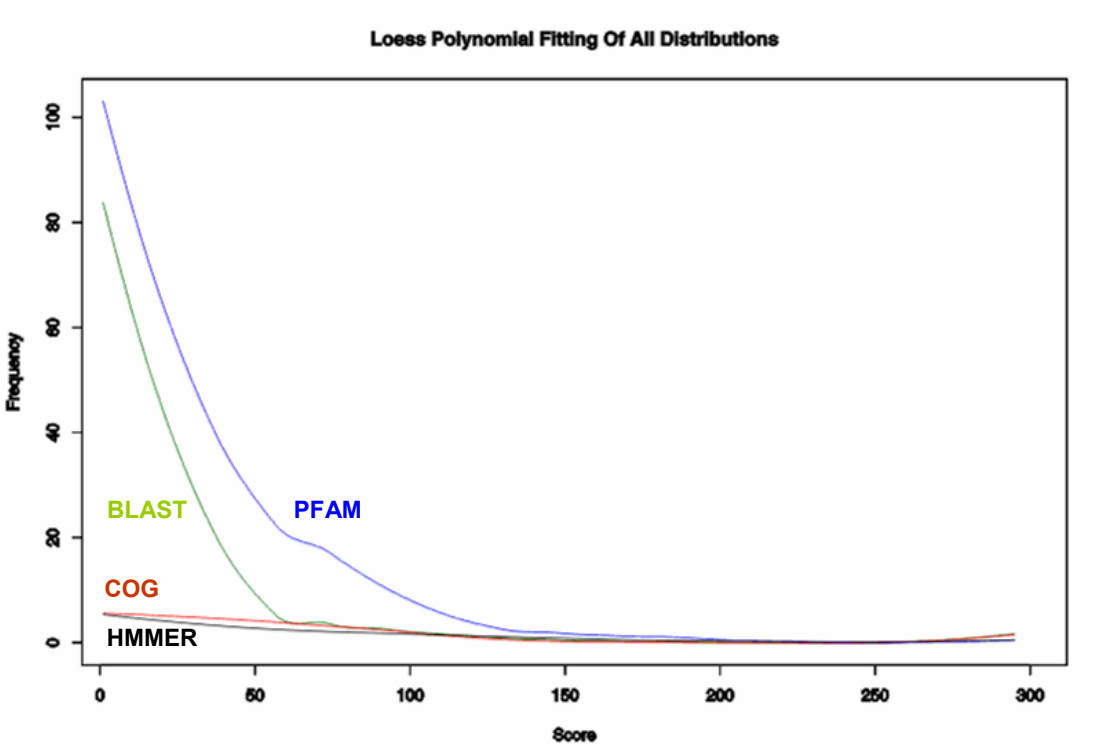


Figure 5.13: **Fitted score distributions for the search hits of all four sequence-based search methods** that were used to assign enzyme functions to the complete set of *H. salinarum* genes. Scores below a cutoff of five (equivalent to an E-value of e^{-05}) were excluded, and the score distributions were fitted to a Loess polynomial function using R.

many assignments for low scores and few for high scores, far less COG and HMMER search results were obtained in the range of low scores <50. As a consequence, a total score had to be calculated from the single scores without standardization of the scores:

Total Score = sum (Single scores) / 4

Finally, the total scores were refined by removing single scores with high deviations to the total score and recalculation of the total scores.

5.8.2 Database structure and implementation of the metabolic database

The database structure of the developed metabolic database Pathnet is discussed in Chapter 5.3. Pathnet tables were implemented as a MySQL database. Reference collections of metabolic *reactions*, *compounds*, and *enzymes* were downloaded in flatfile format from KEGG ligand (version 29), Expasy ENZYME (version 35) and Swiss-Prot (version 44) databases, and integrated into the general tables *reactions*, *compounds*, and *enzymes*. KEGG reaction entries were modified and corrected, and both, the curated data and original KEGG data (*reactions* columns with the prefix 'kegg') stored in order to enable an automatic

update in the future. Data generation for the organism-specific Pathnet tables is described in the next section.

5.8.3 Pathway reconstruction procedure

Metabolic pathways of *H. salinarum* were reconstructed in a two-step procedure. In the first step, relevant *reactions* that take part in the given metabolic pathway were chosen from the complete set of reference *reactions*. The assessment of available *reactions* entries was necessary, since EC numbers are often assigned to more than one *reactions* in the KEGG database (Chapter 5.2.4). The *reactions* for an EC number can be involved in central, but also in specific metabolic pathways which are only relevant for few organisms. Therefore, only KEGG reactions, which are part of standard metabolic pathways relevant for *H. salinarum*, were considered. As an example, the reaction from 2-oxoadipate to 2-aminoadipate catalyzed by 2-aminoadipate transaminase (EC 2.6.1.39) (lysine synthesis) was marked as relevant, whereas the conversion of 6-acetamido 2-oxohexanoate (N-acetyl-lysine synthesis) by the same enzyme was excluded from the reconstruction process. The selection of relevant reactions was realized through a developed web interface (option: 'general reactions'), which selects all EC numbers found on a chosen KEGG map (Figure 5.12)

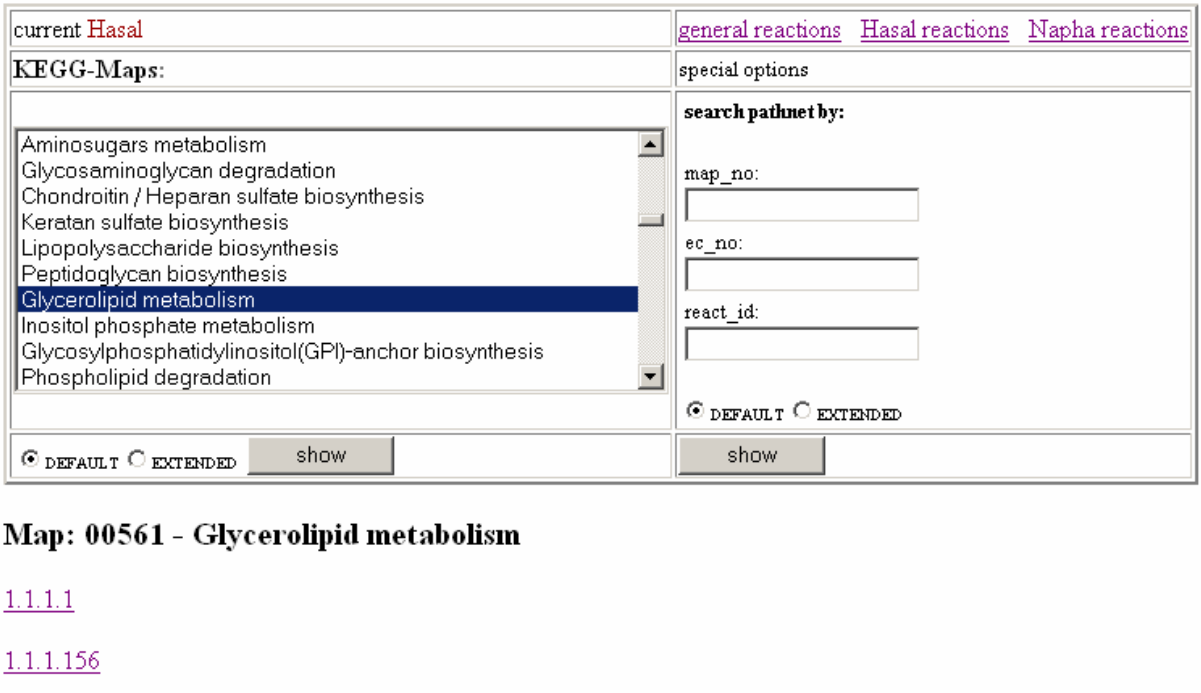


Figure 5.12: **Web interface for metabolic pathway reconstruction**. For a chosen pathway (left window, KEGG-Map: 'glycerolipid metabolism'), all enzymes of the map are listed by their EC numbers (only two EC-No. shown here). Each listed EC-No. is further linked to the according catalyzed reference *reactions* (option: 'general reactions') or to organism-specific database entries (option: 'Hasal/Napha reactions', see Table 5.2 in Chapter 5.3). The Pathnet database can also be searched by a specific KEGG-Map-No., EC number, or Reaction-Id (right window).

and then lists all available *reactions* entries for each of the EC numbers (Muller 2005). The selected *reactions* entries were assessed by setting the reaction-excluded-flag from the default value (flag = 0, unprocessed reaction) to excluded (flag = 1) or to considered (flag = 2). Furthermore, a form can be called to fill in subsystem and pathway list fields for a reaction.

Organism-specific metabolic pathways for *H. salinarum* were reconstructed in the second step of the procedure by creating new database entries for the tables *org_reactions* and *org_enz_genes*. Organism-specific *org_reactions* entries are linked to the general *reactions* entries with their chemical equations and pathway lists, to the reference collection of *enzymes*, and to the involved *enzyme genes* of the organism. Inserts and updates of organism-specific database entries are assisted by the developed web-interface (option: 'Org reactions') (Muller 2005). For each enzyme of a chosen pathway, only the considered reference *reactions* are presented. Furthermore, gene entries of the Halolex genome database (table *org_basicdata*), which are annotated with the respective EC-No., and entries resulting from the enzyme assignment routine (table *org_enz_autoassign*, see Chapter 5.8.1) are retrieved (Table 5.2A-C in Chapter 5.3). Results from bidirectional blast between *Natronomonas* and *Halobacterium* genomes (see Chapter 2.2) are also given (function currently only available for *N. pharaonis*). Based on the presented data collection from different database tables, different fields of *org_reactions* and *org_enz_genes* entries were edited in given forms after triggering insert/update buttons. Often, a *comment* regarding the assessed *org_reaction* or *org_enz_gene* was created, which was stored separately in table *org_comments* upon form submission. When *org_reactions* and *org_enz_genes* entries are available for the processed EC-No., data are shown (Table 5.2DE in Chapter 5.3).

5.8.4 Graphical representation of metabolic pathways

Data of reconstructed pathways can be retrieved from the Pathnet database through the developed web interface (Figure 5.12, Table 5.2 in Chapter 5.3), but a graphical pathway representation gives a more comprehensive view of the organism's metabolism. Therefore, static pathway maps were drawn for *H. salinarum* pathways, which are available through the Halolex website. Furthermore, a graphical routine was developed which generates organism-specific pathway maps on demand by using KEGG web services (Figure 5.7 in Chapter 5.5) (Muller 2005). The program transmits selected Pathnet data to a colouring routine (*color_pathway_by_objects*) of the KEGG API, which generates KEGG maps with customized colours. For each enzyme of a KEGG map, a colour code can be defined for the EC-box and/or EC-label, which is transferred to the KEGG server by using the SOAP protocol. The colouring routine then returns a KEGG map picture (*gif* format) with customized enzyme colouring. In order to link available Pathnet data to each enzyme of the static KEGG

Table 5.6: **Options of the developed graphical program that represents metabolic pathways** based on KEGG maps. The existence of an enzyme (yes: E_{exist}, no: $E_{!exist}$) found on a given KEGG map is determined by the reconstruction data selected from Pathnet or by genome annotation data derived from the Halolex database. Enzymes with experimental data (E_{exp}) retrieved from the literature or from genome-wide 'omics' approaches are specifically colour-marked. The graphical program transfers the defined colour codes for the EC-boxes (Box) and EC-labels (Label) to the external KEGG API that generates the coloured KEGG map. Availability of the data (yes: Y, no: N) differs for the three halophilic strains, *H. salinarum* (HS), *N. pharaonis* (NP), and *H. walsbyii* (HQ).

Data	HS	NP	HQ	E_{exist}	$E_{!exist}$	E_{exp}
genome annotation	Y	Y	Y	Box: yellow	Box: grey	-
metabolic reconstruction	Y	Y	N	Label: green; Box: yellow (manual), grey (automatic)	Label: red; Box: yellow (manual), grey (automatic)	Box: dark yellow
qualitative proteomics	Y	Y	N	Box: yellow	Box: grey	Label: green
quantitative proteomics	Y	N	N	Box: yellow	Box: grey	Label: green (up), red (down), pink (inconsistent)
transcriptomics	Y	N	N	Box: yellow	Box: grey	Label: green (up), red (down), pink (inconsistent)

map, the map is overlaid by an image map whose coordinates are defined by EC-box coordinates of the given KEGG map (as downloaded from the KEGG ftp page). The image map coordinates link to the respective Pathnet data for the given EC-No. of the overlaid EC-boxes (Table 5.2 in Chapter 5.3).

The graphical program was extended to use EC number annotations from complete genomes, and was applied to three halophilic genomes. Further, colouring of KEGG pathways by other data (e.g. proteomics and transcriptomics) was implemented. The different options of the graphical program are summarized in Table 5.6.

5.8 Supplemental material

Supplemental Table 5.7: **(A) List of defined pathways with their metabolic subsystems (S).**

(B) Statistical overview of the metabolic pathway reconstruction for *H. salinarum*. An overall entry statistics of *H. salinarum*-specific Pathnet tables is given followed by statistics for individual pathways. Numbers (#) of created *org_reactions* (R), *org_enz_genes* (GEN), and *org_comments* (CMT) entries are listed. *Org_reactions* entries are distinguished into existing (exist) and non-existing (!exist) reactions. Some of the existing *reactions* have no genetic evidence (!gene). Further, numbers of listed enzymes (ENZ) and literature references (LIT) are given. Subsystem definitions are introduced in Table 5.3.

A:

S	Pathway	Pathway name	Paths
C	*EM_pathw*	Embden-Meyerhof pathway	Glc/Frc → Frc1,6P2 →pyruvate (and reverse)
C	*ED_pathw*	Entner-Douderoff pathway	Glc → gluconate(6P) → GAP + pyruvate
C	*PP_pathw*	pentose-phosphate pathway	Glc6P → ribulose5P → GAP + Frc6P
C	*glycerol_met*	glycerol metabolism	glycerol → glycerone(P) → GAP glycerol → glyceraldehyde → glycerate3P
C	*pyruv_met*	pyruvate metabolism	pyruvate → acetyl-CoA OA/malate → PEP/pyruvate pyruvate → acetate/lactate/Ala
C	*TCA_cycle*	tricarbon acid/ citrate cycle	acetyl-CoA + citrate → 2-OG → malate → citrate
L	*mevalon_pathw*	mevalonate pathway	2 acetyl-CoA + C2 → mevalonate → IPP/DMAPP
L	*isopren_synth*	isoprenoid biosynthesis	IPP → GFPP IPP → dolichol IPP → squalene/phytoene
L	*carot_synth*	carotene biosynthesis	GGPP → phytoene → beta-carotene → retinal
P	*purine_synth*	purine biosynthesis	ribose5P → AIR → IMP
P	*purine_met*	purine metabolism	IMP → orotate → dATP/dGTP
P	*pyrim_synth*	pyrimidine biosynthesis	carbamoylP → orotate orotate + PRPP → UMP
P	*pyrim_met*	pyrimidine metabolism	UMP → dCTP/dTTP/dUTP
A	*Glu_fam_synth*	glutamate-family biosynthesis (Glu, Gln, Pro, Arg)	2-OG → Glu → Gln Glu → Lys Glu → Glu-5-semialdehyde → Pro Glu → ornithine
A	*Arg_met*	arginine metabolism/ urea cycle	ornithine + carbamoylP → citrulline → Arg → ornithine Arg → citrulline → ornithine + carbamoylP
A	*Asp_fam_synth*	aspartate-family biosynthesis (Asp, Asn), Ala	OA → Asp → Asn Asp → Asp-4-semialdehyde → homoserine pyruvate → Ala
A	*homoser_met*	homoserine metabolism (Thr, Met)	homoserine → OP-homoserine → Thr homoserine → homocysteine → methionine
A	*Ser_fam*	serine-family biosynthesis (Ser, Gly, Cys)	glycerate(3P) → (3P)hydroxypyruvate → Ser
A	*branch_aa_synth*	branched-chain amino acid biosynthesis (Val, Leu, Ile)	2 pyruvate → 2-oxoisovalerate → Val pyruvate + 2-oxobutyrate → 2-oxo-3-methylvalerate → Ile 2-oxoisovalerate → 2-oxo-isocaproate → Leu
A	*Lys_synth*	lysine biosynthesis	Asp → Asp-4-semialdehyde → dihydropicolinate → diaminopimelate → Lys (bacteria) acetyl-CoA + 2-OG → homocitrate → 2-oxoadipate → saccharopine → Lys (fungi)
A	*His_synth*	histidine biosynthesis	PRPP → imidozole-glycerol3P + AICAR → His
A	*shikim_pathw*	shikimate pathway	Ery4P + PEP → 3-dehydroquinate → shikimate → chorismate
A	*arom_aa_synth*	aromatic amino acid biosynthesis (Phe, Tyr, Trp)	chorismate → prephenate → phenylpyruvate → Phe prephenate → 4-hydroxyphenylpyruvate → Tyr chorismate → anthranilate → Trp
V	*quin_synth*	quinone biosynthesis	chorismate → 4-hydroxybenzoate → ubiquinone chorismate → isochorismate → mena/phylloquinone
V	*CoA_synth*	coenzyme A biosynthesis	2-oxoisovalerate → pantoate → 4Ppantetheine → CoA
V	*NAD_synth*	NAD biosynthesis	Asp → quinolinate → NAD+/NADP+
V	*thiam_synth*	thiamine biosynthesis	pyruvate + GAP + AIR → thiaminePP
V	*pyridox_synth*	pyridoxal biosynthesis	Ery4P → pyridoxine(P) → pyridoxal5P
V	*biotin_synth*	biotin biosynthesis	malonyl-CoA (?) → pimeloyl-CoA → biotin
V	*riboflavin_synth*	riboflavin biosynthesis	GTP → riboflavin → FMN/FAD
V	*fol_synth*	folate biosynthesis	GTP → dihydropteroate → tetrahydrofolate
V	*porph_synth*	porphyrine biosynthesis	Glu → Glu-tRNA → 5-aminolevulinate → UroIII UroIII → heme/chlorophyll UroIII → precorrin 2 → sirohydrochlorin → siroheme
V	*cobal_synth*	cobalamine biosynthesis	precorrin → (Co-)precorrin 8X → Cob(II)yrinate a,c diamide → adenosylcobinamide + Thr + alpha-ribazole → cobalamin coenzyme

B:

S	Pathway	KEGG Map	# R_{exist}	# $R_{!exist}$	# $R_{!gene}$	# GEN	# ENZ	# CMT	# LIT
C	*EM_pathw*	00010	15	11	1	12	28	28	6
C	*ED_pathw*	00030	4	7	3	1	12	8	5
C	*PP_pathw*	00030	3	7	2	1	9	8	3
C	*glycerol_met*	00561	12	6	0	14	18	19	5
C	*pyruv_met*	00620	14	14	2	12	29	27	7
C	*TCA_cycle*	00020	21	7	2	14	23	27	8
L	*mevalon_pathw*	00100	7	0	0	9	8	5	2
L	*isopren_synth*	00100	15	0	0	10	11	8	2
L	*carot_synth*	00100	9	0	1	5	5	6	4
P	*purine_synth*	00230	12	0	0	12	11	4	1
P	*purine_met*	00230	13	7	1	9	16	4	0
P	*pyrim_synth*	00240	7	0	0	8	7	4	1
P	*pyrim_met*	00240	16	10	0	9	14	4	0
A	*Glu_fam_synth*	00251 00330 00220	9	15	1	10	22	11	10
A	*Arg_met*	00330	7	1	0	7	8	7	2
A	*Asp_fam_synth*	00252	8	4	0	9	12	5	0
A	*homoser_met*	00271	15	5	0	12	16	7	2
A	*Ser_fam*	00260	7	3	0	7	10	3	0
A	*branch_aa_synth*	00290	3	18	0	2	11	3	0
A	*Lys_synth*	00300	1	24	0	1	20	2	0
A	*His_synth*	00340	10	0	1	9	9	4	1
A	*shikim_pathw*	00400	6	1	0	6	7	6	3
A	*arom_aa_synth*	00400	15	8	0	17	18	16	1
V	*quin_synth*	00130	9	7	0	10	11	5	0
V	*CoA_synth*	00770	14	6	4	7	18	1	0
V	*NAD_synth*	00760	7	2	0	5	8	2	0
V	*thiam_synth*	00730	2	6	0	3	6	2	0
V	*pyridox_synth*	00750	0	12	0	2	7	3	0
V	*biotin_synth*	00780	2	6	0	2	6	2	0
V	*riboflavin_synth*	00740	9	3	4	5	11	3	0
V	*fol_synth*	00790	13	6	10	3	11	2	0
V	*porph_synth*	00860	9	6	0	9	14	1	0
V	*cobal_synth*	00860	36	1	15	18	26	14	0

Supplemental Table 5.8: **Definition of Pathnet tables (bold) and attributes (italic).** Data sources for Pathnet tables are indicated in grey. PRI - primary key, SEC - contains secondary keys, semlist - semicolon separated list.

Table/Field	Type	Key	Format	Source	Comment
reactions				reaction file (KEGG ligand)	entries > R10000 are self-defined reactions
react_id	varchar(20)	PRI	R/d{5}	ENTRY line	
react_excl	int(1)		[012]		0 - default, reaction not checked yet 1 - reaction excluded for the reconstruction process 2 - reaction considered for the reconstruction process
total_equ	varchar(255)	SEC		EQUATION line	
main_equ	varchar(255)	SEC		see kegg_main_equ	might be modified
orig_main_equ	varchar(255)	SEC		reaction_main.lst (KEGG ligand)	total_equ, if not defined by KEGG
ec_list	varchar(255)	SEC	semlist	see kegg_ec_list	might be modified
orig_ec_list	varchar(255)	SEC	semlist	ENZYME line	
subsystem	varchar(20)		[CLPAV]	see Table 5.3	
path_list	varchar(255)		semlist	see Suppl. Table 5.7	
kegg_map_list	varchar(255)		semlist	PATHWAY line	
compounds				compound file (KEGG ligand)	
comp_id	varchar(6)	PRI	C/d{5}	ENTRY line	
comp_name	varchar(255)			see orig_comp_name	might be modified
orig_comp_name	varchar(255)			first NAME line	
comp_abbrev	varchar(20)				
enzymes				enzyme.dat (ENZYME)	
ec_no	varchar(15)	PRI	\d\.\d{1,2}\.\d{1,2}\.\d{1,3}	ID line	
ec_name	varchar(255)			DE line	
ec_abbrev	varchar(20)				
total_swiss_seq	int(4)			parsed DE line of SwissProt entries for EC numbers	
arch_swiss_seq	int(3)			see total_swiss_seq, only archaeal Swiss-Prot entries considered	

Table/Field	Type	Key	Format	Source	Comment
org_reactions					
react_id	varchar(20)	PRI	R/d{5}	see reactions.react_id	
react_exist	int(1)		[01]	0 - reaction does not exist in org (conf_level = 1-3), 1 - reaction exists in org (conf_level = 4-6)	
react_conf_level	int(1)		[123456]		confidence whether reaction exists in org (manual), confidence level: unlikely => 1, question => 2, possible => 3, insecure => 4, likely => 5, secure => 6
gene_list	varchar(255)	SEC	semlist	-	
ec_list	varchar(255)	SEC	semlist	-	
react_attr	varchar(255)		semlist	-	manual/autom - manual/automatic entry, no_genet_evid - no gene found for existing reaction, exper - experimental evidence whether reaction exists, altern_enz - alternative enzymes for reaction, no_seq/few_seq - 0/<10 sequences in SwissProt, gap_arch/few_arch - 0/<3 archaeal sequence in SwissProt
comm_id	varchar(20)	SEC	L/d{5}	-	
org_enz_genes					
gene_id	varchar(20)	PRI		gene id (Halolex)	
gene_conf_level	int(1)		[123456]		confidence whether correct EC number assigned to gene (manual), definitions as for react_conf_level
ec_list	varchar(255)	SEC	semlist	-	
gene_name	varchar(20)			gene abbreviation (Halolex)	might be modified
prot_name	varchar(255)			protein name (Halolex)	might be modified
gene_attr	varchar(255)				manual/autom - manual/automatical entry, exper - experimental evidence for gene function, subunit - probable subunit of an enzyme complex, paralog - paralogous gene found
comm_id	varchar(20)	SEC	L/d{5}		

Table/Field	Type	Key	Format	Source	Comment
org_comments					
comm_id	varchar(20)	PRI	L/d{5}		
comment	text				some abbreviations: pos_enz_activ – positive enzyme activity test (analog neg_enz_activ) NMR - NMR labelling studies TU - transcription unit
ref_list	varchar(255)		semlist	mostly PMIDs	
pic_list	varchar(255)		semlist	-	
comm_date	date		automatic		
org_enz_autoassign					
ec_no	varchar(20)	PRI	\d\.\d{1,2}\. \d{1,2}\. \d{1,3}		combined key since n:m relationship between EC number and gene id
gene_id	varchar(20)	PRI		gene id (Halolex)	
auto_conf_level	int(1)		[123456]		confidence whether correct EC number assigned to gene (automatic), definitions as for react_conf_level
best_ec_gene_comb	char(2)				E - best assignment for EC number O - best assignment for gene id
total_score	float(5,1)				average of single scores
BLAST_score	float(5,1)				score derived from E-value (-logx) of performed blast search
COG_score	float(5,1)				score from COG search
HMMER_score	float(5,1)				score from HMMER search
PFAM_score	float(5,1)				score from Pfam search

CHAPTER 6

Comparative Metabolic Analysis for Halophilic Archaea

Gene equipments of four halophilic archaea were compared amongst each other and differences in predicted metabolic capabilities were analysed. While *Natronomonas pharaonis* contains genes for the synthesis of all proteinogenic amino acids, *Halobacterium salinarum* lacks gene clusters for lysine, arginine, and branched chain amino acid synthesis. Furthermore it was predicted that different pathways for serine and proline biosynthesis exist in the two halophiles. Genes for folate biosynthesis and metabolism differ greatly between halophilic strains, and it is suggested that reduced folate levels in *H. salinarum* effect other metabolic pathways such as glycine biosynthesis. Respiratory chains of halophilic and other respiratory archaea share similar genes for pre-quinone electron transfer steps, but show great diversity with respect to genes encoding complex III and complex IV analogs. Thus, a high inter- as well as intraspecies variability of electron transfer modes likely occurs, which indicates adaptation to changing environmental conditions in extreme habitats. Haloarchaea seem to have adopted several strategies to utilize abundant cell material available in brines such as acquisition of catabolic enzymes, often plasmid-encoded *(H. salinarum, Haloarcula marismortui)*, secretion of hydrolytic enzymes *(H. salinarum, N. pharaonis)*, and elimination of biosynthetic gene clusters *(H. salinarum)*.

6.1 Introduction

In spite of their common halophilic environment, haloarchaea were described to use different types of carbon sources. While *Halobacterium salinarum* mainly grows on amino acids, other halophiles are known to utilize a variety of sugars (Altekar and Rangaswamy 1992). These carbohydrate-utilizing strains such as *Haloferax mediterranei* and *Haloarcula vallismortis* apply modified Embden-Meyerhof (EM) (Altekar and Rangaswamy 1992) and Entner-Douderoff (ED) (Danson and Hough 1992) pathways in order to catabolize fructose and glucose, respectively. *Halobacteria* can not only be distinguished by their catabolic

capabilities, but differ also in their capabilities to synthesize amino acids and cofactors. Whereas synthetic media for *H. salinarum* contain at least 10 amino acids and 3 vitamins (Oesterhelt and Krippahl 1973), *Natronomonas. pharaonis* does not require any cofactors and only one amino acid (leucine) for growth (Chapter 3.2.3).

Haloarchaea employ various systems to derive metabolic energy. Aerobic respiration has been observed but under anaerobic conditions respiration with nitrate (Lledo et al. 2004) and DMSO (Muller and DasSarma 2005) as terminal electron acceptor has also been shown in halophiles. Amongst all archaea, an even wider spectrum of respiratory redox reactions has been described, e.g. oxygen respiration while growing under chemolithotrophic conditions (H_2, S, FeS) as well as anaerobic respiration with sulphur or sulphate (Schafer et al. 1996). *H. salinarum* acquired two further systems to generate energy. When grown under phototrophic conditions halobacterial cells can establish a proton gradient by utilizing the light-driven proton pump, bacteriorhodopsin (Oesterhelt and Tittor 1989). Under anaerobic conditions, *H. salinarum* ferments arginine to ornithine to generate ATP using plasmid-encoded enzymes of the arginine deiminase pathway (Ruepp and Soppa 1996).

In this chapter, different capabilities of *H. salinarum* and *N. pharaonis* for the biosynthesis of amino acids were interpreted on the genomic level. Furthermore encoded respiratory chain components of halophilic and other aerobic archaea were compared amongst each other considering genome and literature data. The complete gene equipments of the four completely sequenced halophilic archaea have also been analysed in order to identify genomic differences that result in the observed differences in catabolic and anabolic capabilities in halophiles. Furthermore, comparative analysis of halophilic gene equipments points out different alternate enzymes (e.g. for β-carotene cleavage) and indicates a variety of metabolic pathways presumably employed by haloarchaea (e.g. nitrate assimilation and carbon dioxide fixation), which will be of interest for future investigations.

6.2 Amino acid metabolism in *Natronomonas* and *Halobacterium*

Biosynthesis pathways leading to all 20 proteinogenic amino acids were reconstructed for *H. salinarum* and *N. pharaonis* and stored within the metabolic database Pathnet. The so created organism-specific *reaction* entries for the two halophiles were then compared amongst each other. As a result, different capabilities for amino acid synthesis could be analyzed in detail (Table 6.1).

Table 6.1: **Comparison between amino acid biosynthesis pathways** in *H. salinarum* (HS) and *N. pharaonis* (NP). For each amino acid, the existence of synthesis pathways was assessed using metabolic reconstruction and experimental data (exp) from labeling studies (the respective literature references are given in the text). Amino acids supplemented in the synthetic medium (med) of *H. salinarum* are marked.

Amino acid synthesis	HS		Comment	NP	Comment
Glutamate family					
glutamate	**Y**	exp		**Y**	
glutamine	**Y**			**Y**	
proline	(**Y**)	exp	- experimental, but no genetic evidence	**Y**	
arginine	N	med	- no ornithine synthesis; - ornithine → Arg via urea cycle enzymes - Arg deiminase pathway present	**Y**	- one enzyme in ornithine synthesis missing; - ornithine → Arg via urea cycle enzymes
Aspartate family and alanine					
aspartate	**Y**	exp		**Y**	
asparagine	**Y**			**Y**	- also asparaginase found
lysine	N	med		**Y**	via diaminopimelate pathway; - three enzymes missing
threonine	**Y**	exp/med		**Y**	
methionine	N	med	- homoserine metabolism present, but missing enzyme for L-homo-Cys → Met	**Y**	
alanine	**Y**	exp	- via aspartate-pyruvate transaminase	**Y**	see HS
Serine family					
serine	**Y**	exp/med	- via the phosphorylated pathway, - one enzyme missing	**Y**	- via the phosphorylated and the unphosphorylated pathway
glycine	**Y**		- from Ser and Thr	**Y**	- from Ser
cysteine	**Y**			**Y**	
Branched chain aminoacids					
valine	N	med		**Y**	
leucine	N	med		(N)	- 2-isopropylmalate synthase (EC 2.3.3.13) seems to be N-terminally truncated
isoleucine	N	med		**Y**	
Aromatic amino acids					
phenylalanine	**Y**	exp/med	- first step of shikimate pathway unknown	**Y**	see HS
tyrosine	**Y**	med		**Y**	
tryptophane	**Y**			**Y**	
histidine	**Y**	exp	- one enzyme missing	**Y**	see HS

From the gene equipment of *Natronomonas*, a complete independence of this organism from supplemented amino acids was concluded and subsequently proven by the development of a synthetic medium without amino acids except leucine (Chapter 3.2.3). Leucine requirement is probably due to a mutation in the N-terminal region of 2-isopropylmalate synthase (EC 2.3.3.13, NP2206A). In contrast to *Natronomona*s, *Halobacterium* lacks gene clusters for the synthesis of branched chain amino acids, lysine, praline, and of the arginine precursor ornithine. However, the biosynthesis of proline has already been proven experimentally (Ghosh and Sonawat 1998). Methionine is also predicted to be an essential amino acid,

since the gene encoding THF-dependent methionine synthase is absent in the *H. salinarum* genome.

For amino acid synthesis pathways existing in both halophiles, the gene equipment is analogous except for enzymes involved in serine and glycine biosynthesis. *Natronomonas* seems to synthesize serine via non-phosphorylated and phosphorylated pathways, whereas *Halobacterium* possesses only enzymes for the latter pathway. Furthermore, glycine is derived from serine in both species, but *H. salinarum* might additionally synthesize glycine through the cleavage of threonine. Detailed differences in amino acid biosynthetic pathways between the two halophiles are discussed in the following sections.

6.2.1 Glutamate family

The biosynthesis of glutamate from the TCA cycle intermediate 2-oxoglutarate (2-OG) is an important metabolic conversion in *Halobacterium* as shown by NMR labeling studies. Labels from pyruvate, alanine (Ghosh and Sonawat 1998), acetate, and glycerol (Ekiel et al. 1986) were mainly found to be incorporated into glutamate, and a considerable part of the flux through the TCA cycle was shown to be channelled to glutamate. Three paralogous glutamate dehydrogenases genes were found in *Natronomonas* (NP1582A, NP1806A, NP6184A) and *Halobacterium* (OE1270F, OE2728R, OE1943F), and activity of two glutamate dehydrogenases with $NADP^+$ (OE1943F) and NAD^+ (OE1270F) cofactor specifity was proven for *H. salinarum* (Bonete et al. 1987; Bonete et al. 1989; Bonete et al. 1990; Perez-Pomares et al. 1999; Hayden et al. 2002). The derived glutamate can be used for glutamine synthesis by glutamate-ammonia ligase (EC 6.3.1.2) in both halophiles (OE3922R, NP0076A, NP4376A). Apart from the three glutamate dehydrogenase paralogs, *Natronomonas* possesses also a gene for glutamate synthase (NP1794A), which converts 2-OG and glutamine to glutamate and which is part of the proposed ammonia assimilation pathway (Figure 3.2 in Chapter 3.2.4).

Glutamate is the precursor of further amino acids. For proline synthesis, glutamate is phosphorylated and subsequently reduced to L-glutamate 5-semialdehyde by glutamate kinase (ProB) and glutamate 5-semialdehyde dehydrogenase (ProA), respectively. Glutamate 5-semialdehyde forms a pyrroline cycle (in a spontaneous reaction), which is reduced to proline by pyrroline-5-carboxylate reductase (ProC). In *Natronomonas*, a proline synthesis cluster *proCBA* was found, which is absent in the *Halobacterium* genome. However, proline signals were derived from labelled alanine and pyruvate in *Halobacterium* (Ghosh and Sonawat 1998) indicating that this organism might be capable of proline synthesis via another pathway, e.g. via 1-pyrroline-5-carboxylate dehydrogenase (EC 1.5.1.2) and proline dehydrogenase (EC 1.5.99.8) (PutA). The latter enzyme was found to be encoded in *H. salinarum* (OE3955F), but is missing in *N. pharaonis*. Notably, *Haloquadratum*

walsbyi also possesses a *proCBA* gene cluster and lacks a proline dehydrogenase homolog as *Natronomonas*, while the *Haloarcula marismortui* gene equipment is analogous to *Halobacterium* (Supplemental Table 6.3).

Natronomonas and *Halobacterium* also differ significantly in their gene equipment for arginine synthesis and metabolism. Whereas *Natronomonas* encodes an *argXCBD* cluster with ornithine synthesis enzymes and a probable transcription regulator ArgX, *Halobacterium* lacks the complete gene set for *de novo* synthesis of arginine. Instead halobacterial arginine requirements are covered by uptake of external arginine via an experimentally verified arginine-ornithine antiporter (OE5204R) (J. Tittor, pers. comm.). Adjacent to this transporter gene, an *arcRACB* cluster is found on the plasmid PHS3, which encodes enzymes and a probable transcription regulator for the described arginine deiminase pathway (Ruepp and Soppa 1996). This fermentative pathway leading from arginine to ornithine is employed by *H. salinarum* cells under anaerobic conditions. In all halophiles, ornithine can be likely converted to arginine via the urea cycle enzymes ornithine carbamoyltransferase (EC 2.1.3.3), argininosuccinate synthase (EC 6.3.4.5), and argininosuccinate lyase (EC 4.3.2.1).

H. salinarum and *H. marismortui* (but not *N. pharaonis* and *H. walsbyi)* might also be capable of glutamate fermentation via the β-methylaspartate pathway as observed for *Clostridium tetanomorphum* (Brecht et al. 1993), since they exhibit required methylaspartate mutase (EC 5.4.99.1) and methylaspartate ammonia-lyase (EC 4.3.2.1) homologs that are encoded in the *mam* gene cluster (Figure 6.4 in Chapter 6.2.4). Glutamate is likely degraded to mesaconate and then to citramalate, which was shown to be subsequently split to pyruvate and acetate in thermophilic anaerobic bacteria (Plugge et al. 2001). However, the required enzymes for the latter degradation steps are unknown yet.

6.2.2 Aspartate family

Aspartate is derived from the TCA cycle intermediate oxaloacetate. Aspartate signals have been detected from labelled acetate and glycerol in *Halobacterium* (Ekiel et al. 1986). Several aspartate transaminase (EC 2.6.1.1) homologs were identified in *N. pharaonis* (NP0824A, NP1666A, NP4024A, NP4410A) and *H. salinarum* (OE1755F, OE1944R, OE2619F). These might not only be involved in oxaloacetate conversion to aspartate but also in pyruvate transamination to alanine, since experiments showed that the halobacterial enzyme involved in alanine synthesis uses aspartate not glutamate as amino group donor (Bhaumik and Sonawat 1994). Consistent with this, no homologs of glutamate-pyruvate transaminase (EC 2.6.1.2) have been found in the two halophilic strains. Aspartate is likely transaminated to asparagine via asparagine synthase (EC 6.3.5.4) in haloarchaea.

In a pathway analogous to proline synthesis from glutamate, aspartate can be phosphorylated and subsequently reduced to L-aspartate 4-semialdehyde. This intermediate is then metabolized to homoserine in both halophiles. In *Natronomonas*, homoserine is further converted to lysine via the diaminopimelate (DAP) pathway using enzymes of the respective biosynthesis gene cluster *dapABD* as well as enzymes encoded by *lysA* and *argG* located elsewhere in the genome. However, the haloalkaliphile lacks some genes (*dapE, dapF*) required within the DAP pathway which are found in the lysine biosynthesis cluster of *H. walsbyi* (*dapABD-lysA-dapFE*) (Supplemental Table 6.3). Unknown alternative non-orthologous enzymes likely bridge observed pathway gaps in *N. pharaonis* (and *H. marismortui)* lysine biosynthesis pathways. Although *H. salinarum* possesses a *dapA* homolog encoding the first DAP pathway enzyme all other required lysine biosynthesis genes are missing.

Homoserine is phosphorylated and hydrolyzed to threonine by homoserine kinase (EC 2.7.1.39) and threonine synthase (EC 4.2.3.1) in haloarchaea, and consistent with this, threonine signals were detected from labeled acetate in *Halobacterium*, if threonine was omitted from the growth medium (Ekiel et al. 1986). Homoserine can also be converted to O-acetyl-L-homoserine (but not to O-succinyl-L-homoserine) and then to the methionine precursor, L-homocysteine, via a variety of thiol-lyase and -synthase reactions. Several paralogous thiol-lyases were found in *Halobacterium* and *Natronomonas*, which cluster into two orthologous groups (Figure 5.3 in Chapter 5.2.2). However, exact specifity of these paralogs is difficult to determine by similarity search. Thus, O-acetyl-L-homoserine might be converted to L-homocysteine directly or via cystathionine depending on the actual specifity of found thiol-lyases. *Natronomonas*, but not *Halobacterium*, is further able to convert L-homocysteine by THF-dependent methionine synthase (EC 2.1.1.14), of which two adjacent paralogs *metE_1* and *metE_2* have been found in the *N. pharaonis* genome.

6.2.3 Serine family

Serine can be derived through two different pathways (Figure 6.1). Within the phosphorylated pathway, glycerate 3-phosphate, a glycolytic intermediate, is oxidized to 3P-hydroxypyruvate, which is further converted to serine in a transaminase and a phosphatase reaction. Alternatively, unphosphorylated glycerate is directly oxidized to hydroxypyruvate and then transaminated. Both, *Halobacterium* and *Natronomonas* possess adjacent hydratase (EC 1.1.1.95) and phosphatase (EC 3.1.3.3) genes (*serAB*) required for the phosphorylated synthesis pathway. However, no phosphoserine transaminase (EC 2.6.1.52) genes for the intermediate pathway step could be identified. However, since serine synthesis in *Halobacterium* has been shown by NMR (Ekiel et al. 1986), a functional serine synthesis pathway has to exist in this organism. In both halophiles homologs to serine-pyruvate

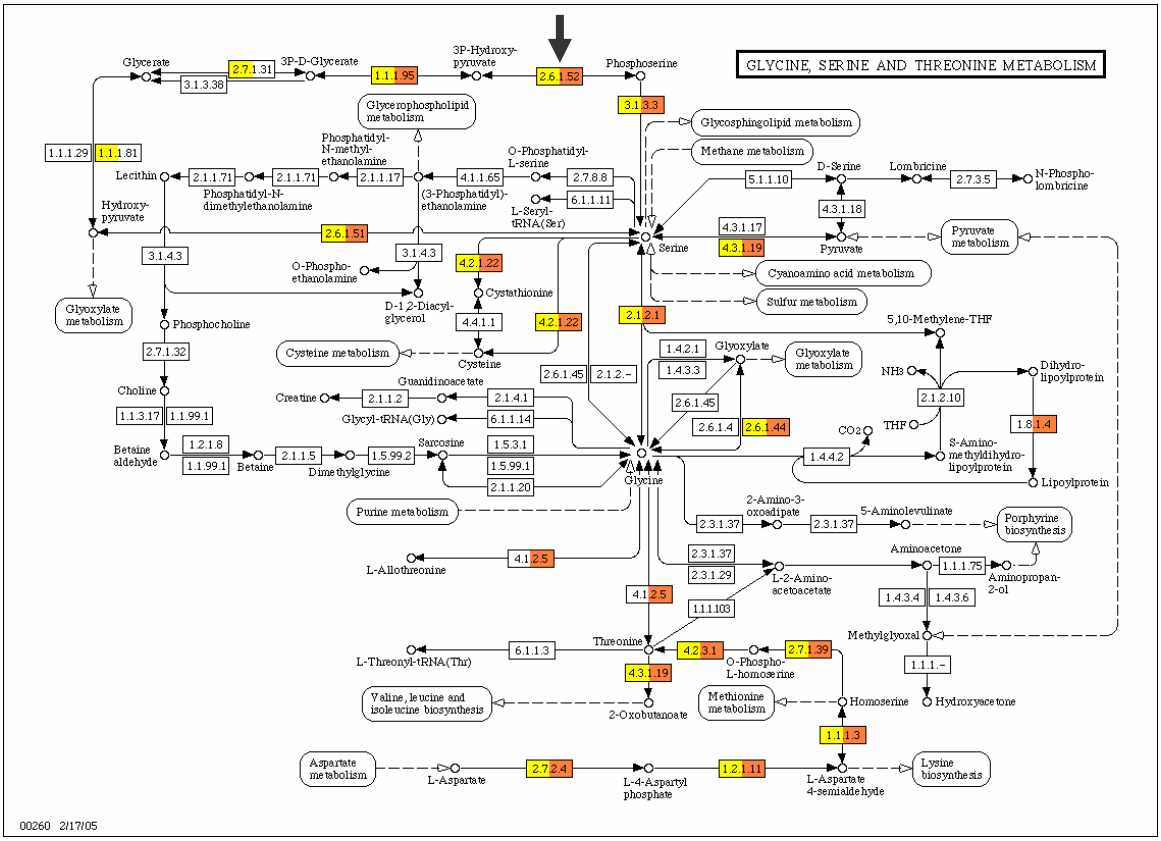


Figure 6.1: **Comparison of glycine, serine, and threonine metabolism** (KEGG-Map00260) in *N. pharaonis* (yellow) and *H. salinarum* (orange) (from Muller 2005). Enzymes for which genes were identified in the genome are colour-coded. The vertical arrow marks an enzyme (EC 2.6.5.52) of the phosphorylated pathway for serine synthesis, which is proposed to be catalyzed by an enzyme (EC 2.6.1.51, OE4391F, NP0884A, NP2578A) involved in the unphosphorylated pathway variant.

aminotransferase (EC 2.6.1.51, OE4391F, NP0884A, NP2578A) have been identified, and the encoded enzyme might not only catalyze the conversion of hydroxypyruvate but also the analogous reaction with phosphorylated hydroxypyruvate (EC 2.6.1.52). Proximity of the serine-pyruvate aminotransferase gene with *serAB* on the halobacterial genome further supports this suggestion. Since *Natronomonas* possesses also a hydroxypyruvate reductase (EC 1.1.1.81) gene (NP1162A), not only the phosphorylated but also the unphosphorylated serine synthesis pathway is indicated for this species.

Natronomonas and *Halobacterium* are capable to convert serine to glycine in a THF-dependent reaction (EC 2.1.2.1). However, *Halobacterium* might also convert threonine to glycine by threonine aldolase (EC 4.1.2.5) (OE4436R). In the future, it will be interesting to determine, to which extent glycine is derived from serine and threonine in *Halobacterium.* Furthermore, metabolic fluxes for the two potential serine synthesis pathways in *Natronomonas* should be analysed.

Similar to the reactions of the homoserine metabolism, serine can be converted to O-acetyl-serine in both halophiles and then to cysteine by the incorporation of sulfide (Figure 6.2). Sulfide needs to be assimilated from sulfate by reduction of sulfate via adenylylsulfate (APS), 3P-adenylylsulfate (PAPS), and sulfite. Although a gene for the small subunit *(cysD)* of the sulfate adenylyltransferase (EC 2.7.7.4) is present in *Natronomonas* (NP4570A) and *Halobacterium* (OE1684F), the gene encoding the GTPase subunit *(cysN)* is missing in the two halophiles and other archaea. In some archaea such as *Archaeoglobus fulgidus* dissimilatory sulfate adenylyl-transferases (Sat) (Sperling et al. 1998) have been described which are also absent in *N. pharaonis* and *H. salinarum*. Homologs for enzymes required for two subsequent sulfate assimilation steps (*cysC, cysH*) are present in some archaea but were not found in other of the haloarchaeal genomes. However, in both halophiles sulfite dehydrogenase (EC 1.8.2.1) homologs (NP2718A, NP2044A, OE2569R) for sulfite oxidation were found. These might also function in reverse mode reducing sulfate. In *Natronomonas*, several reductase genes (e.g. NP4004A) are present, which encode potential sulfite reductases producing the required sulfide for cysteine synthesis.

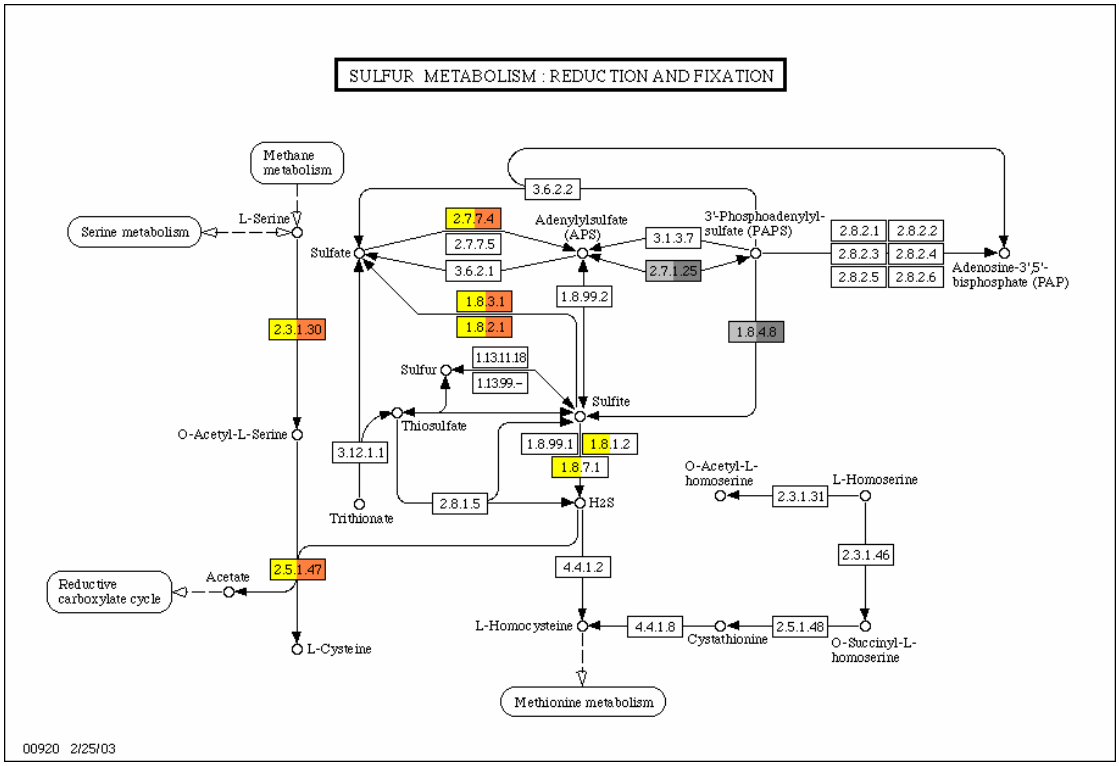


Figure 6.2: **Sulfur metabolism in *N. pharaonis* (yellow) and *H. salinarum* (orange)** (from Muller 2005). Enzymes for which no genes were identified in the genome are shown in grey. Cysteine synthase (EC 2.5.1.57) introduces a thiol-group into O-acetyl-L-serine. The required sulfide needs to be assimilated from sulfate. However, most genes for the sulfate assimilation pathway via APS and PAPS are absent in the two halophiles.

6.2.4 Biosynthesis of branched chain amino acids

The *de novo* synthesis of valine, leucine, and isoleucine from pyruvate occurs only in *N. pharaonis*, whose genome contains a gene cluster for branched-chain amino acid synthesis (Figure 6.3). However, the 2-isopropylmalate synthase (EC 2.3.3.13) gene *leuA_1* (NP2206A) found within this cluster seems to be corrupted at the start of the gene. Tblastn searches with 2-isopropylmalate synthases from other species indicate a highly conserved sequence in front of the first valid start codon (ATG, GTG) for the *leuA_1* gene. Thus, the function of the encoded 2-isopropylmalate synthase is likely limited or abolished, and leucine synthesis pathway subsequently interrupted in *Natronomonas*. Consistent with this, leucine cannot be omitted from the synthetic growth medium for this species (Chapter 3.2.3). Thus, two other *leuA* paralogs (NP0624A, NP2994A) found elsewhere in the *N. pharaonis* genome or encoded thiol-lyases catalyzing similar reactions as 2-isopropylmalate synthase seem not be sufficient to complement *leuA_1*. *H. salinarum* lacks all genes required for branched-chain amino acid synthesis except for the *leuC* gene (EC 4.2.1.33). Its genome contains further an *ilvE* gene, which is required for both, biosynthesis and degradation of branched-chain amino acids.

For isoleucine synthesis, a second precursor, 2-oxobutyrate, is required, which is synthesized from threonine via threonine-ammonia lyase (EC 4.3.1.19) in *N. pharaonis* (NP1076A) but also in *H. salinarum* (OE3931R). Labelling studies in *Haloarcula hispanica* showed that 2-oxobutyrate was not only derived from threonine but also from glutamate via the β-methylaspartate pathway (Figure 6.4) (Hochuli et al. 1999). Enzymes of this pathway were only found in *H. salinarum* and *H. marismortui*, though. The fact that *Halobacterium* encodes several enzymes for 2-oxobutyrate synthesis but lacks further enzymes for

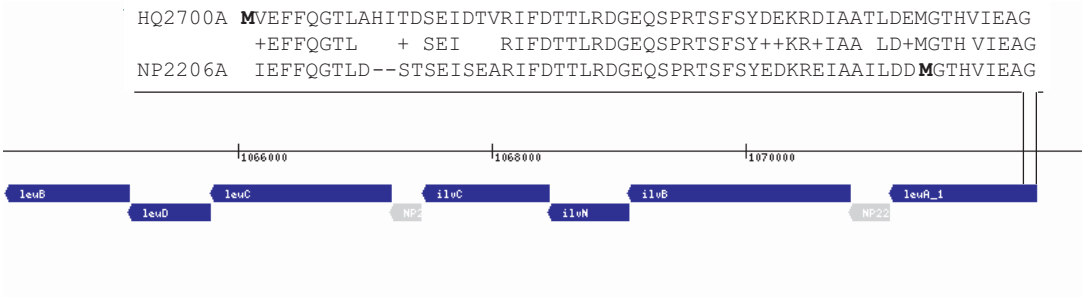


Figure 6.3: **Gene cluster for the biosynthesis of branched-chain amino acids** in *N. pharaonis*. Genes encoding enzymes of the branched-chain amino acid biosynthesis pathway (blue arrows) are organized in three probable transcription units (*leuA_1*, *ilvBNC*, and *leuCDB*). Grey arrows show spurious ORFs within the gene cluster. The alignment of the N-terminal region of 2-isopropylmalate synthase (EC 2.3.3.13) from *H. walsbyi* (HQ) with the genomic region of the *N. pharaonis* (NP) *leuA_1* gene by tblastn shows that the first available start codon, which correspond to the marked Met residues (bold) differs in the two halophiles. In the *N. pharaonis* sequence, the first possible start is located already within the region with similarity to the HQ sequence. Thus, a mutation at the start of the *leuA_1* gene is indicated. This probably results in a non-functional 2-isopropylmalate synthase, and the interruption of the leucine biosynthesis pathway in *Natronomonas*.

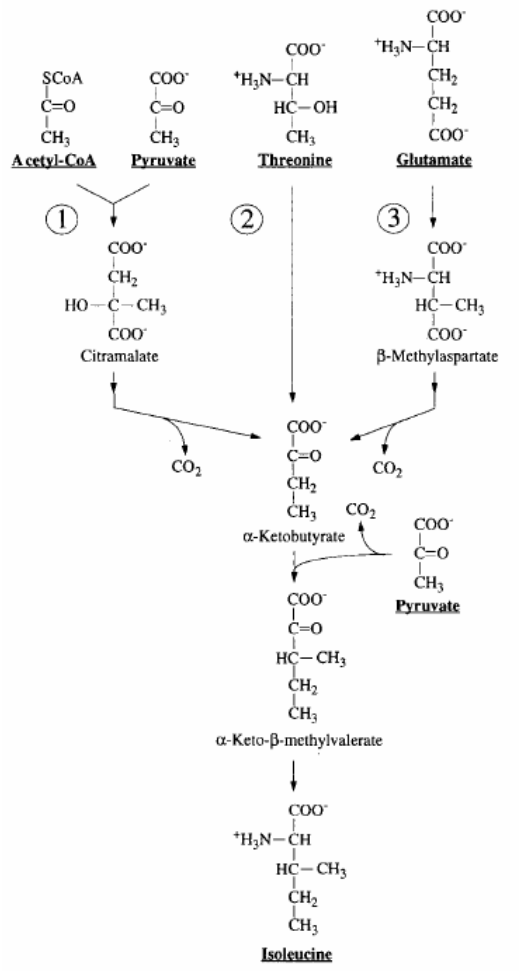


Figure 6.4: **Different routes for isoleucine biosynthesis** (from Hochuli et al. 1999). Pyruvate and α-ketobutyrate (= 2-oxobutyrate) are converted to α-keto-β-methylvalerate and subsequently to isoleucine. 2-oxobutyrate can be delivered via several pathways. Pyruvate and acetyl-CoA are the precursors for 2-oxobutyrate synthesis via the pyruvate pathway (route 1), threonine is the precursor for the synthesis via the threonine pathway (route 2), and glutamate is the starting point of the glutamate pathway better known as β-methyl-aspartate pathway or glutamate fermentation (route 3). Threonine ammonia-lyase (EC 4.3.1.19) is required for route 2, and was found in all four completely sequenced halophiles. Route 1 employs citramalate lyase (EC 4.1.3.22), for which no sequences are available yet in public databases. Route 3 employs cobalamin-dependent methylaspartate mutase (EC 5.4.99.1) and methylaspartate ammonia-lyase (EC 4.3.1.2) as described for *Clostridium tetanomorphum* (Brecht et al. 1993). The genes that encode these two enzymes were found to be clustered in *H. salinarum* and *H. marismortui (mamABC)*, but are absent in *N. pharaonis* and *H. walsbyi.* NMR studies in *Haloarcula hispanica* indicated that 56% of isoleucine is synthesized via the threonine pathway and 44% via the pyruvate pathway (Hochuli et al. 1999).

isoleucine synthesis implies that 2-oxobutyrate might also be a precursor for other metabolic pathways. 2-oxobutyrate may further be derived from methionine by methionine gamma-lyase (EC 4.4.1.11) whose activity has been detected in *H. salinarum* (Nordmann et al. 1994). As a result of this reaction, volatile methanthiol is released from the cells. This phenomenon is related to the halobacterial sensory system, but the specific function of methanethiol release by *Halobacterium* is yet unknown.

6.2.5 Biosynthesis of aromatic amino acids

Phenylalanine, tyrosine, and tryptophan are synthesized from two precursors, erythrose-4P and PEP, via the shikimate pathway (Figure 6.5). However, most archaea lack enzymes for the non-oxidative pentose phosphate pathway, which is the canonical pathway leading to synthesis of erythrose-4P. For methanogenic archaea, an alternative triose carboxylation pathway was proposed to deliver erythrose-4P (Choquet et al. 1994) and NMR labeling studies in *Methanococcus maripaludis* even suggested that archaeal precursors and early steps of aromatic amino acid synthesis differ from the known shikimate pathway (Tumbula et al. 1997). Consistent with these previous findings, no 3-deoxy-7-phosphoheptulonate

synthase (EC 2.5.1.54) and 3-dehydroquinate synthase (EC 4.2.3.4) homologs are found in methanogenes, halophiles, and *A. fulgidus*. Only in crenarchaeota, Pyrococci, and *Thermoplasma* these two enzymes could be identified. Recently, an alternative non-orthologous 3-dehydroquinate synthase (EC 4.2.3.4) gene (COG1465, OE1475F, NP2238A) was proposed for halophilic and methanogenic archaea, which is often encoded adjacent to the 3-dehydroqinate dehydratase (EC 4.2.1.10) gene (COG0710, OE1477R, NP2240A). This alternative enzyme exhibits a reverse phylogenetic profile to the known bacterial variant of 3-dehydroquinate synthase (COG0337), but its function has not been experimentally verified yet. Genes for all enzymes of later shikimate pathway steps leading from 3-dehydroquinate to phenylalanine, tyrosine, and tryptophane via the pathway branch points, chorismate and prephenate, were found in haloarchaea (Figure 6.5). Consistent with this, phenylalanine was shown to be synthesized *de novo* in *H. salinarum* (Ekiel et al. 1986).

A second pathway apart from the *de novo* synthesis for the synthesis of aromatic amino acids has recently been described for *M. maripaludis* (Porat et al. 2004). In this archaeon, aromatic amino acids are also synthesized via incorporation of exogenous aryl acids via two-subunit indolepyruvate oxidoreductase (Ior), which is homologous to other ferredoxin-dependent oxidoreductases. Interestingly, NMR labeling studies also exclude the sole operation of the standard shikimate pathway for the synthesis of tyrosine for *H. hispanica* (Hochuli et al. 1999). The occurrence of aromatic amino acid synthesis via indolepyruvate oxidoreductase is unlikely for halophilic archaea, though, since only two ferredoxin-dependent oxidoreductases complexes, pyruvate- and 2-oxoglutarate-ferredoxin oxidoreductase, are encoded within their genomes.

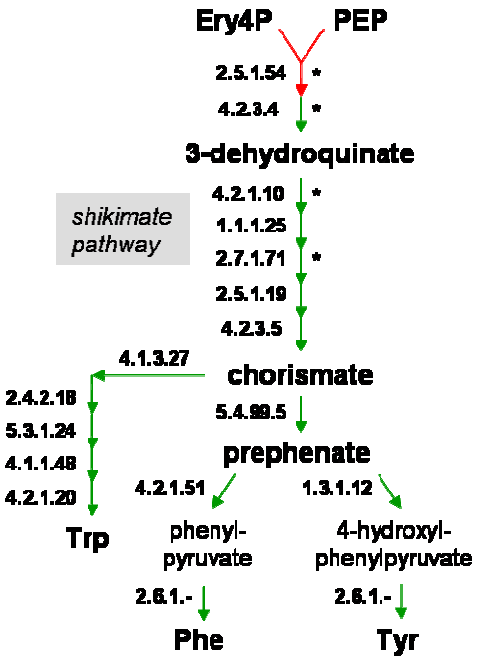


***Figure 6.5:* Aromatic amino acids synthesis via the shikimate pathway.** The precursers for the archaeal shikimate pathway remain enigmatic, since genes encoding non-oxidative pentose phosphate pathway enzymes that are required for erythrose-4P synthesis and the enzyme for the first step of the shikimate pathway (EC 2.5.1.54) are absent in most archaea. An alternative first step in the shikimate pathway of *Methanococcus maripaludis* was indicated by labeling studies (Tumbula et al. 1997).

* Non-orthologous enzymes in bacteria and archaea:
- 3-deoxy-7-phospho-heptulonate synthase (EC 2.5.1.54): three types (COG0722/COG3200/COG2876) in bacteria, COG2876 also in some archaea, archaeal type unkown;
- 3-dehydroquinate synthase (EC 4.2.3.4): COG0337 in bacteria and some archaea, COG1465 proposed for archaea;
- 3-dehydroquinate dehydratase (EC 4.2.1.10): COG0757 in bacteria, COG0710 in bacteria and archaea
- shikimate kinase (EC 2.7.1.71): COG0703 in B, COG1685 in archaea (Daugherty et al. 2001)

Histidine biosynthesis genes were found partly clustered on the halophilic genomes. One pathway gap was observed for the histidine biosynthesis pathway starting from ribose-5P, but activity of an unknown alternative histidinol-phosphatase (EC 3.1.3.15) is likely, since histidine signals were observed from labeled substrates in *H. salinarum* (Ekiel et al. 1986).

6.3 Respiratory chains of haloarchaea and other archaea

Classical respiratory chains described for mitochondria consist of 4 membrane complexes which catalyze a series of redox reactions; NADH dehydrogenase (complex I), succinate dehydrogenase (complex II), ubiquinone-cytochrom c reductase (complex III), and cytochrome-*c*-oxidase (complex IV). Electrons from various sources are fed into complexes I and II, subsequently transferred to a ubiquinone carrier, to complex III, to a mobile cytochrome-*c*, and finally to complex IV. Along the electron transport path, energy is released which is utilized to pump protons across the membrane (complexes I, III, and IV). The resulting proton gradient is used by the ATP synthase complex to generate ATP. The final complex IV transfers electrons onto molecular oxygen which is reduced to water. As discussed below, respiratory chains from respiratory archaea, e.g. *Halobacteriales*, *Sulfolobales*, and *Thermoplasmatales*, (and from repiratory bacteria) differ significantly from classical chains in respect to complex compositions and functions.

6.3.1 Respiratory chains of haloarchaea

For all four completely sequenced haloarchaeal species genes encoding respiratory chain and ATP synthase components have been identified (Figure 6.6). Thus, haloarchaea are likely capable of ATP production under oxygen consumption as experimentally shown for *N. pharaonis* (Chapter 3.2.5) and *H. salinarum.* In halophiles and other aerobic archaea, *nuo* genes with similarity to type I NADH dehydrogenase subunits have been found. However, *nuoE, nuoF*, and *nuoG* encoding subunits, which form the NADH acceptor module, are missing, indicating that NADH cannot be dehydrogenated by the *nuo* complex. For *H. salinarum*, a functional type I NADH dehydrogenase has been excluded experimentally, since typical resonances of Fe-S cluster were absent upon NADH addition to the halobacterial membrane (Sreeramulu et al. 1998). Instead, inhibitor studies indicated a type II NADH dehydrogenase in *Halobacterium* (Sreeramulu et al. 1998) and other aerobic archaea (Schafer et al. 1996). NADH dehydrogenase type II is able to feed electrons derived from NADH into the respiratory transport chain as complex I, but cannot translocate protons across the membrane. Consistent with this, candidate genes (OE2307F, NP3508A) for type II

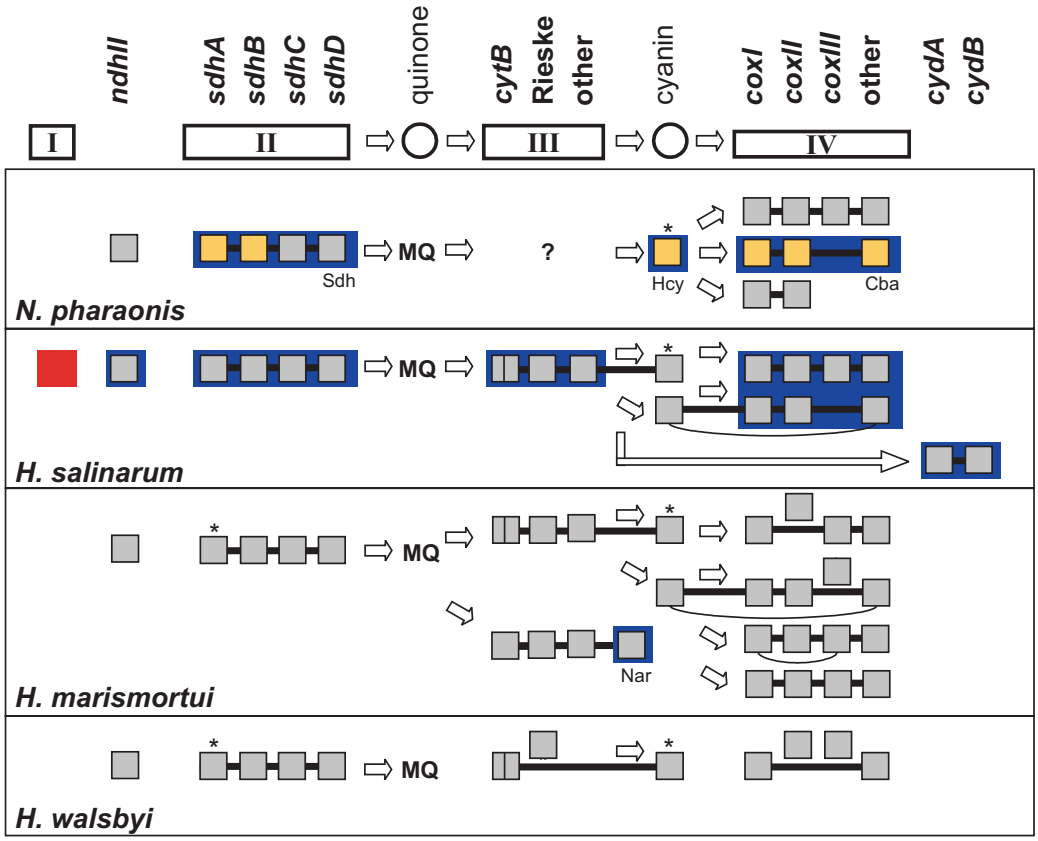


Figure 6.6: **Respiratory chain profile for halophilic archaea.** Subunits (boxes) of respiratory complexes are often encoded adjacent in the genomes (straight connections) and can be fused to each other (arcuate connections). The proposed electron flow between respiratory complexes is indicated by arrows. Complexes that have been characterized experimentally are marked by blue boxes, and sequenced subunits of isolated complexes are coloured yellow. The asterisks indicate respiratory chain components, for which paralogs were found. Homologs to complex I subunits have been found, but genes encoding the NADH acceptor module are missing. A functional NADH-dehydrogenating complex I has been excluded for *H. salinarum* by inhibitor studies (red box) (Sreeramulu et al. 1998). Instead, NADH dehydrogenase type II has been suggested, which is not capable of proton translocation. MQ - menaquinone.

NADH dehydrogenase were detected in *H. salinarum* and all other haloarchaea, when searching with the experimentally verified *Acidianus ambivalens* sequence *(noxA)* (Gomes et al. 2001).

Succinate dehydrogenase complexes (Sdh) of *Natronomonas* and *Halobacterium* have been characterised experimentally, and were shown to contain heme *b* and FAD cofactors (Mattar, PhD thesis 1996) and to be affected by the typical Sdh inhibitor malonate (Sreeramulu et al. 1998), respectively. The *sdh* genes are highly similar amongst different halophiles, and are co-transcribed in the same order *sdhCDBA* (Mattar, PhD thesis 1996). Some subunits of the *N. pharaonis* Sdh complex (SdhA, SdhB) have been validated by N-terminal protein sequencing. Genes for the electron transfer flavoprotein (Etf) has been detected in all halophiles except *H. walsbyi*. Etf complexes are specific electron acceptor complexes that transfer electrons resulting from fatty acid degradation into the respiratory chain via Etf

dehydrogenase, homologs of which were found in halophiles (e.g. NP4564A). Electrons originating from NADH dehydrogenase as well as from Sdh and Etf complexes are presumably transferred to menaquinone in halophiles (*men* gene clusters for menaquinone synthesis present).

Clustered *pet* genes which probably encode complex III analogs have been found in the genomes of *H. salinarum*, *H. walsbyi*, and *H. marismortui* (Figure 6.6). Probable subunits of this complex resemble Rieske, cytochrome-*b6*, and 17K proteins of bacteria. However, as in all other archaea, no cytochrome-*c1* homologs were found. Preparations of both, *H. salinarum* and *N. pharaonis* complex III analogs (Sreeramulu et al. 1998; Scharf et al. 1997), suggested the presence *b*- and *c*-type cytochromes but only the *Halobacterium* complex exhibited a typical Rieske Fe-S cluster. Unexpectedly, no homologs to complex III components were found in the *N. pharaonis* genome, and it can be speculated that this species acquired an alternative complex III analog (Chapter 3.2.5). On the other hand, the *H. marismortui* genome exhibits genes for a second complex III analog, namely a Rieske Fe-S protein and a cytochrome-*b*-like subunit. The genes are located together with the dissimilatory nitrate reductase (Nar) of this organism (Yoshimatsu et al. 2000). Therefore, the encoded genes for a second complex III analog might be involved in an anaerobic respiratory chain.

Since cytochrome-*c* homologs are absent in all halophiles, an alternative carrier between complexes III and IV is required. Copper-containing halocyanin has been suggested to assume this function, since characterized terminal oxidase complex of *Natronomonas*

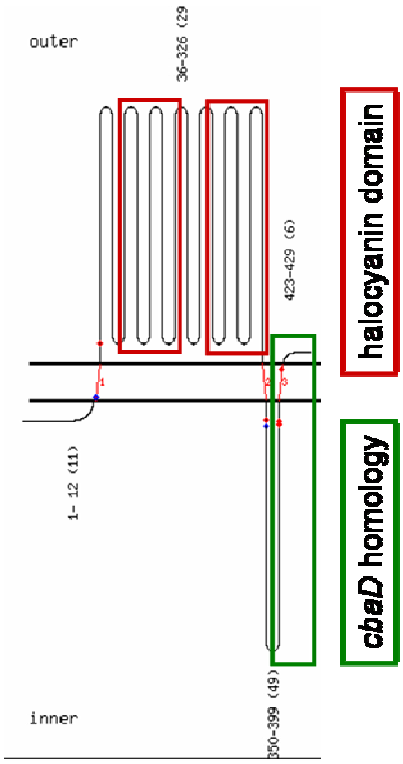


Figure 6.7 **Schematic representation of a fusion of respiratory components** observed for *H. salinarum* and *H. marismortui.* Two halocyanin domains (red) are fused to a *cbaD* domain (green) encoding a subunit of a terminal oxidase complex. The fusion indicates that halocyanin functions as electron carrier between the complex III analogon and the terminal oxidase complex in halophiles. Interestingly, the *cbaD* gene of *H. salinarum* str. NRC-1 which is closely related to *H. salinarum* str. R1 reveals only one halocyanine domain.

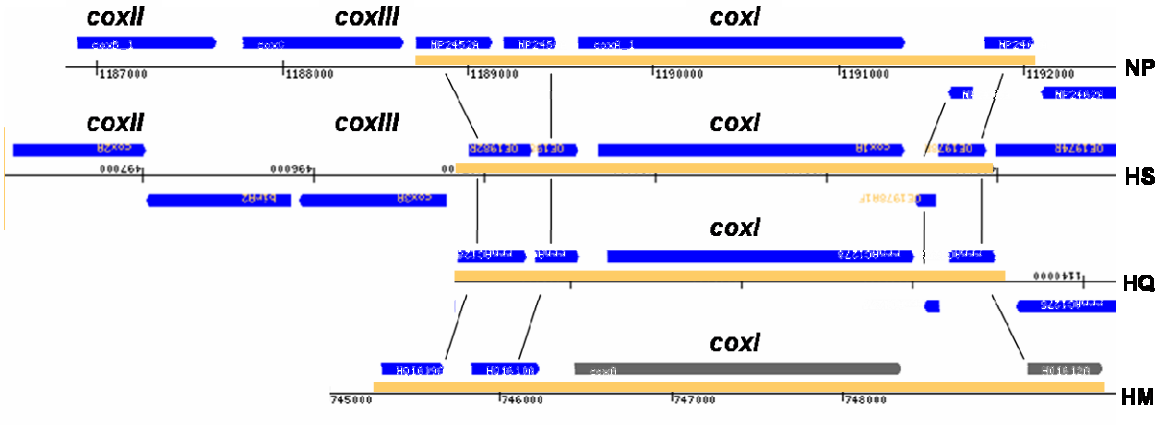


Figure 6.8: **Conserved gene cluster encoding a probable proton-translocating terminal oxidase** (Cox) in four halophilic archaea (NP - *N. pharaonis*, HS - *H. salinarum*, HQ - *H. wasbyi*, HM - *H. marismortui)*. Around the *coxI* subunits (orange line) several small genes are conserved in all strains (black vertical lines). These conserved genes probably encode minor subunits of halophilic Cox complexes.

requires halocyanin (Mattar, PhD thesis 1996). Neighbourhood of halocyanin genes to the *pet* cluster in the haloarchaea, and fusion of a terminal oxidase subunit *cbaD* with two halocyanin domains in *H. salinarum* str. R1 and *H. marismortui* further support this suggestion (Figure 6.7). Interestingly, one halocyanin domain was eliminated from the fused gene by a 420 bp in-frame deletion in *H. salinarum* str. NRC-1 (F. Pfeiffer, pers. comm.).

The four haloarchaeal genomes show clustered genes for cytochrome-*c*-oxidase-like complexes (Cox) (Figure 6.8). The gene cluster contains *coxI*, in *N. pharaonis* and *H. salinarum* also *coxII* and *coxIII* genes, as well as conserved hypothetical proteins that might encode additional small subunits of this Cox complex. The *coxI* homolog reveals an NHN signature, which is usually found in proton-translocating terminal oxidases (Schafer et al. 1996). In all halophiles except *H. walsbyi*, a second *cox* cluster has been identified containing a *coxI* homolog lacking an NHN motif. For *N. pharaonis*, this second terminal oxidase complex, an ba_3-type complex (Cba), has already been characterized in detail (Mattar and Engelhard 1997). *Natronomonas* and *Haloarcula* encode one and two further Cox complexes, respectively, and the *H. salinarum* likely exhibits a plasmid-encoded cytochrome-*d*-oxidase complex. Indeed, a *d*-type cytochrome was found when growing *H. salinarum* at low oxygen tension (Sreeramulu et al. 1998).

6.3.2 Respiratory chains of archaea

Obligate and facultative aerobes are found amongst crenarchaeota *(Sulfolobus, Acidianus)* and euryarchaeota *(Halobacteriales, Thermoplasma)*, and components of their respiratory chains have been described previously (Schafer et al. 1996). Since complete genome sequences are available except for *Acidianus*, published experimental results on archaeal

respiratory chains could be compared with data derived from genome analysis in order to obtain a pathway profile for respiratory archaea (Figure 3.3A in Chapter 3.2.5).

Respiratory chains of all aerobic archaea resemble each other in the early steps of electron transport (pre-quinone steps). As described for haloarchaea above, only an incomplete set of *nuo* genes could be identified in all respiratory archaea, and it was proposed through experimental data from *Acidianus* and *Sulfolobus* that type II NADH dehydrogenase catalyzes dehydrogenation of NADH instead (Schafer et al. 1996). This enzyme was isolated for *Acidianus ambivalens* and could be assigned to the *noxA* gene. All species except *Thermoplasma acidophilum* exhibit homologs of this gene, thus, a non-proton translocating NADH dehydrogenase seems indeed to be ubiquitous in archaea.

Genomes of *Thermoplasma acidophilum*, *Sulfolobus solfataricus*, and *Acidianus ambivalens* contain succinate dehydrogenase gene clusters with the same gene order, *sdhABCD*, differing from the gene order observed in halophiles. Although similar on the genomic level, succinate dehydrogenase complexes of non-halophilic archaea differ on the molecular level. Whereas three classical Fe-S clusters were observed for Sdh complexes of *T. acidophilum* and *Sulfolobus tokadaii*, complexes of *Sulfolobus acidocaldarius* and *A. ambivalens* revealed unusual Fe-S compositions (Sreeramulu et al. 1998). Organism-specific quinones likely function as mobile electron carriers in archaeal respiratory chains, and for *A. ambivalens* caldariella quinone was reported to transfer electrons (Gomes et al. 2001).

Acidianus, *Thermoplasma*, and *Sulfolobus* apply completely different complexes in the later steps of their electron transport chains (post-quinone steps). *A. ambivalens* reveals the simplest electron transfer chain consisting of NADH dehydrogenase (NoxA), caldariella quinone, and an aa_3-type quinole oxidase (Dox) (Gomes et al. 2001). This quinole oxidase directly accepts electrons from the quinone (no complex III analogon required), and functions as terminal oxidase to reduce molecular oxygen. Four subunits of the Dox complex have been validated by protein sequencing. Except for *doxB* (*coxI*-like), *dox* genes are not similar to *cox* genes. *Acidianus* might contain another Dox complex since paralogous *doxA* (Q8NKT8, P97224) and *doxD* (Q8NKT7, P97207) genes were found in public databases.

S. acidocaldarius is characterized by so called supercomplexes (Sox), which combine subunits of both, the classical complexes III and IV (Castresana et al. 1995). So far, subunits from three alternative Sox complexes (SoxABCD, SoxM, SoxLN) have been validated by protein sequencing. The three complexes are composed of different complex III and complex IV subunits, amongst them Rieske proteins, cytochrome-*b*-like subunits, and cytochrome-*c*-oxidase-like subunits. On the molecular level, the SoxABCD complex is an aa_3-type terminal oxidase including a cytochrome-a_{587} (cytochrome-*b* homolog). Within the SoxM complex, electrons are presumably transferred from Rieske protein and cytochrome-a_{587} to sulfocyanin, a mobile carrier analog to halocyanin, and then to a ba_3-type terminal oxidase

component (Komorowski and Schafer 2001). Apart from Sox complexes, *dox* genes encoding a probable quinole oxidase were also found within the *S. solfataricus* genome. Thus, *Sulfolobus* species might additionally apply a minimal respiratory chain as observed for *A. ambivalens* under certain conditions.

In the *T. acidophilum* genome, two gene clusters with potential Rieske proteins and cytochromes-*b* subunits have been found. One of the cytochromes-*b* homologs seems to be a fusion protein, to which a domain with similarity to the halobacterial protein (OE1866F, close to the *pet* cluster) is fused to. *T. acidophilum* lacks *cox*, *dox*, and *sox* genes, but a functional respiratory chain might be established using a cytochrome-*d* oxidase complex as terminal oxidase (two *cydA* genes but no *cydB* gene found). Consistent with this, *d*-type cytochrome was detected experimentally in this species (Sreeramulu et al. 1998).

6.4 Variations in the metabolism of haloarchaea

Gene equipments of the four sequenced haloarchaeal strains *(H. salinarum, N. pharaonis, H. walsbyi*, and *H. marismortui)* were compared amongst each other as described in Methods. Retrieved gene lists with genes present in a halophilic strain but absent in at least one of the other haloarchaea were checked, and different metabolic features between the halophiles were summarized in Supplemental Table 6.3.

6.4.1 Sugar and central metabolism

Amongst the four compared halophilic strains, only *H. marismortui* has acquired all genes that are potentially involved in uptake and catabolism of a variety of sugars such as fructose, glucose, maltose, and sucrose. Consistent with biochemical findings for carbohydrate-utilizing haloarchaea (Altekar and Rangaswamy 1992; Danson and Hough 1992), a 1-phosphofructokinase (EC 2.7.1.56) gene needed for fructose degradation via the Embden-Meyerhof (EM) pathway has been identified in the *H. marismortui* genome sequence but a 6-phosphofructokinase (EC 2.7.1.11) homolog for glucose degradation via the EM pathway is lacking. Instead, glucose is degraded via the semi-phosphorylated Entner-Douderoff (ED) pathway involving 2-keto-3-deoxygluconate kinase (EC 2.7.1.45) and aldolase (EC 4.1.2.14). Other sequences haloarchaea do not exhibit diverse sugar metabolism genes found in *H. marismortui*. However, *H. walsbyi* encodes 2-keto-3-deoxygluconate kinase and aldolase as well as components of a sugar phosphotransferase system; thus, glucose degradation is likely in this species. Although *H. salinarum* contains a 2-keto-3-deoxygluconate kinase

gene, this strain seems to have lost its capability to catabolize glucose due to the lack of other enzyme genes for the ED pathway.

All of the analysed halophiles except *N. pharaonis* likely utilize glycerol as carbon source applying glycerol kinase (EC 2.7.1.30) and glycerol-3P dehydrogenase (EC 1.1.99.5) to obtain glycerone-P, which is an intermediate of the lower EM pathway and precursor of archaeal lipids (Figure 6.9). The plasmid-encoded glycerol dehydrogenase (EC 1.1.1.6) gene of *H. salinarum* as well as the dihydroxyacetone kinase (EC 2.7.1.29) gene found in *H. walsbyi* indicate a complex glycerol metabolism in halophiles that should be elucidated in future.

For the lower EM pathway, pyruvate metabolism, and TCA cycle only few differences between the haloarchaea were observed. While all haloarchaea possess glyceraldehyde dehydrogenase (OE1154F, NP0012A, HQ1360A, HQ2025A, rrnAC2262), *H. marismortui* and *H. walsbyi* possess another distantly related type of this enzyme (HQ1394A, rrnAC2363). Furthermore, PEP carboxylase (EC 4.1.1.31) was found in these two strains, thus, trioses resulting from different carbon substrates can be fed into the TCA cycle via three potential pathways. First, phosphoenolpyruvate (PEP) is mainly converted to acetyl-CoA degraded in the TCA cycle for energy generation (Figure 5.8 in Chapter 5.6.1) but PEP

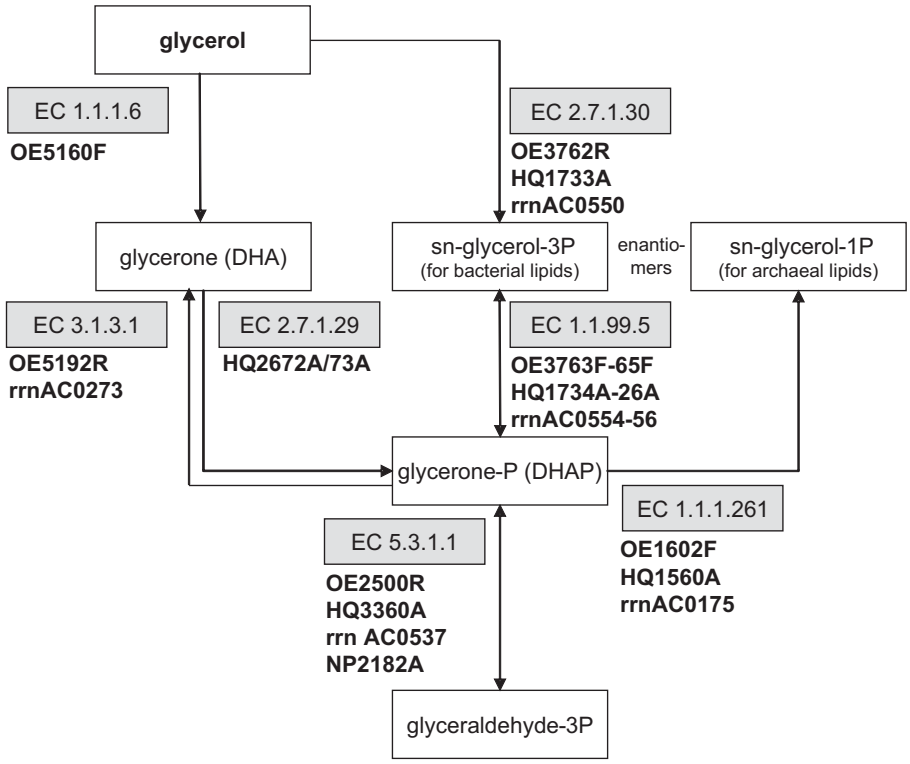


Figure 6.9: **Glycerol metabolism in halophiles.** Glycerol is probably utilized via glycerol kinase (EC 2.7.1.30) and glycerol-3-phosphate dehydrogenase (EC 1.1.99.5). Further enzymes, glycerol dehydrogenase (EC 1.1.1.6) and dihydroxyacetone kinase (EC 2.7.1.29), are present in *H. salinarum* in *H. walsbyi*, respectively. The derived glycerone-P is either fed into the lower Embden-Meyerhof pathway or into archaeal lipid synthesis.

and pyruvate can also be converted to the TCA cycle intermediate oxalacetate. These reactions are needed to fill up the oxalacetate pool of the cycle since TCA cycle compounds are drawn off for biosynthetic purposes.

No differences in respect to TCA cycle genes were found, but isocitrate lyase (EC 4.1.3.1) genes required for a functional glyoxylate cycle are only present in *H. walsbyi* and *N. pharaonis*. Utilization of reduction equivalents produced by TCA cycle enzymes by the respiratory chain has already been discussed in the previous section. Finally, a RUBSICO subunit has been found in *N. pharaonis*, and activity of this enzyme should be determined in future as already done for other halophiles (Rajagopalan and Altekar 1994).

6.4.2 Nucleotide and lipid metabolism

The four haloarchaeal strains show only few differences in their nucleotide and lipid metabolism, namely in the occurrence of pyrimidine kinases (EC 2.7.1.21/48), thymidine phosphorylase (EC 2.4.2.4), and the archaeal-type of IPP isomerase (EC 5.3.3.2) (Supplemental Table 6.3). The latter is involved in the *de novo* synthesis of isoprenoids via the mevalonate pathway, and replaces IPP isomerase found in bacteria. However, in contrast to non-halophilic archaea, all halobacteria encode the bacterial-type IPP isomerase, but *N. pharaonis* and *H. salinarum* possess additionally the archaeal-type of this enzyme. It might be speculated that the bacterial IPP isomerase is more efficient and enable halobacteria to cover increased isoprenoid demands. Isoprenoids are not only required for archaeal membrane lipids and quinone side chains as in other archaea but also for a variety of other membrane components such as dolichol, carotenoid pigments, and retinal that is incorporated in rhodopsins.

For *de novo* synthesis of retinal a β-carotene cleavage enzyme, β,β-carotene 15,15'-monooxygenase (EC 1.14.99.36) is needed. However, a gene encoding this enzyme was only detected on the *H. walsbyi* chromosome and on a *H. marismortui* plasmid. Therefore, *H. salinarum* and *N. pharaonis* must possess a non-orthologous gene encoding an alternative β,β-carotene 15,15'-monooxygenase (EC 1.14.99.36). A probable candidate gene might be *thiC*, which is present in the *H. salinarum*, *N. pharaonis*, and *H. marismortui* genomes but absent in *H. walsbyi* (reverse phylogenetic profile to the canonical enzyme). This gene was also found to be amongst regulated genes in transcriptomics experiments when comparing *H. salinarum* cultures grown under phototrophic conditions (bacterio-rhodopsin is active) with cultures grown in the dark (J. Twellmeyer, pers. comm.).

6.4.3 Amino acid and nitrogen metabolism

Capabilities to synthesize amino acids differ significantly between *H. salinarum* and the other three halophiles (for details see Chapter 6.2). *Halobacterium* lacks genes for the synthesis of branched amino acids, lysine, and ornithine. Further, methionine synthase (EC 2.5.1.49) and hydroxypyruvate reductase (EC 1.1.1.81) are missing, which are involved in methionine and serine synthesis, respectively. The proline synthesis gene cluster *proCBA* was only found in *N. pharaonis* and *H. walsbyi*. However, as discussed in Chapter 6.2, alternative synthesis pathways likely exist in *H. salinarum*, which would explain observed serine and proline synthesis in NMR studies (Ekiel et al. 1986; Ghosh and Sonawat 1998). Some genes required for arginine and lysine synthesis, e.g. amino-acid acetyltransferase (EC 2.3.1.1) and diaminopimelate epimerase (EC 5.1.1.7) are absent in all halophiles, but the occurence of yet unknown non-orthologous genes replacing the missing enzyme genes is likely.

Gene equipment for amino acid degradation pathways differs only slightly between halophiles with tryptophanase (EC 4.1.99.1) and glutamate decarboxylase (EC 4.1.1.15) genes only present in some of the strains. In contrast, nitrogen metabolism varies greatly between the haloarchaea. While *H. marismortui*, *N. pharaonis*, and *H. walsbyi* reveal all required genes for urea conversion and nitrate assimilation, *H. salinarum* lacks these genes. *H. marismortui* encodes a previously described respiratory nitrate reductase (Yoshimatsu et al. 2000) as well as a nitrous-oxide reductase. Urea cycle genes for the conversion of ornithine to arginine were found in all halophilic strains, but arginase (EC 3.5.3.1) genes (rrnAC0383, rrnAC0453) for arginine degradation are only present in *H. marismortui*. Only in *H. salinarum*, plasmid-encoded enzymes of the arginine deiminase pathway are found, which are used for fermentative arginine degradation with concomitant ATP production (Ruepp and Soppa 1996).

6.4.4. Cofactor metabolism

Halophiles also seem to possess different capabilities for cofactor synthesis (Supplemental Table 6.3). While *H. walsbyi* lacks genes for biotin, and *H. salinarum* furthermore several genes for thiamine synthesis, *N. pharaonis* and *H. marismortui* probably synthesize all major cofactors (except pyridoxal-5P). However, two biotin *(bioD)* and thiamine *(thiM)* synthesis genes that are found in *N. pharaonis* remain to be detected in *H. marismortui*. Biosynthesis pathways for biotin and thiamine are not well studied yet, and it might be possible that yet unassigned genes (e.g. *thiD* and *thi1)* found in all sequenced halophiles might replace the known enzymes.

By analysing the four halophilic genomes, many differences in gene equipment for folate biosynthesis as well as folate metabolism were observed (Figure 6.10, Table 6.2). The folate biosynthesis pathway starts with a GTP precursor, which is transformed to a pteridine

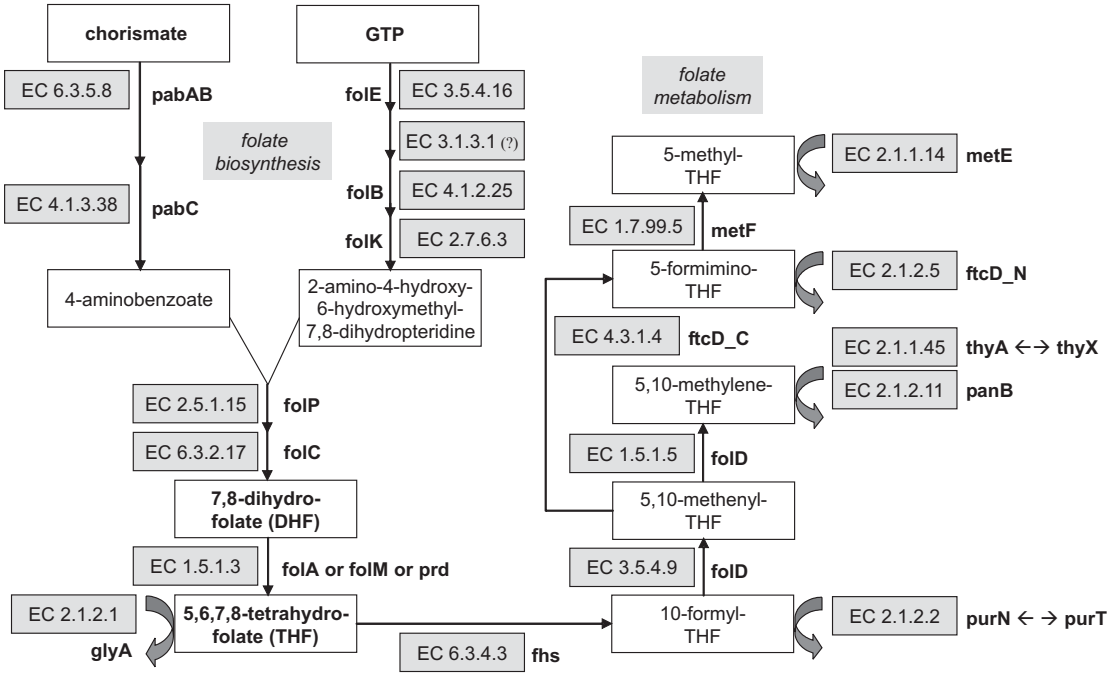


Figure 6.10: **Folate biosynthesis and folate metabolism.** Early steps of folate synthesis starting from GTP are not completely understood yet. DHF is reduced to THF by one of the three alternative dihydrofolate reductases (EC 1.5.1.3). THF is then converted to several other folate coenzymes required for the C1 metabolism. As indicated by grey arrows, folate coenzymes are involved in the biosynthesis of amino acids (EC 2.1.1.14, EC 2.1.2.1), nucleotides (EC 2.1.2.2, EC 2.1.1.45), and coenzyme A (EC 2.1.2.11) as well as in the histidine degradation pathway (EC 2.1.2.5). However, in some instances alternative enzymes (ThyX, PurT) developed, which use alternative coenzymes instead of folate. The gene equipment for folate biosynthesis and metabolism differs considerably between halophilic strains (Table 6.2).

Table 6.2: Folate biosynthesis (A) and metabolism (B) genes in haloarchaea (Havol - *H. volcanii*, Napha - *N. pharaonis*, Hqwal - *H. walsbyi*, Hamar - *H. marismortui*, Hasal - *H. salinarum* str. R1). Gene codes are given in case a gene is present. Clustered genes in the genomes are marked by grey shading. Since the complete *H. volcanii* sequence is not available yet, the given profile might be incomplete.

(A) *FolE, folB*, and *folK* genes were not found in any archaeal genome except for a *folE* gene in the plasmid-encoded folate metabolism cluster of *H. marismortui.* An ISH element disrupts the *pabA* gene of *H. salinarum* str. R1, so that folate biosynthesis is likely to be completely omitted.

(B) Dihydrofolate reductase (EC 1.5.1.3) activity is encoded by the *folA* gene and by the non-orthologous *prd* domain in haloarchaea (Levin et al. 2004). Genomes lacking a *folA* gene usually also lack *thyA* genes encoding the folate-dependent thymidylate synthase (EC 2.1.1.45) because FolA is required for cofactor recycling of ThyA. Species without FolA/ThyA encode the folate-independent enzyme thymidylate synthase ThyX instead, so that ThyA and ThyX exhibit reverse phylogenetic profiles (below the horizontal line).

A: Gene	COG	Napha	Hqwal	Hamar	Hasal
pabA	COG0512	NP0800A	HQ1783A	rrnAC0710 pNG7325	OE1570F/ISH/ OE1573A1F
pabB	COG0147	NP0802A	HQ1784A	rrnAC0709 pNG7327	OE1568F
pabC	COG0115	NP0798A	HQ1247A	rrnAC0711	OE1574F
folE	COG0302	-	-	pNG7382	-
folB	COG1539	-	-	-	-
folK	COG0801	-	-	-	-

B: Gene	COG	Havol	Napha	Hqwal	Hamar	Hasal
folP	COG0294		NP3770A	HQ2465A	rrnAC0246	OE2921R
folC-prd-folP	COG0285/-/ COG0294	AY676165* *(folC-P, but no prd domain)*	NP1478A	HQ1767A	rrnAC0184	OE1615R
folA	COG0262	J05088* *(hdrA)*	-	HQ1842A	(rrnAC0859)	-
		AAF23265† *(hdrB)*	NP2922A	HQ2455A	pNG7359	-
fhs	COG2759		-	HQ1768A	pNG7380	-
folD	COG0190		NP2054A	HQ2790A	rrnAC0996	OE3038F
ftcD_C	COG3404		-	-	pNG7381	-
metF	COG0685		-	HQ1756A	pNG7363	-
thyA	COG0207	AAF23264† *(hts)*	NP2924A	HQ2456A	-	-
thyX	COG1351		-	-	rrnAC1121	OE2898R

* - Codes were retrieved from the nucleotide database of NCBI.
† - Codes were retrieved from the protein database of NCBI.

intermediate by a series of reactions. Genes required for these conversions *(folE, folB, folK)* are missing in all archaeal genomes, so that an alternative pathway seems likely in archaea. However, in *H. marismortui* a *folE* gene encoding GTP cyclohydrolase I (EC 3.5.4.16) was identified within a plasmid-encoded folate metabolism cluster, which was likely acquired by horizontal gene transfer. Interestingly, other cyclohydrolase genes involved in purine *(purH)* and riboflavin synthesis *(ribA)* were also found to be absent in many archaea, and the archaeal *purO* gene has been shown to replace the gene *(purH)* for the bacterial IMP cyclohydrolase (EC 3.5.4.10).

The second folate precursor, 4-aminobenzoate, is synthesized from chorismate, which is an intermediate of the shikimate pathway for *de novo* synthesis of aromatic amino acids (Figure 6.5 in Chapter 6.2.5). For the synthesis of 4-aminobenzoate or para-aminbenzoate (pab), three *pab* genes are required that were found to be clustered in all of the halophilic genomes. *PabAB* encoding the aminodeoxychorismate synthase complex (EC 6.3.5.8) are often misannotated as components of the anthranilate synthase (EC 4.1.3.27), though, which is a similar enzyme involved in aromatic amino acid biosynthesis. The *pabA* homolog of *H. salinarum* str. R1 was found to be disrupted by an ISH element, so that folate biosynthesis in this strain is likely to be completely omitted.

The last step in the biosynthesis of tetrahydrofolate (THF) is catalyzed by dihydrofolate reductase (EC 1.5.1.3). This enzyme is encoded by the *folA* gene, which was found in the chromosomes of *N. pharaonis*, *H. walsbyi* and *Haloferax volcanii* (two paralogs) and on a plasmid in *H. marismortui* but is absent in *H. salinarum*. FolA can be replaced by an analogous enzyme (FolM) or by a linker domain, Prd, which is located in between FolC and FolP domains in FolC-Prd-FolP fusion proteins (Levin et al. 2004). While *folM* is absent in haloarchaeal genomes, *folC-prd-folP* genes are found in all completely sequenced halophiles except *H. volcanii*. For a deletion strain of *H. volcanii*, it was recently shown that the Prd linker domain complements deleted *hdrA* and *hdrB* genes (= *folA* paralogs), and it was

proposed that the dihydrofolate reductase linker domain Prd is involved in an alternative pathway with reduced folate synthesis capability (Levin et al. 2004). In consistence with this, folate is added to the synthetic medium for *H. salinarum* (Oesterhelt and Krippahl 1973), but it is unclear whether *Halobacterium* requires folate due to reduced folate synthesis via Prd or due to the *pabA* mutation found in strain R1.

The dihydrofolate reductase FolA is not only involved in THF synthesis, but also in the regeneration of dihydrofolate derived from a reaction catalyzed by thymidylate synthase (EC 2.1.1.45) encoded by *thyA*. In case of an absent *folA* gene, species were found to replace the folate-dependent thymidylate synthase by an alternative formate-dependent enzyme encoded by *thyX* (Levin et al. 2004). In accordance with this, a *thyA* gene was found adjacent to the *folA* gene in *N. pharaonis* and *H. walsbyi* but is missing in *H. salinarum* and *H. marismortui* that lack *folA* in their chromosomes and contain *thyX* instead of *thyA* (reverse phylogenetic profiles for non-orthologous enzymes ThyA and ThyX).

The *H. marismortui* plasmid pNG700 encodes not only folate biosynthesis genes *folE* (EC 3.5.4.16) and *folCP2* (EC 6.3.2.17/2.5.1.15), but also folate metabolism genes *fhs* (EC 6.3.4.3) and *ftcD*_C (EC 4.3.1.4). The latter genes encode enzymes required for the conversion of THF to other folate coenzymes (Figure 6.10) such as 10-formyl-THF (*fhs*) and 5-formimimo-THF (*ftcD_C*) and occur rarely in other archaea, e.g. *Thermoplasma* species and *H. walsbyi* (only *fhs* found). However, another folate metabolism gene, *folD*, is commonly present in archaea and folate coenzymes are generally required within amino acid nucleotide metabolism pathways (Figure 6.10), so that alternative enzymes are likely to bridge observed gaps in the folate metabolism of archaea.

6.5 Conclusions

Synthesis capabilities for amino acids differ greatly between *N. pharaonis* and *H. salinarum*, and it is not clear whether biosynthesis gene clusters were gained by the *Natronomonas* or lost by *Halobacterium*. However, the second scenario is supported by (i) the occurrence of gene clusters for arginine, lysine, and branched-chain amino acids biosynthesis in the other haloarchaea *(H. walsbyi*, *H. marismortui)* as well as (ii) the presence of some remaining genes (*leuC*, *dapA*) of these pathways in the *H. salinarum* genome. *Halobacterium* strains grow under extreme salt conditions where high amounts of cell material from moderate halophilic organisms are available. These can be sensed by the motile *H. salinarum* cells, processed by its secretion enzymes (halolysin, chitinase, carboxypeptidase) and subsequently utilized. The constant availability of external amino acids might have led to the

loss of gene clusters for amino acid synthesis and to the acquisition of a catabolic gene clusters for arginine and glutamate fermentation. However, also for alkaline lakes, in which *N. pharaonis* thrives in, high levels of C- and N-sources have been reported (Soliman and Truper 1982). In contrast to *Halobacterium*, *N. pharaonis* grows at lower salt concentrations, so that *Natronomonas* likely competes for external compounds with other halophilic species including bacteria and algae. In future, it would be interesting to study if secreted subtilisin-like proteases and amino acid transporters are downregulated upon a shift from the complex to the synthetic medium with fewer nutrients, and if amino acids biosynthesis is subsequently upregulated in *N. pharaonis*.

Some of the differences between the amino acid metabolism in *H. salinarum* and other haloarchaea might be due to differences in the gene equipment for folate metabolism. While *Natronomonas* encodes the dihydrofolate reductase gene *folA, Halobacterium* reveals only a *folC-prd-folP* gene, of which the Prd linker domain was recently proven to act as an alternative dihydrofolate reductase with reduced activity (Levin et al. 2004). *Halobacterium* uses a format-dependent instead of a folate-dependent thymidylate synthase and lacks the folate-dependent methionine synthase. Furthermore, glycine synthesis seems not only to occur via glycine hydroxymethyltransferase (EC 2.1.2.1) in *H. salinarum*, but this folate-dependent enzyme might be circumvented by an alternate pathway via threonine aldolase (EC 4.1.2.5), which is lacking in other haloarchaea. Future studies might elucidate the effects of reduced availability of folate in detail.

Respiratory chains of all aerobic archaea resemble each other in the pre-quinone steps of electron transport (complex I and II), while post-quinone electron transfer steps differ completely in the studied archaeal species. *A. ambivalens* reveals the simplest electron transfer chain by using a quinol oxidase directly as terminal oxidase (no complex III). The *S. solfataricus* genome indicates occurrence of an analogous electron transfer pathway, but so far only other electron transfer modes via the three Sox complexes replacing complex III and IV have been described. The *Sulfolobus* pathway via the SoxM complex is similar to one of the postulated electron transport modes for *H. salinarum* probably transferring electrons from Rieske/cytochrome-*b* components via a mobile cyanin to a ba_3-type terminal oxidase complex. Finally, *T. acidophilum* and *H. salinarum* might apply a cytochrome-*d* oxidase complex as terminal oxidase. For *N. pharaonis*, no complete respiratory chain could be reconstructed, so that even more electron transfer modes are likely to be discovered in archaea.

From phylogenetic trees for subunits I from archaeal and bacterial terminal oxidases (*cox*, *dox*, *sox, qox*), Rieske proteins, and *b*-type cytochromes it was concluded that respiratory chains must have originated prior to the split of the three domains of life (Schafer et al. 1996). However, although respiratory chain components themselves are relatively conserved

in all domains, archaeal respiratory complexes and electron transfer chains reveal a high plasticity. Subunits of respiratory chain complexes for the later electron transfer steps are combined to various complexes (e.g. Sox complexes), and all of the studied respiratory archaea were found to encode multiple alternative complex III and/or complex IV analogs in their genomes. Different types of respiratory complexes within a species and amongst the archaeal domain probably reflect versatile extreme environments, in which conditions can change rapidly. Halobacteria, for example, are exposed to varying oxygen tensions, and cells might cope by flexibly combining alternative respiratory complexes; thus, switching between different modes of electron flow.

The compared gene equipments of the four haloarchaeal genomes fit well to the previously described metabolic pathways of halophiles, e.g. for sugar metabolism in *H. marismortui*. In most cases, a straightforward decision can be reached for or against a metabolic capability in the considered halophile, since the complete set of enzyme genes involved in a metabolic pathway is usually present (proline synthesis gene cluster in *H. walsbyi*) or absent (e.g. glycerol degradation genes in *N. pharaonis*). However, some pathways such as the *N. pharaonis* respiratory chain and lysine and arginine biosynthesis pathways were found to be incomplete. In most instances, it can be assumed, though, that yet unknown non-orthologous enzymes complete these pathways.

Catabolic pathways of haloarchaea differ considerably. While *H. marismortui* and *H. walsbyi* likely utilize sugars, *N. pharaonis* and *H. salinarum* probably produce secretion enzymes to degrade external proteins and peptides. For anabolism, *H. salinarum* assumes a special position as it is the only one of the four sequenced haloarchaea that lacks enzymes for the nitrogen assimilation pathway and several amino acid and coenzyme synthesis pathways. On the other hand, *H. salinarum* acquired many plasmid-encoded metabolic enzymes (e.g. for siderophore biosynthesis, glycerol dehydrogenase) that are unique amongst haloarchaea. *H. marismortui* also reveals many plasmid-encoded enzyme genes (urease, GTP cyclohydrolase I), which might have been acquired by horizontal gene transfer to overcome metabolic deficiencies. It appears that halophiles have adopted different strategies (plasmid-encoded enzymes, secretion enzymes, gain/loss of metabolic pathways) to adapt their metabolism to the nutritional conditions found in halophilic environments which should be analysed in-depth in future comparative studies.

6.6 Methods

6.6.1 Amino acid metabolism in *Natronomonas* and *Halobacterium*

The amino acid metabolism of *H. salinarum* str. R1 was reconstructed as described in Chapter 5.8.3. For the *N. pharaonis* amino acid metabolism, the following automatic reconstruction routine was applied. For each EC-Number of KEGG maps involved in amino acid metabolism, *gene* entries annotated with the given EC-No. were selected from the Halolex database. In case a gene for the given enzyme was found, *org_reactions* (react_exist = 1) and *org_enz_genes* entries were created in Pathnet. The react_exist flag was set to 0 for *org_reactions* entries where no genes were detected. However, some of the automatically generated entries were modified in order to bridge pathway gaps, e.g. in lysine biosynthesis. Furthermore, some comments were created for *org_reactions* and *org_enz_genes* entries. The accuracy of the metabolic reconstruction resulting from the automatic routine strongly depends on the quality of the function annotation of the sequenced genome, especially the EC-No. annotation.

Following reconstruction, a comparison routine was run that compares created *H. salinarum*- and *N. pharaonis*-specific *reactions* entries for a given pathway as defined by a KEGG map. Reactions with differing exist flags for the two halophiles were marked and checked to identify metabolic differences.

6.6.2 Respiratory chains of haloarchaea and other archaea

Published data on archaeal respiratory chains was collected and sequences of described respiratory chain components extracted from Swiss-Prot. The remaining subunits of respiratory chain complexes were identified by blastp search against the complete genomes and their gene context subsequently analysed.

6.6.3 Variations in the metabolism of haloarchaea

Protein databases (fasta) of the four completely sequenced haloarchaeal genomes, *H. salinarum* str. R1*, N. pharaonis, H. walsbyi*, and *H. marismortui*, were blasted against each other and against nr (or Swiss-Prot/TrEMBL) database. For example, each predicted protein sequence of *N. pharaonis* (species 1) was blasted against the *H. salinarum* (species 2) database and nr. Then, the difference of blast scores (-log(E-value)) between the best nr hit and the best *H. salinarum* hit was determined. *N. pharaonis* genes were listed by descending blast score differences in order to find genes lacking in the *H. salinarum* genome. In order to reduce false positive 'missing genes' caused by distant sequence similarity between *N. pharaonis* and *H. salinarum* orthologs only genes with a *H. salinarum* hit below score 20 (E-value $>e^{-20}$) were considered to be absent in *H. salinarum*. The

differential blast analysis routine was applied in both directions (species 1 to species 2 and *vice versa*) for all combinations of the set of four halophilic genomes. The obtained lists with differences in gene equipments were checked for enzyme genes, combined, and summarized in Supplemental Table 6.3.

6.7 Supplemental Material

Supplemental Table 6.3: **Differences in metabolic gene equipment for the four completely sequenced halophilic archaea.** Grey shading marks present enzymes. Enzymes that are replaced by an alternative enzyme with the same function in another haloarchaea are marked by asterisks. Enzymes with reverse phylogenetic profiles, which might replace each other are underlined.

	H. marismortui	*H. walsbyi*	*N. pharaonis*	*H. salinarum*
Central/sugar metabolism and transport				
2-keto-3-deoxygluconate kinase (EC 2.7.1.45)	yes	yes	no	yes
2-keto-3-deoxygluconate aldolase (EC 4.1.2.14)	yes	yes	no	no
1-phosphofructokinase (EC 2.7.1.56)	yes	no	no	no
glycerol dehydrogenase (EC 1.1.1.6)	no	no	no	yes (PL)
dihydroxyacetone kinases (EC 2.7.1.29)	no	yes	no	no
glycerol kinase (EC 2.7.1.30)	yes	yes	no	yes
glycerol-3P dehydrogenase (EC 1.1.99.5)	yes	yes	no	yes
sn-glycerol-3-phosphate ABC transport system	yes	yes	no	yes
GAP dehydrogenase (EC 1.2.1.59)* (also another type in all strains)	yes	yes	no	no
PEP carboxylase (EC 4.1.1.31)	yes	yes	no	no
isocitrate lyase (EC 4.1.3.1), glyoxylate cycle	no	yes	yes (2nd on PL)	no
RUBISCO (EC 4.1.1.39), large subunit	no	no	yes	no
electron transfer flavoprotein	yes	no	yes	yes
pet genes (complex III)	yes	yes	no	yes
cytochrome d oxidase	no	no	no	yes
glucosidases	yes	no	no	no
rhamnulokinase	yes	no	no	no
sucrose-6P hydrolase	yes (PL)	no	no	no
maltose O-acetyltransferase	yes	no	no	no
trehalose-6P synthase	yes	no	no	no
sugar phosphotransferase system, ptsI subunit (EC 2.7.3.9)	yes (PL)	yes	no	no
ptsIIB/C subunits	yes (PL)	no	no	no
maltose ABC transporter	yes	no	no	no
Amino acid and nitrogen metabolism				
methionine synthase (EC 2.5.1.49)	yes	yes	yes	no
hydroxypyruvate reductase (EC 1.1.1.81), Ser biosynthesis	yes	yes	yes	no
glycine cleavage systems (EC 1.4.4.2)	yes	no	yes	yes
threonine aldolase (EC 4.1.2.5), Gly metabolism	no	no	no	yes
proline synthesis enzymes (*proCBA*)	no	yes	yes	no
proline dehydrogenase (EC 1.5.99.8)	yes	no	no	yes
lysine biosynthesis enzymes	yes	yes	yes	no
succinyl-dap desuccinylase (EC 3.5.1.18) *(dapE)*	yes	yes	no	no
diaminopimelate (dap) epimerase (EC 5.1.1.7) *(dapF)*	no	yes	no	no
arginine biosynthesis enzymes	yes	yes	yes	no
arginine deiminase pathway enzymes	no	no	no	yes (PL)
argininases (EC 3.5.3.1)	yes	no	no	no

	H. marismortui	***H. walsbyi***	***N. pharaonis***	***H. salinarum***
branched-chain amino acid biosynthesis enzymes	yes	yes	yes	no
urocanate metabolism (*hut* genes), His degradation	yes	no	no	yes
methylaspartate mutase (EC 5.4.99.1), Glu fermentation methylaspartate ammonia-lyase (EC 4.3.1.2)	yes yes	no no	no no	yes yes
glutamate decarboxylase (EC 4.1.1.15), Glu degradation	yes	no	yes	yes
tryptophanase (EC 4.1.99.1)	yes	no	no	yes
urease and ABC urea transporter nitrate assimilation pathway enzymes and transporter nitrate reductase (respiratory) nitrous-oxide reductase nitrilase nitric-oxide synthase (EC 1.14.13.39), oxygenase subunit N-methylhydantoinase (ATP-hydrolyzing) (EC 3.5.2.14) formamidase (EC 3.5.1.49)	yes (PL) yes yes yes (PL) yes (2nd on PL) no yes yes (PL)	yes yes no no no no yes no	yes yes no no yes yes yes (2nd on PL) yes (PL)	no no no no no no no no
Cofactor, nucleotide, and lipid metabolism				
thiamine-P pyrophosphorylase (EC 2.5.1.3) *(thiE)* hydroxyethylthiazole kinase (EC 2.7.1.50) *(thiM)*	yes no	yes yes	yes yes	no no
biotin biosynthesis enzymes *(bioBF)* dethiobiotin synthase (EC 6.3.3.3) *(bioD)*	yes no	no no	yes yes	no no
GTP cyclohydrolase I (EC 3.5.4.16) dihydrofolate reductase (EC 1.5.1.3)	yes (PL) yes (2nd on PL)	no yes	no yes	no no
thymidylate synthase (EC 2.1.1.45) *(thyA)* thymidylate synthase (EC 2.1.1.45) *(thyX)* uridine kinase (EC 2.7.1.48) thymidine kinase (EC 2.7.1.21) thymidine phosphorylase (EC 2.4.2.4)	no yes no no no	yes no no yes no	yes no no no yes	no yes yes yes no
IPP isomerase (EC 5.3.3.2), archaeal type* (also bacterial-type) β, β-carotene 15,15'-monooxygenase (EC 1.14.99.36) thiamine biosynthesis protein *(thiC)*	no yes (PL) *yes*	no yes no	yes no yes	yes (PL) no yes
Secretion enzymes				
phospholipase C	yes	yes	yes	no
alkaline phosphatase* (3rd type in Hamar) alkaline phosphatase D	no yes	no no	no no	yes (PL) no
chitinase halolysin subtilisin-like proteases carboxypeptidase	no no no no	no no no no	no no yes yes	yes yes no no

Miscellaneous				
cell surface glycoprotein halomucin	yes (2nd on PL) no	yes yes	no no	yes no
MO-ST cluster (*fla/che* genes) transducer proteins	yes yes	no no	yes yes	yes yes
siderophore biosynthesis cluster	no	no	no	yes (PL)
gas vesicle cluster	no	yes	no	yes (PL)
catalase (EC 1.11.1.6)/ peroxidase (EC 1.11.1.7)	yes	no	yes	yes
aerobic carbon monoxide dehydrogenase	no	yes	no	no
aldehyde ferredoxin oxidoreductase (EC 1.2.7.5)	yes	no	yes	no
trp-tRNA ligase (EC 6.1.1.22)*	no	no	no	yes
L-lactate permease	yes	yes	no	yes
Kdp transport system	no	no	no	yes (PL)
arsenite/heavy metal efflux pump transcription regulator *(arsD)*	no no	no yes	yes (PL) no	no yes (PL)

ABBREVIATIONS

BRE	TFB-recognition element
CAnch	C-terminal protein anchor
CLip	C-terminal lipid anchor
COG	cluster of orthologous groups
Csg	cell surface glycoprotein; S-layer protein
Cyt/ExCyt	cytoplasmic/extracytoplasmic
EC-no	enzyme classification number
ED	Entner-Douderoff pathway
EM	Embden-Meyerhof pathway
EV	expert-validated gene and start set
ini-Met	N-terminal methionine residue
ISH	haloarchaeal insertion sequence
KEGG	Kyoto Encyclopaedia of Genes and Genomes
MO-ST	motility and signal transduction
MS	mass spectrometry
N1/N2	unprocessed/processed N-terminal peptide by ini-Met cleavage
NAc	N-acetylation
NLip	N-terminal lipid anchor
NMR	nuclear magnetic resonance spectroscopy
ORF	open reading frame
pI	isoelectric point
PP	pentose-phosphate pathway
PRPP	5-phospho-α-D-ribose 1-diphosphate
PV	proteomics-verified gene and start subset
Sec	general secretory pathway
Tat	twin-arginine pathway
TBP	TATA-box binding protein
TCA cycle	tricarboxylic acid cycle
TFB	transcription factor B
TM	transmembrane domain

Species abbreviations

Arcfu, AF	*Archaeoglobus fulgidus*
Bacsu	*Bacillus subtilus*
Deira	*Deinococcus radiodurans*
Ecoli	*Escherichia coli*
Haln1, HN	*Halobacterium salinarum* strain NRC-1
Hamar, HM	*Haloarcula marismortui*
Hasal, HS, OE	*Halobacterium salinarum* strain R1
Hqwal, HQ	*Haloquadratum walsbyi*
Metac	*Methanosarcina acetivorans*
Metma, MM	*Methanosarcina mazei*
Metka	*Methanopyrus kandleri*
Metth	*Methanothermobacter thermoautotrophicus*
Napha, NP	*Natronomonas pharaonis*
Pyrae	*Pyrobaculum aerophilum*
Pyrfu, PF	*Pyrococcus furiosus*
Strco	*Streptomyces coelicolor*
Sulso, SS	*Sulfolobus solfataricus*
Theac	*Thermoplasma acidophilum*

REFERENCES

Aitken, D.M., and Brown, A.D. (1969) Citrate and glyoxylate cycles in the halophil, Halobacterium salinarium. *Biochimica et Biophysica Acta* **177**: 351-&.

Allers, T., and Mevarech, M. (2005) Archaeal genetics - The third way. *Nature Reviews Genetics* **6**: 58-73.

Altekar, W., and Rangaswamy, V. (1990) Indication of a modified EMP pathway for fructose breakdown in a halophilic archaebacterium. *FEMS Microbiology Letters* **69**: 139-143.

Altekar, W., and Rangaswamy, V. (1992) Degradation of endogenous fructose during catabolism of sucrose and mannitol in halophilic archaebacteria. *Archives of Microbiology* **158**: 356-363.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389-3402.

Avetisyan, A.V., Kaulen, A.D., Skulachev, V.P., and Feniouk, B.A. (1998) Photophosphorylation in alkalophilic halobacterial cells containing halorhodopsin: chloride-ion cycle? *Biochemistry (Moscow)* **63**: 625-628.

Badger, J.H., and Olsen, G.J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Molecular Biology and Evolution* **16**: 512-524.

Baliga, N.S., Bonneau, R., Facciotti, M.T., Pan, M., Glusman, G., Deutsch, E.W. et al. (2004) Genome sequence of Haloarcula marismortui: A halophilic archaeon from the Dead Sea. *Genome Research* **14**: 2221-2234.

Barkley, S.J., Cornish, R.M., and Poulter, C.D. (2004) Identification of an archaeal type II isopentenyl diphosphate isomerase in Methanothermobacter thermautotrophicus. *Journal of Bacteriology* **186**: 1811-1817.

Bell, S.D., and Jackson, S.P. (2001) Mechanism and regulation of transcription in archaea. *Current Opinion in Microbiology* **4**: 208-213.

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology* **340**: 783-795.

Berquist, B.R., and DasSarma, S. (2003) An archaeal chromosomal autonomously replicating sequence element from an extreme halophile, Halobactetium sp strain NRC-1. *Journal of Bacteriology* **185**: 5959-5966.

Bhaumik, S.R., and Sonawat, H.M. (1994) Pyruvate metabolism in Halobacterium salinarium studied by intracellular 13C nuclear magnetic resonance spectroscopy. *Journal of Bacteriology* **176**: 2172-2176.

Bieger, B., Essen, L.O., and Oesterhelt, D. (2003) Crystal structure of halophilic dodecin: a novel, dodecameric flavin binding protein from Halobacterium salinarum. *Structure* **11**: 375-385.

Bolhuis, A. (2002) Protein transport in the halophilic archaeon Halobacterium sp. NRC-1: A major role for the twin-arginine translocation pathway? *Microbiology* **148**: 3335-3346.

Bolhuis, H., Poele, E.M.T., and Rodriguez-Valera, F. (2004) Isolation and cultivation of Walsby's square archaeon. *Environmental Microbiology* **6**: 1287-1291.

Bonete, M.J., Camacho, M.L., and Cadenas, E. (1987) A new glutamate dehydrogenase from Halobacterium halobium with different coenzyme specificity. *International Journal of Biochemistry* **19**: 1149-1155.

Bonete, M.J., Camacho, M.L., and Cadenas, E. (1989) Kinetic mechanism of Halobacterium halobium NAD+-glutamate dehydrogenase. *Biochimica et Biophysica Acta* **990**: 150-155.

Bonete, M.J., Camacho, M.L., and Cadenas, E. (1990) Analysis of the kinetic mechanism of halophilic NADP-dependent glutamate dehydrogenase. *Biochimica et Biophysica Acta* **1041**: 305-310.

Bradshaw, R.A., Brickey, W.W., and Walker, K.W. (1998) N-terminal processing: the methionine aminopeptidase and N-alpha-acetyl transferase families. *Trends in Biochemical Sciences* **23**: 263-267.

Brecht, M., Kellermann, J., and Pluckthun, A. (1993) Cloning and sequencing of glutamate mutase component E from Clostridium tetanomorphum. *FEBS Letters* **319**: 84-89.

Castresana, J., Lubben, M., and Saraste, M. (1995) New archaebacterial genes coding for redox proteins: implications for the evolution of aerobic metabolism. *Journal of Molecular Biology* **250**: 202-210.

Choquet, C.G., Richards, J.C., Patel, G.B., and Sprott, G.D. (1994) Purine and pyrimidine biosynthesis in methanogenic bacteria. *Archives of Microbiology* **161**: 471-480.

Choquet, C.G., Richards, J.C., Patel, G.B., and Sprott, G.D. (1994) Ribose biosynthesis in methanogenic bacteria. *Archives of Microbiology* **161**: 481-488.

Corcelli, A., Lattanzio, V.M.T., Mascolo, G., Papadia, P., and Fanizzi, F. (2002) Lipid-protein stoichiometries in a crystalline biological membrane: NMR quantitative analysis of the lipid extract of the purple membrane. *Journal of Lipid Research* **43**: 132-140.

Danson, M.J., and Hough, D.W. (1992) The enzymology of archaebacterial pathways of central metabolism. *Biochemical Society Symposium*: 7-21.

DasSarma, S., and Arora, P. (2001) Halophiles. In *Encyclopedia of Life Science*. Chichester: John Wiley & Sons, Ltd.

Daugherty, M., Vonstein, V., Overbeek, R., and Osterman, A. (2001) Archaeal shikimate kinase, a new member of the GHMP-kinase family. *Journal of Bacteriology* **183**: 292-300.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999) Improved microbial gene identification with Glimmer. *Nucleic Acids Research* **27**: 4636-4641.

Deppenmeier, U., Johann, A., Hartsch, T., Merkl, R., Schmitz, R.A., Martinez-Arias, R. et al. (2002) The genome of Methanosarcina mazei: Evidence for lateral gene transfer between bacteria and archaea. *Journal of Molecular Microbiology and Biotechnology* **4**: 453-461.

Desmarais, D., Jablonski, P.E., Fedarko, N.S., and Roberts, M.F. (1997) 2-Sulfotrehalose, a novel osmolyte in haloalkaliphilic archaea. *Journal of Bacteriology* **179**: 3146-3153.

Dilks, K., Rose, R.W., Hartmann, E., and Pohlschroder, M. (2003) Prokaryotic utilization of the twin-arginine translocation pathway: a genomic survey. *Journal of Bacteriology* **185**: 1478-1483.

D'Souza, S.E., and Altekar, W. (1998) A class II fructose-1,6-bisphosphate aldolase from a halophilic archaebacterium Haloferax mediterranei. *Journal of General and Applied Microbiology* **44**: 235-241.

Ebert, K., Goebel, W., and Pfeifer, F. (1984) Homologies between heterogeneous extrachromosomal DNA populations of Halobacterium halobium and 4 new halobacterial isolates. *Molecular & General Genetics* **194**: 91-97.

Ekiel, I., Sprott, G.D., and Smith, I.C.P. (1986) Mevalonic acid is partially synthesized from amino acids in Halobacterium cutirubrum: a 13C nuclear magnetic resonance study. *Journal of Bacteriology* **166**: 559-564.

Essen, L.O., Siegert, R., Lehmann, W.D., and Oesterhelt, D. (1998) Lipid patches in membrane protein oligomers: crystal structure of the bacteriorhodopsin-lipid complex. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 11673-11678.

Ghosh, M., and Sonawat, H.M. (1998) Kreb's TCA cycle in Halobacterium salinarum investigated by C-13 nuclear magnetic resonance spectroscopy. *Extremophiles* **2**: 427-433.

Gomes, C.M., Bandeiras, T.M., and Teixeira, M. (2001) A new type-II NADH dehydrogenase from the archaeon Acidianus ambivalens: Characterization and in vitro reconstitution of the respiratory chain. *Journal of Bioenergetics and Biomembranes* **33**: 1-8.

Gordon, D., Abajian, C., and Green, P. (1998) Consed: a graphical tool for sequence finishing. *Genome Research* **8**: 195-202.

Gradin, C.H., Hederstedt, L., and Baltscheffsky, H. (1985) Soluble succinate dehydrogenase from the halophilic archaebacterium, Halobacterium halobium. *Archives of Biochemistry and Biophysics* **239**: 200-205.

Graupner, M., Xu, H.M., and White, R.H. (2002) New class of IMP cyclohydrolases in Methanococcus jannaschii. *Journal of Bacteriology* **184**: 1471-1473.

Green, M.L., and Karp, P.D. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5**.

Green, M.L., and Karp, P.D. (2005) Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Research* **33**: 4035-4039.

Grey, V.L., and Fitt, P.S. (1976) Improved synthetic growth medium for Halobacterium cutirubrum. *Canadian Journal of Microbiology* **22**: 440-442.

Gruber, C., Legat, A., Pfaffenhuemer, M., Radax, C., Weidler, G., Busse, H.J., and Stan-Lotter, H. (2004) Halobacterium noricense sp nov., an archaeal isolate from a bore core of an alpine Permian salt deposit, classification of Halobacterium sp NRC-1 as a strain of H-salinarum and emended description of H-salinarum. *Extremophiles* **8**: 431-439.

Hartmann, R., Sickinger, H.D., and Oesterhelt, D. (1980) Anaerobic growth of halobacteria. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* **77**: 3821-3825.

Hayashi, S., and Wu, H.C. (1990) Lipoproteins in bacteria. *Journal of Bioenergetics and Biomembranes* **22**: 451-471.

Hayden, B.M., Bonete, M.J., Brown, P.E., Moir, A.J.G., and Engel, P.C. (2002) Glutamate dehydrogenase of Halobacterium salinarum: evidence that the gene sequence currently assigned to the NADP(+)-dependent enzyme is in fact that of the NAD(+)-dependent glutamate dehydrogenase. *FEMS Microbiology Letters* **211**: 37-41.

Heath, C., Jeffries, A.C., Hough, D.W., and Danson, M.J. (2004) Discovery of the catalytic function of a putative 2-oxoacid dehydrogenase multienzyme complex in the thermophilic archaeon Thermoplasma acidophilum. *FEBS Letters* **577**: 523-527.

Hirasawa, M., Dose, M.M., Kleis-Sanfrancisco, S., Hurley, J.K., Tollin, G., and Knaff, D.B. (1998) A conserved tryptophan at the ferredoxin-binding site of ferredoxin:nitrite oxidoreductase. *Archives of Biochemistry and Biophysics* **354**: 95-101.

Hirasawa, M., Rubio, L.M., Griffin, J.L., Flores, E., Herrero, A., Li, J. et al. (2004) Complex formation between ferredoxin and Synechococcus ferredoxin: Nitrate oxidoreductase. *Biochimica et Biophysica Acta* **1608**: 155-162.

Hochuli, M., Patzelt, H., Oesterhelt, D., Wuethrich, K., and Szyperski, T. (1999) Amino acid biosynthesis in the halophilic archaeon Haloarcula hispanica. *Journal of Bacteriology* **181**: 3226-3237.

Hofacker, A., Schmitz, K.M., Cichonczyk, A., Sartorius-Neef, S., and Pfeifer, F. (2004) GvpE- and GvpD-mediated transcription regulation of the p-gvp genes encoding gas vesicles in Halobacterium salinarum. *Microbiology-Sgm* **150**: 1829-1838.

Hoff, W.D., Jung, K.H., and Spudich, J.L. (1997) Molecular mechanism of photosignaling by archaeal sensory rhodopsins. *Annual Review of Biophysics and Biomolecular Structure* **26**: 223-258.

Horikoshi, K. (1999) Alkaliphiles: some applications of their products for biotechnology. *Microbiology and Molecular Biology Reviews* **63**: 735-750.

Hou, S.B., Freitas, T., Larsen, R.W., Piatibratov, M., Sivozhelezov, V., Yamamoto, A. et al. (2001) Globin-coupled sensors: A class of heme-containing sensors in archaea and bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 9353-9358.

Huet, J., Schnabel, R., Sentenac, A., and Zillig, W. (1983) Archaebacteria and eukaryotes possess DNA-dependent RNA-polymerases of a common type. *EMBO Journal* **2**: 1291-1294.

Hulo, N., Sigrist, C.J.A., Saux, V.L., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A. et al. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Research* **32**: D134-D137.

Jolley, K.A., Maddocks, D.G., Gyles, S.L., Mullan, Z., Tang, S.L., Dyall-Smith, M.L. et al. (2000) 2-oxoacid dehydrogenase multienzyme complexes in the halophilic archaea? Gene sequences and protein structural predictions. *Microbiology* **146**: 1061-1069.

Julenius, K., Molgaard, A., Gupta, R., and Brunak, S. (2005) Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology* **15**: 153-164.

Kalmokoff, M.L., Koval, S.F., and Jarrell, K.F. (1992) Relatedness of the flagellins from methanogens. *Archives of Microbiology* **157**: 481-487.

Kamekura, M., Seno, Y., Holmes, M.L., and Dyallsmith, M.L. (1992) Molecular cloning and sequencing of the gene for a halophilic alkaline serine protease (halolysin) from an unidentified halophilic archaea strain (172P1) and expression of the gene in Haloferax volcanii. *Journal of Bacteriology* **174**: 736-742.

Kandler, O., and Konig, H. (1998) Cell wall polymers in Archaea (Archaebacteria). *Cellular And Molecular Life Sciences* **54**: 305-308.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Research* **32**: D277-D280.

Kauri, T., Wallace, R., and Kushner, D.J. (1990) Nutrition of the halophilic archaebacterium, Haloferax volcanii. *Systematic and Applied Microbiology* **13**: 14-18.

Kennedy, S.P., Ng, W.V., Salzberg, S.L., Hood, L., and DasSarma, S. (2001) Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Research* **11**: 1641-1650.

Kerscher, L., and Oesterhelt, D. (1981) The catalytic mechanism of 2-oxoacid-ferredoxin oxidoreductases from Halobacterium halobium. One-electron transfer at 2 distinct steps of the catalytic cycle. *European Journal of Biochemistry* **116**: 595-600.

Kerscher, L., and Oesterhelt, D. (1981) Purification and properties of two 2-oxoacid-ferredoxin oxidoreductases from Halobacterium halobium. *European Journal of Biochemistry* **116**: 587-594.

Kerscher, L., and Oesterhelt, D. (1982) Pyruvate-ferredoxin oxidoreductase: new findings on an ancient enzyme. *Trends in Biochemical Sciences* **7**: 371-374.

Kikuchi, A., Sagami, H., and Ogura, K. (1999) Evidence for covalent attachment of diphytanylglyceryl phosphate to the cell-surface glycoprotein of Halobacterium halobium. *Journal of Biological Chemistry* **274**: 18011-18016.

Klare, J.P., Gordeliy, V.I., Labahn, J., Buldt, G., Steinhoff, H.J., and Engelhard, M. (2004) The archaeal sensory rhodopsin II/transducer complex: a model for transmembrane signal transfer. *FEBS Letters* **564**: 219-224.

Klein, C., Garcia-Rizo, C., Bisle, B., Scheffer, B., Zischka, H., Pfeiffer, F. et al. (2005) The membrane proteome of Halobacterium salinarum. *Proteomics* **5**: 180-197.

Kobayashi, T., Kanai, H., Aono, R., Horikoshi, K., and Kudo, T. (1994) Cloning, expression, and nucleotide sequence of the alpha-smylase gene from the haloalkaliphilic archaeon Natronococcus sp. strain Ah-36. *Journal of Bacteriology* **176**: 5131-5134.

Koch, M.K., and Oesterhelt, D. (2005) MpcT is the transducer for membrane potential changes in Halobacterium salinarum. *Molecular Microbiology* **55**: 1681-1694.

Kokoeva, M.V., Storch, K.F., Klein, C., and Oesterhelt, D. (2002) A novel mode of sensory transduction in archaea: binding protein-mediated chemotaxis towards osmoprotectants and amino acids. *EMBO Journal* **21**: 2312-2322.

Komorowski, L., and Schafer, G. (2001) Sulfocyanin and subunit II, two copper proteins with novel features, provide new insight into the archaeal SoxM oxidase supercomplex. *FEBS Letters* **487**: 351-355.

Konrad, Z., and Eichler, J. (2002) Lipid modification of proteins in Archaea: attachment of a mevalonic acid-based lipid moiety to the surface-layer glycoprotein of Haloferax vokanii follows protein translocation. *Biochemical Journal* **366**: 959-964.

Koonin, E.V., Makarova, K.S., and Aravind, L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology* **55**: 709-742.

Krieger, C.J., Zhang, P.F., Mueller, L.A., Wang, A., Paley, S., Arnaud, M. et al. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research* **32**: D438-D442.

Krishnan, G., and Altekar, W. (1993) Halophilic class I aldolase and glyceraldehyde-3-phosphate dehydrogenase: some salt-dependent structural features. *Biochemistry* **32**: 791-798.

Krogh, A., Larsson, B., Heijne, G.v., and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology* **305**: 567-580.

Kushwaha, S.C., Kates, M., and Porter, J.W. (1976) Enzymatic synthesis of C40 carotenes by cell-free preparation from Halobacterium cutirubrum. *Canadian Journal of Biochemistry* **54**: 816-823.

Lanyi, J.K., Duschl, A., Hatfield, G.W., May, K., and Oesterhelt, D. (1990) The primary structure of a halorhodopsin from Natronobacterium pharaonis. Structural, functional and evolutionary implications for bacterial rhodopsins and halorhodopsins. *Journal of Biological Chemistry* **265**: 1253-1260.

Lechner, J., and Sumper, M. (1987) The primary structure of a procaryotic glycoprotein. Cloning and sequencing of the cell surface glycoprotein gene of halobacteria. *Journal of Biological Chemistry* **262**: 9724-9729.

Lechner, J., Wieland, F., and Sumper, M. (1985) Biosynthesis of sulfated saccharides N-glycosidically linked to the protein via glucose. Purification and identification of sulfated dolichyl monophosphoryl tetrasaccharides from halobacteria. *Journal of Biological Chemistry* **260**: 860-866.

Levin, I., Giladi, M., Altman-Price, N., Ortenberg, R., and Mevarech, M. (2004) An alternative pathway for reduced folate biosynthesis in bacteria and halophilic archaea. *Molecular Microbiology* **54**: 1307-1318.

Lledo, B., Martinez-Espinosa, R.M., Marhuenda-Egea, F.C., and Bonete, M.J. (2004) Respiratory nitrate reductase from haloarchaeon Haloferax mediterranei: biochemical and genetic analysis. *Biochimica et Biophysica Acta* **1674**: 50-59.

Lowe, T.M., and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**: 955-964.

Lutz, I., Sieg, A., Wegener, A.A., Engelhard, M., Boche, I., Otsuka, M. et al. (2001) Primary reactions of sensory rhodopsins. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 962-967.

Madigan, M.T., Martinko, J.M., and Parker, J. (2000) *Brock biology of microorganisms*. Upper Saddle River, New Jersey: Prentice -Hall, Inc.

Mahillon, J., and Chandler, M. (1998) Insertion sequences. *Microbiology and Molecular Biology Reviews* **62**: 725-774.

Martinez-Espinosa, R.M., Marhuenda-Egea, F.C., and Bonete, M.J. (2001) Assimilatory nitrate reductase from the haloarchaeon Haloferax mediterranei: Purification and characterisation. *FEMS Microbiology Letters* **204**: 381-385.

Mattar, S. (1996) *Molekularbiologische und biochemische Charakterisierung zweier Komplexe der Atmungskette von Natronomonas pharaonis*. Dortmund.

Mattar, S., and Engelhard, M. (1997) Cytochrome ba(3) from Natronobacterium pharaonis: an archaeal four-subunit cytochrome-c-type oxidase. *European Journal of Biochemistry* **250**: 332-341.

Mattar, S., Scharf, B., Kent, S.B.H., Rodewald, K., Oesterhelt, D., and Engelhard, M. (1994) The primary structure of halocyanin, and archaeal blue copper protein, predicts a lipid anchor for membrane fixation. *Journal of Biological Chemistry* **269**: 14939-14945.

McHardy, A.C., Goesmann, A., Puehler, A., and Meyer, F. (2004) Development of joint application strategies for two microbial gene finders. *Bioinformatics (Oxford)* **20**: 1622-1631.

Mevarech, M., and Werczberger, R. (1985) Genetic transfer in Halobacterium volcanii. *Journal of Bacteriology* **162**: 461-462.

Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J. et al. (2003) GenDB: an open source genome annotation system for prokaryote genomes. *Nucleic Acids Research* **31**: 2187-2195.

Michal, G. (1999) Biochemical pathways. *Spektrum, Akad. Verl., Heidelberg, Berlin*.

Mizuki, T., Kamekura, M., DasSarma, S., Fukushima, T., Usami, R., Yoshida, Y., and Horikoshi, K. (2004) Ureases of extreme halophiles of the genus Haloarcula with a unique structure of gene cluster. *Bioscience Biotechnology and Biochemistry* **68**: 397-406.

Muller, K. (2005) *Aminosaeurestoffwechsel in halophilen Archaea*. Munich.

Muller, J.A., and DasSarma, S. (2005) Genomic analysis of anaerobic respiration in the archaeon Halobacterium species NRC-1: dimethyl sulfoxide and trimethylamine N-oxide as terminal electron acceptors. *Journal of Bacteriology* **187**: 1659-1667.

Nakamura, S., Mizutani, S., Wakai, H., Kawasaki, H., Aono, R., and Horikoshi, K. (1995) Purification and partial characterization of cell surface gycoprotein from extremely halophilic archaeon Haloarcula japonica strain TR-1. *Biotechnology Letters* **17**: 705-706.

Ng, W.V., Kennedy, S.P., Mahairas, G.G., Berquist, B., Pan, M., Shukla, H.D. et al. (2000) Genome sequence of Halobacterium species NRC-1. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 12176-12181.

Nielsen, H., Engelbrecht, J., Brunak, S., and Heijne, G.v. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* **10**: 1-6.

Nordmann, B., Lebert, M.R., Alam, M., Nitz, S., Kollmannsberger, H., Oesterhelt, D., and Hazelbauer, G.L. (1994) Identification of volatile forms of methyl groups released by Halobacterium salinarium. *Journal of Biological Chemistry* **269**: 16449-16454.

Nutsch, T., Marwan, W., Oesterhelt, D., and Gilles, E.D. (2003) Signal processing and flagellar motor switching during phototaxis of Halobacterium salinarum. *Genome Research* **13**: 2406-2412.

Oesterhelt, D. (1976) Isoprenoids and bacteriorhodopsin in halobacteria. *Progress in Molecular and Subcellular Biology* **4**: 133-166.

Oesterhelt, D., and Krippahl, G. (1973) Light inhibition of respiration in Halobacterium halobium. *FEBS Letters* **36**: 72-76.

Oesterhelt, D., and Krippahl, G. (1983) Phototropic growth of halobacteria and its use for isolation of photosynthetically-deficient mutants. *Annales De Microbiologie* **B134**: 137-150.

Oesterhelt, D., and Tittor, J. (1989) Two Pumps, one principle: light-driven ion transport in halobacteria. *Trends in Biochemical Sciences* **14**: 57-61.

Oren, A. (2002) *Halophilic microorganisms and their environments*. Dordrecht: Kluwer Academic Publishers.

Pallen, M.J., Chaudhuri, R.R., and Henderson, I.R. (2003) Genomic analysis of secretion systems. *Current Opinion in Microbiology* **6**: 519-527.

Patenge, N., Berendes, A., Engelhardt, H., Schuster, S.C., and Oesterhelt, D. (2001) The fla gene cluster is involved in the biogenesis of flagella in Halobacterium salinarum. *Molecular Microbiology* **41**: 653-663.

Peck, R.F., Johnson, E.A., and Krebs, M.P. (2002) Identification of a lycopene beta-cyclase required for bacteriorhodopsin biogenesis in the archaeon Halobactetium salinarum. *Journal of Bacteriology* **184**: 2889-2897.

Peck, R.F., Echavarri-Erasun, C., Johnson, E.A., Ng, W.V., Kennedy, S.P., Hood, L. et al. (2001) Brp and blh are required for synthesis of the retinal cofactor of bacteriorhodopsin in Halobacterium salinarum. *Journal of Biological Chemistry* **276**: 5739-5744.

Perez-Pomares, F., Ferrer, J., Camacho, M., Pire, C., Llorca, F., and Bonete, M.J. (1999) Amino acid residues involved in the catalytic mechanism of NAD-dependent glutamate dehydrogenase from Halobacterium salinarum. *Biochimica et Biophysica Acta-General Subjects* **1427**: 417-417.

Perrier, J., Durand, A., Giardina, T., and Puigserver, A. (2005) Catabolism of intracellular N-terminal acetylated proteins: involvement of acylpeptide hydrolase and acylase. *Biochimie* **in press**.

Pfeiffer, T., Sanchez-Valdenebro, I., Nuno, J.C., Montero, F., and Schuster, S. (1999) METATOOL: for studying metabolic networks. *Bioinformatics* **15**: 251-257.

Plugge, C.M., van Leeuwen, J.M., Hummelen, T., Balk, M., and Stams, A.J.M. (2001) Elucidation of the pathways of catabolic glutamate conversion in three thermophilic anaerobic bacteria. *Archives of Microbiology* **176**: 29-36.

Pohlschroder, M., Dilks, K., Hand, N.J., and Rose, R.W. (2004) Translocation of proteins across archaeal cytoplasmic membranes. *FEMS Microbiology Reviews* **28**: 3-24.

Polevoda, B., and Sherman, F. (2003) N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins. *Journal of Molecular Biology* **325**: 595-622.

Porat, I., Waters, B.W., Teng, Q., and Whitman, W.B. (2004) Two biosynthetic pathways for aromatic amino acids in the archaeon Methanococcus maripaludis. *Journal of Bacteriology* **186**: 4940-4950.

Preston, G.M., Studholme, D.J., and Caldelari, I. (2005) Profiling the secretomes of plant pathogenic Proteobacteria. *FEMS Microbiology Reviews* **29**: 331-360.

Purschke, W.G., Schmidt, C.L., Petersen, A., and Schafer, G. (1997) The terminal quinol oxidase of the hyperthermophilic archaeon Acidianus ambivalens exhibits a novel subunit structure and gene organization. *Journal of Bacteriology* **179**: 1344-1353.

Radax, C., Gruber, C., and Stan-Lotter, H. (2001) Novel haloarchaeal 16S rRNA gene sequences from Alpine Permo-Triassic rock salt. *Extremophiles* **5**: 221-228.

Rajagopalan, R., and Altekar, W. (1994) Characterization and purification of ribulose-bisphosphate carboxylase from heterotrophically grown halophilic archaebacterium, Haloferax mediterranei. *European Journal of Biochemistry* **221**: 863-869.

Rawal, N., Kelkar, S.M., and Altekar, W. (1988) Alternative routes of carbohydrate metabolism in halophilic archaebacteria. *Indian Journal of Biochemistry & Biophysics* **25**: 674-686.

Rose, R.W., Bruser, T., Kissinger, J.C., and Pohlschroder, M. (2002) Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway. *Molecular Microbiology* **45**: 943-950.

Rosenshine, I., Tchelet, R., and Mevarech, M. (1989) The mechanism of DNA transfer in the mating system of an archaebacterium. *Science* **245**: 1387-1389.

Roth, J.R., Lawrence, J.G., and Bobik, T.A. (1996) Cobalamin (coenzyme B-12): Synthesis and biological significance. *Annual Review of Microbiology* **50**: 137-181.

Rudolph, J., and Oesterhelt, D. (1996) Deletion analysis of the che operon in the archaeon Halobacterium salinarium. *Journal of Molecular Biology* **258**: 548-554.

Rudolph, J., Tolliday, N., Schmitt, C., Schuster, S.C., and Oesterhelt, D. (1995) Phosphorylation in halobacterial signal transduction. *EMBO Journal* **14**: 4249-4257.

Ruepp, A., and Soppa, J. (1996) Fermentative arginine degradation in Halobacterium salinarium (formerly Halobacterium halobium): Genes, gene products, and transcripts of the arcRACB gene cluster. *Journal of Bacteriology* **178**: 4942-4947.

Sartorius-Neef, S., and Pfeifer, F. (2004) In vivo studies on putative Shine-Dalgarno sequences of the halophilic archaeon Halobacterium salinarum. *Molecular Microbiology* **51**: 579-588.

Schafer, G. (1996) Bioenergetics of the archaebacterium Sulfolobus. *Biochimica et Biophysica Acta* **1277**: 163-200.

Schafer, G., Purschke, W.G., Gleissner, M., and Schmidt, C.L. (1996) Respiratory chains of archaea and extremophiles. *Biochimica et Biophysica Acta - Bioenergetics* **1275**: 16-20.

Scharf, B., Wittenberg, R., and Engelhard, M. (1997) Electron transfer proteins from the haloalkaliphilic archaeon Natronobacterium pharaonis: possible components of the respiratory chain include cytochrome bc and a terminal oxidase cytochrome ba(3). *Biochemistry* **36**: 4471-4479.

Seidel, R., Scharf, B., Gautel, M., Kleine, K., Oesterhelt, D., and Engelhard, M. (1995) The primary structure of sensory rhodopsin II: a member of an additional retinal protein subgroup is coexpressed with its transducer, the halobacterial transducer of rhodopsin II. *Proceedings of the National Academy of Sciences of the United States of America* **92**: 3036-3040.

Skulachev, V.P. (1992) The laws of cell energetics. *European Journal of Biochemistry* **208**: 203-209.

Skulachev, V.P., Kobayashi, H., Krulwich, T.A., Schafer, G., Fillingame, R.H., Poole, R.K. et al. (1999) Bacterial energetics at high pH: what happens to the H+ cycle when the extracellular H+ concentration decreases? *Bacterial Response to pH - Novartis Foundation Symposium* **221**: 200-217.

Smit, A., and Mushegian, A. (2000) Biosynthesis of isoprenoids via mevalonate in archaea: the lost pathway. *Genome Research* **10**: 1468-1484.

Soliman, G.S.H., and Truper, H.G. (1982) Halobacterium pharaonis sp. nov., a new, extremely haloalkaliphilic archaebacterium with low magnesium requirement. *Zentralblatt Fur Bakteriologie Mikrobiologie Und Hygiene I Abteilung Originale C* **3**: 318-329.

Sonawat, H.M., Srivastava, S., Swaminathan, S., and Govil, G. (1990) Glycolysis and Entner-Doudoroff pathways in Halobacterium halobium: some new observations based on 13C NMR spectroscopy. *Biochemical and Biophysical Research Communications* **173**: 358-362.

Soppa, J., and Oesterhelt, D. (1989) Halobacterium sp. GRB: a species to work with!? *Canadian Journal of Microbiology* **35**: 205-209.

Sperling, D., Kappler, U., Wynen, A., Dahl, C., and Truper, H.G. (1998) Dissimilatory ATP sulfurylase from the hyperthermophilic sulfate reducer Archaeoglobus fulgidus belongs to the group of homo-oligomeric ATP sulfurylases. *FEMS Microbiology Letters* **162**: 257-264.

Sreeramulu, K., Schmidt, C.L., Schafer, G., and Anemuller, S. (1998) Studies of the electron transport chain of the euryarcheon Halobacterium salinarum: Indications for a type II NADH dehydrogenase and a complex III analog. *Journal of Bioenergetics and Biomembranes* **30**: 443-453.

Stan-Lotter, H., Doppler, E., Jarosch, M., Radax, C., Gruber, C., and Inatomi, K.-i. (1999) Isolation of a chymotrypsinogen B-like enzyme from the archaeon Natronomonas pharaonis and other halobacteria. *Extremophiles* **3**: 153-161.

Stoeckenius, W., Lozier, R.H., and Bogomolni, R.A. (1979) Bacteriorhodopsin and the purple membrane of halobacteria. *Biochimica et Biophysica Acta* **505**: 215-278.

Sumper, M. (1987) Halobacterial glycoprotein biosynthesis. *Biochimica et Biophysica Acta* **906**: 69-80.

Sumper, M. (1993) S-layer glycoproteins from moderately and extremely halophilic archaeobacteria. In *Advances in bacterial paracrystalline surface layers*. Beveridge, T.J., and Koval, S.F. (eds). New York: Plenum Press.

Tansill, B. (1984) *Bergey's manual of systematic bacteriology*. Baltimore: Williams & Wilkins.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997) A genomic perspective on protein families. *Science* **278**: 631-637.

Tebbe, A., Klein, C., Bisle, B., Siedler, F., Scheffer, B., Garcia-Rizo, C. et al. (2005) Analysis of the cytosolic proteome of Halobacterium salinarum and its implication for genome annotation. *Proteomics* **5**: 168-179.

Thomas, N.A., Bardy, S.L., and Jarrell, K.F. (2001) The archaeal flagellum: a different kind of prokaryotic motility structure. *FEMS Microbiology Reviews* **25**: 147-174.

Thompson, D.K., Palmer, J.R., and Daniels, C.J. (1999) Expression and heat-responsive regulation of a TFIIB homologue from the archaeon Haloferax volcanii. *Molecular Microbiology* **33**: 1081-1092.

Tindall, B.J., Ross, H.N.M., and Grant, W.D. (1984) Natronobacterium gen. nov. and Natronococcus gen. nov., two new genera of haloalkaliphilic archaebacteria. *Systematic and Applied Microbiology* **5**: 41-57.

Tjalsma, H., Bolhuis, A., Jongbloed, J.D.H., Bron, S., and van Dijl, J.M. (2000) Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. *Microbiology and Molecular Biology Reviews* **64**: 515-+.

Torarinsson, E., Klenk, H.P., and Garrett, R.A. (2005) Divergent transcriptional and translational signals in Archaea. *Environmental Microbiology* **7**: 47-54.

Tumbula, D.L., Teng, Q., Bartlett, M.G., and Whitman, W.B. (1997) Ribose biosynthesis and evidence for an alternative first step in the common aromatic amino acid pathway in Methanococcus maripaludis. *Journal of Bacteriology* **179**: 6010-6013.

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* **31**: 258-261.

Vreeland, R.H., Rosenzweig, W.D., and Powers, D.W. (2000) Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature* **407**: 897-900.

Walsby, A.E. (1980) Square bacterium. *Nature* **283**: 69-71.

Weik, M., Patzelt, H., Zaccai, G., and Oesterhelt, D. (1998) Localization of glycolipids in membranes by in vivo labeling and neutron diffraction. *Molecular Cell* **1**: 411-419.

Woese, C.R., and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* **74**: 5088-5090.

Woodson, J.D., and Escalante-Semerena, J.C. (2004) CbiZ, an amidohydrolase enzyme required for salvaging the coenzyme B-12 precursor cobinamide in archaea. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 3591-3596.

Wulff, K., Hubbard, J.S., and Miller, A.B. (1972) Reversible inactivation of isocitrate dehydrogenase from an obligate halophile: changes in the secondary structure. *Archives of Biochemistry and Biophysics* **148**: 318-&.

Yamashita, S., Hemmi, H., Ikeda, Y., Nakayama, T., and Nishino, T. (2004) Type 2 isopentenyl diphosphate isomerase from a thermoacidophilic archaeon Sulfolobus shibatae. *European Journal of Biochemistry* **271**: 1087-1093.

Yao, V.J., and Spudich, J.L. (1992) Primary structure of an archaebacterial transducer, a methyl-accepting protein associated with sensory rhodopsin I. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 11915-11919.

Yoshimatsu, K., Sakurai, T., and Fujiwara, T. (2000) Purification and characterization of dissimilatory nitrate reductase from a denitrifying halophilic archaeon, Haloarcula marismortui. *FEBS Letters* **470**: 216-220.

Zhang, W.S., Brooun, A., Mueller, M.M., and Alam, M. (1996) The primary structures of the Archaeon Halobacterium salinarium blue light receptor sensory rhodopsin II and its transducer, a methyl-accepting protein. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 8230-8235.

Weblinks

BRENDA	http://www.brenda.uni-koeln.de; Comprehensive Enzyme Information System
COG	http://www.ncbi.nlm.nih.gov/COG/new/; Cluster of Orthologous Groups DB
ENZYME	http://www.expasy.org/enzyme/; Enzyme Nomenclature Database
Halolex	http://www.halolex.mpg.de; Information System for Halophilic Archaea
JVirGel	http://www.jvirgel.de; Calculation of virtual two-dimensional protein gels
KEGG	http://www.genome.jp/kegg/pathway.html#metabolism; KEGG PATHWAY DB
KEGG API	http://www.genome.jp/kegg/soap/
MetaCyc/EcoCyc	http://metacyc.org and http://ecocyc.org; DB of Nonredundant, Experimentally Elucidated Metabolic Pathways
Migenas	http://www.migenas.org; Microbial Genome Analysis System of the Max-Planck Society
MolE	http://medlib.med.utah.edu/masspec/mole.htm; Molecular Mass Calculator
NCBI Taxonomy	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy
NetNGlyc/NetOGlyc	http://www.cbs.dtu.dk/services; Center for Biological Sequence Analysis Prediction Server
PDB	http://www.rcsb.org/pdb/; RCSB Protein Data Bank
Pfam	http://www.sanger.ac.uk/Software/Pfam/; Protein Families Database of Alignments and HMMs
PROSITE	http://www.expasy.org/prosite/; Database of Protein Families and Domains
Smart	http://smart.embl-heidelberg.de/; Simple Modular Architecture Research Tool
Swiss-Prot	http://www.expasy.org/sprot/; Protein Knowledgebase

Publications

The work presented in this thesis will contribute to the following publications:

Michaela Falb, Friedhelm Pfeiffer, Peter Palm, Karin Rodewald, Volker Hickmann, Jörg Tittor, Dieter Oesterhelt

Living with two extremes: Conclusions from the genome sequence of *Natronomonas pharaonis*. Genome Research. *In press.*

Michaela Falb, Friedhelm Pfeiffer, Dieter Oesterhelt

Genome-wide analysis of gene start codons and N-terminal peptide modifications in halophilic archaea. *In preparation.*

Friedhelm Pfeiffer, Stefan Schuster, Peter Palm, Karin Rodewald, Michaela Falb, Jörg Tittor, Jörg Soppa, Andreas Ruepp, Dieter Oesterhelt

The genome squence of *Halobacterium salinarum* strain R1. *In preparation.*

Henk Bolhuis, Peter Palm, Francisco Rodriguez-Valera, Michaela Falb, Friedhelm Pfeiffer, Dieter Oesterhelt

The genome of square haloarchaeon: Survival at the limits of water activity. *In preparation.*

Poster presentation:

Michaela Falb, Friedhelm Pfeiffer, Dieter Oesterhelt

Gene and start codon selection in GC-rich genomes. 12th International Conference on Intelligent Systems for Molecular Biology (ISMB)/3rd European Conference on Computational Biology (ECCB). Glasgow. July 31 - August 4, 2004.

ACKNOWLEDGEMENTS

I am especially grateful to Prof. Dr. Dieter Oesterhelt who was a great source of inspiration for me. I would like to thank him for accompanying this PhD project with his scientific interest and guidance, for his generous support as well as for helpful discussions and encouragement.

I am indepted to Prof. Dr. Erich Bornberg-Bauer for introducing me to the project and for offering me the opportunity to broaden my knowledge in the Bioinformatics Lab at the University of Manchester. I also thank him for numerous constructive discussions, his continuous support as well as for helpful tips about planning and conducting a PhD study.

I especially want to thank my supervisor Dr. Friedhelm Pfeiffer for his steady help relating technical and biological questions of my work. I am very grateful for his continuous support and guidance, in particular for his help with 'tedious' gene start assignments, his contribution to the integration of proteomics data into my work, and his patience when reviewing manuscripts of mine.

I very much appreciated the work of Kerstin Müller, Anke Klein, and Björn Hammesfahr who implemented various web tools to integrate some of my data into Halolex.

I further would like to acknowledge Markus Rampp, Reinhard Tisma, and Thomas Soddemann from the computing centre at Garching as well as Günter Raddatz for their frequent help with technical problems and immediate support in case of 'emergency'.

I am obliged to Jörg Tittor and Susanne von Gronau who performed experiments accompanying the genome analysis of *N. pharaonis*. I would also like to appreciate the lab assistence by Bea Scheffer and Brigitte Kessler as well as the critical review of the *N. pharaonis* manuscript by Dr. Doug Griffith.

I am also grateful to my former and present colleagues for their help, friendship, and the nice working atmosphere, especially Jens Twellmeyer and Gregorius Amoutzias for interesting discussions, encouragement, and some distraction with tea & tanks.

Ganz besonders möchte ich meinen Eltern und Schwestern danken, die mich in den vergangenen Jahren jederzeit und auf vielfältige Weise unterstützt und mir stets Mut gemacht haben, dass ich diese Doktorarbeit erfolgreich beenden werde.

Michaela Falb

Personal Details

Address Agnes-Bernauer-Str. 244, 81241 Munich
Nationality German
Date of Birth 4 March 1978
Place of Birth Heiligenstadt/Eichsfeld

Education

Oct 2001 – Aug 2005
PhD in Bioinformatics
Max-Planck-Institute of Biochemistry, Martinsried near Munich
Thesis **Computational genome and pathway analysis of halophilic archaea**
Supervised by Prof D Oesterhelt

Oct 1996 - Sep 2001
Diploma in Biotechnology
Technical University of Braunschweig
Thesis **Development of a database for prokaryotic transcription factors**
Supervised by Prof D Jahn
Exam Subjects Technical Biochemistry, Technical Microbiology, Genetics, Chemical Engineering, and Technical Chemistry

Sep 1992 – Jun 1996
Abitur
Lingemann Gymnasium, Heiligenstadt/Eichsfeld
Exam Subjects Mathematics, Biology, English, Economics & Law

Research projects in the UK

Sep 2002 – Jun 2003
Marie Curie Training Site in Bioinformatics, University of Manchester
Supervised by Prof E Bornberg-Bauer
Thesis **Computational pathway analysis of *Halobacterium salinarum***

Jun 2000 - Sep 2000
Summer Research Project, University of Manchester
Supervised by Dr MN Jones
Thesis **The targeting of phospholipid vesicles to biofilms of *Streptococcus oralis* and their use as carriers of the bactericide chlorhexidine**

Skills

Computing
o SQL programming, setup and maintenance of relational databases
o highly proficient in Perl, basic knowledge in Java and Pascal
o profound experience with the GenDB genome annotation software
o familiar with Unix/Linux, Windows, and MS office

Languages German (native speaker), English (fluent, written and spoken), Russian (intermediate level), Spanish & Arabic (basics)