J Mol Evol (2001) 52:471–489
DOI: 10.1007/s002390010178

JOURNAL OF MOLECULAR EVOLUTION

Phylogenetic Analysis of Metabolic Pathways

Christian V. Forst,* Klaus Schulten

Theoretical Biophysics Group, University of Illinois at Urbana–Champaign, Beckman Institute, MC-251, 405 North Mathews Avenue, Urbana, IL 61801, USA

Received: 14 August 2000 / Accepted: 4 January 2001

Abstract. The information provided by completely sequenced genomes can yield insights into the multi-level organization of organisms and their evolution. At the lowest level of molecular organization individual enzymes are formed, often through assembly of multiple polypeptides. At a higher level, sets of enzymes group into metabolic networks. Much has been learned about the relationship of species from phylogenetic trees comparing individual enzymes. In this article we extend conventional phylogenetic analysis of individual enzymes in different organisms to the organisms' metabolic networks. For this purpose we suggest a method that combines sequence information with information about the underlying reaction networks. A distance between pathways is defined as incorporating distances between substrates and distances between corresponding enzymes. The new analysis is applied to electron-transfer and amino acid biosynthesis networks yielding a more comprehensive understanding of similarities and differences between organisms.

Key words: Metabolic networks — Phylogeny — Electron transfer — Aminoacid biosynthesis — Microbial genomes

Introduction

The metabolism of living systems and the evolution of metabolism have been investigated for several decades. The first studies were performed in the late 50s and early 60s by Popper (1957, 1963) and Lipmann (1965). These studies were followed by others seeking to understand the origin of life and the evolution of the biosphere: seminal contributions by Haldane (1928), Miller (1953), Oparin (1924), and Orgel (1968) discussing the (prebiotic) chemical environment suitable for a biotic evolution are noteworthy in this context. Based on these discussions, hypotheses on the origin and evolution of metabolism were formulated (Hartman 1975), and questions regarding the emergence of the first cyclic metabolic networks[1] were addressed, e.g., of the citric acid cycle (Wächtershäuser 1990).

The data available from complete genomes permits an analysis of higher-level functional components, such as metabolic networks, as has been demonstrated by Overbeek et al. (1997). There is a need, however, for methods to compare higher-level functional components between, and within, organisms. The availability of complete genomes from phylogenetically diverse representatives of all three known domains (archaea, bacteria, and eukaryotes) will create new possibilities for analysis methods.

In this article we extend the conventional sequence comparison and phylogenetic analysis of individual enzymes to metabolic networks. First, a database suitable for this task is outlined. Second, a method for calculating distances between metabolic networks based on sequence information of the involved biomolecules is presented. The suggested method is similar to methods used in existing packages for performing sequence alignment

* *Present Address:* Bioscience Division, Mailstop M888, Los Alamos National Laboratory, Los Alamos, NM 87545
Correspondence to: C.V. Forst; e-mail: chris@lanl.gov

[1] For a definition see Appendix

and for analyzing phylogenies (Thompson et al. 1994, Felsenstein 1996, Huson 1998). Employing our method, a phylogenetic analysis of the reversible ferredoxin–NADPH reductase pathway[2] is performed. The analysis is extended to seven other pathways involving ferredoxin. In a second example, terminal oxidase complexes are analyzed and the results are compared to conclusions reached earlier by Musser and Chan (1998). The third example is an investigation of the tryptophan biosynthesis pathway, which connects with the serine salvage pathway and the pathway of serine biosynthesis.

Materials and Methods

The analysis of metabolic networks based on the sequence information of enzymes and substrates requires access to suitable databases. Recently such databases, which provide the combined information of sequences and pathways, have been established. One example of metabolic databases is the WIT-system (What Is There) (Overbeek et al. 2000). The WIT-system provides information on gene and operon organization, as well as information about metabolic networks for completely or partially sequenced genomes. Using WIT, researchers are able to perform a so-called metabolic reconstruction of microbial genomes (Overbeek et al. 1997). Independently, Tatusov et al. (1996) deduced the metabolism and evolution of *H. influenzae* from a whole-genome comparison with *E. coli.*

The WIT-database allows searches of unannotated proteins in each of the sequenced genomes. The strength of the WIT-system lies in the interactive annotation of unannotated proteins and assertion of pathways. This system has been employed in our study to obtain sequence information as well as information on the organization of metabolic networks. All annotations for functional genes within the studied pathways included in the WIT-system were compared to annotated sequences in the common sequence databases and were reannotated if necessary.

Currently, 53 genomes of microbial origin and one of a multicellular organism (*C. elegans*) are accessible via the WIT-system. Of these genomes, 42 are today completely sequenced, the remaining genomes are subject of ongoing sequencing projects. All the genomes of organisms that are used throughout the paper are characterized in Table 1.

Both the WIT-system and this paper use a classification of organisms into three primary domains in agreement with Woese (1982, 1998a). We note that Mayr (1998) strongly disagrees with this classification and suggests a return to a classification into prokaryotes and eukaryotes as introduced by Chatton (1937).

Distances Between Metabolic Pathways

Aligning sequences to each other and measuring distances, using, for example, BLOSUM (Henikoff and Henikoff 1992) and PAM (Dayhoff et al. 1978) similarity matrices in multiple sequence alignment algorithms, is a common approach to compare individual enzymes. Either by direct usage of molecular sequence data with parsimony or maximum likelihood methods, or a two step approach via (i) multiple sequence alignment and calculation of a corresponding distance matrix, and (ii) visualization of the distance data as graphs, a phylogenetic graph is constructed. In this paper these methods are extended to define distances between metabolic pathways. For this purpose, we combine sequence information of involved genes with information of the corresponding network. Metabolic pathways are considered as reaction graphs (networks) with specific graph-topological information, such as connectivity. For each functional role of the pathway, all genes in the genomes that code for this functional role are used. The sequences corresponding to the functional roles are combined into a set of sequences. For each set of sequences a multiple sequence alignment is performed using ClustalW v1.74 (Thompson et al. 1994) together with the BLOSUM62 similarity matrix. Alignment parameters are set to default values. Comparisons of pathways with different topologies are performed by introducing gap penalties[3] for each missing functional role and by considering adjacency matrices to address the graph topology. ClustalW provides distance matrices for aligned sequences which are then used for calculations of the pathway distances. Phylogenetic relationships are deduced through phylogenetic graph reconstruction programs such as *SPLITSTREE2* (Huson 1998) or the *PHYLIP* software suite (Felsenstein 1996) employing the Fitch–Margoliash method (Fitch and Margoliash 1967).

The simplest type of metabolic pathway involves a substrate processed by an enzyme. The distance Δ between such pathways is defined through distances between sequences of the same functional role, i.e., substrate S and enzyme E. The corresponding distances, obtained by ClustalW multiple sequence alignments, are denoted by ΔS and ΔE. Cofactors, such as ferredoxins, are referred to in this contribution as substrates. The active agents of ferredoxins are Fe-S clusters that are oxidized or reduced by oxidoreductases; the scaffold that keeps Fe-S clusters in place is provided by the tertiary structure of the ferredoxins. In contrast to ferredoxins and other coenzymes coded in the genome, small organic molecules that are involved in metabolic pathways are not considered as functional roles. These substrates are present in all organisms and, thus, a respective distance of NADH in *E. coli* and NADH in *M. jannaschii,* for instance, vanishes.

In the following we define an overall distance which encompasses distances between enzymes as well as distances between substrates when the latter are proteins. Proteins arise in the form of substrates in the ferredoxin–NADPH reductase pathway, in the ferredoxin utilizing pathways, and in the terminal oxidase supercluster; ferredoxins and Rieske-proteins. For the analysis of tryptophan/serine biosynthesis, only distances between enzymes are considered since the substrates do not involve proteins, i.e., are not genetically coded. A distance Δ between pathways is in general defined by

$$\Delta = \Phi_S \Delta S + \Phi_E \Delta E \tag{1}$$

where $\Phi_i = 1$ for orthologs, and $\Phi_i = f$ for paralogs, $i = S,E;$ for f we choose positive values specified further below.

Orthologs are genes in different species that evolved from a common ancestral gene by specification, paralogs are genes related by duplication within a genome (Fitch 1970). Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if related to the original one (Tatusov et al. 1997). The distinction between orthologs and paralogs per functional role is made to account for the fact that paralogs are more likely than orthologs to have diverged in function.

The program *Gapped BLAST* (Altschul et al. 1997) is used to find homologous sequences. We define a protein x from organism $\mathcal{A}$ as potential ortholog of a sequence y from organism $\mathcal{B}$ if (i) x and y are similar with an E-value, the expected fraction of false positives, smaller than 10^{-4}, and if (ii) there is no other sequence in $\mathcal{B}$ closer to x, and there is no other sequence in $\mathcal{A}$ closer to y. An assessment of orthologs and paralogs can only be done for completed genomes. In the case of genomes that are still involved in ongoing sequence projects, not all genes necessary for the classification between orthologs and paralogs might have been sequenced yet; nevertheless, the distinction between orthologs and paralogs are performed and the results are analyzed in this paper. In case of $f = 1$ (Eq. 1) ortholog and paralog pathway

[2] For a definition see Appendix

[3] See next section

Table 1. Genomes included in analysis

Code	Organism	D[a]	Size[kB][b]	# ORFs[c]	[d]	References[e]
AG	*Archaeoglobus fulgidus*	A	2178.40	2493	x	Klenk *et al.* 1997
AP	*Aeropyrum pernix*	A	1669.70	1631	x	Kawarabayasi *et al.* 1999
TH	*Methanobacterium thermoaut.*	A	1751.38	1866	x	Smith *et al.* 1997
PH	*Pyrococcus horikoshii*	A	1738.51	1825	x	Kawarabayasi *et al.* 1998
PF	*Pyrococcus furiosus*	A	1581.49	1932	o	http://www.ornl.gov/hgmis/publicat/99santa/157.html
MJ	*Methanococcus jannaschii*	A	1739.93	1811	x	Bult *et al.* 1996
AA	*Aquifex aeolicus*	B	1590.78	1774	x	Deckert *et al.* 1998
DR	*Deinococcus radiodurans*	B	3261.20	3771	x	White *et al.* 1999
EC	*Escherichia coli*	B	4639.22	4289	x	Blattner *et al.* 1997
YP	*Yersinia pestis*	B	4501.71	4296	o	http://www.sanger.ac.uk/Projects/Y_pestis/
HI	*Haemophilus influenzae*	B	1830.14	1846	x	Fleischmann *et al.* 1995
PA	*Pseudomonas aeruginosa*	B	6286.26	5642	o	http://www.pseudomonas.com/
NG	*Neisseria gonorrhoea*	B	2063.17	1853	o	ftp://ftp.genome.ou.edu/pub/gono
NM	*Neisseria meningitidis, Z2491*	B	2183.23	1838	x	Parkhill *et al.* 2000a
RC	*Rhodobacter capsulatus, SB1003*	B	2079.41	1989	o	http://capsulapedia.uchicago.edu/
RP	*Rickettsia prowazekii*	B	1101.98	849	x	Andersson *et al.* 1998
HP	*Helicobacter pylori*	B	1667.88	1547	x	Tomb *et al.* 1997
CJ	*Campylobacter jejuni*	B	1644.03	2106	x	Parkhill *et al.* 2000b
CY	*Synechocystis sp.*	B	3573.47	3226	x	Kaneko *et al.* 1996
CQ	*Chlamydia pneumoniae, CWL029*	B	1230.23	993	x	Kalman *et al.* 1999
CT	*Chlamydia trachomatis, serovar D*	B	1057.45	867	x	Stephens *et al.* 1998
BB	*Borrelia burgdorferi*	B	1519.86	1666	x	Fraser *et al.* 1997
TP	*Treponema pallidum*	B	1138.82	1031	x	Fraser *et al.* 1998
CA	*Clostridium acetobutylicum*	B	4030.73	3967	o	http://www.genomecorp.com/genesequences/clostridium/clospage.html
ML	*Mycobacterium leprae*	B	2420.76	1568	o	http://www.sanger.ac.uk/Projects/M_leprae/
MT	*Mycobacterium tuberculosis*	B	4411.53	3924	x	Cole *et al.* 1998
MG	*Mycoplasma genitalium*	B	580.07	532	x	Fraser *et al.* 1995
MP	*Mycoplasma pneumoniae*	B	816.39	674	x	Himmelreich *et al.* 1996
PN	*Streptococcus pneumoniae*	B	2104.82	1844	o	http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?organism =s_pneumoniae
ST	*Streptococcus pyogenes*	B	1799.24	1599	o	http://www.genome.ou.edu/strep.html
EF	*Enterococcus faecalis*	B	3209.12	2967	o	http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?organism =e_faecalis
BS	*Bacillus subtilis*	B	4214.81	4093	x	Kunst *et al.* 1997
SC	*Saccharomyces cerevisiae*	E	12057.28	6125	x	Goffeau *et al.* 1997
CE	*Caenorhabditis elegans*	E	165227.99	16332	x	The *C. elegans* Sequencing Consortium, 1998

[a] Domain: A . . . archaea, B . . . bacteria, E . . . eukarya
[b] For ongoing sequence projects subject to change
[c] ORF: open reading frame
[d] Ongoing sequence projects are marked by o; completely sequenced genomes are marked by x
[e] Either citation or URLs of sequencing institution

representations (PRs[4]) are treated in the same way; for $f < 1$ ($f > 1$) the total distance Δ between paralog PRs is smaller (larger) than $\sum_i \Delta X_i$, where X_i stands for a functional role, as S or E. The parameter f is chosen such that the calculated distance matrix (which yields the phylogenetic tree) has a minimum number of distance triples that violate the triangle inequality. The optimal value for f has been obtained by visual inspection.

Definition (1) can be generalized to n functional roles per pathway. We denote by Γ and Γ' metabolic pathways of identical topology involving n functional roles I_i, I_i', $i = 1 \ldots n$ and by $\Delta X_i = \delta(I_i, I_i')$ distances between functional roles I_i and I_i' calculated utilizing an alignment δ. A distance Δ between Γ and Γ' is then defined through

$$\Delta = \sum_{i=1}^{n} \Phi_i \cdot \Delta X_i, \quad \Phi_i = \begin{cases} 1 & \text{for ortholog pair } i \\ f & \text{for paralog pair } i \end{cases} \qquad (2)$$

where $f > 0$. The chosen forms for Eq. (1) and Eq. (2) are motivated by assuming independent and, thus, additive contributions of distances between sequences per functional role to a pathway distance.

Graphs and Gaps

The graph-topology of a metabolic network is represented as an adjacency matrix, an $n \times n$-matrix for n functional roles, with non-zero elements for pairs of functional roles that are connected by a reaction using a common substrate. Substrates can be non-genome coded, e.g., NADPH, or genome coded, e.g., ferredoxin. In the latter case the substrate makes a contribution to the distance between pathways. If two networks with different graph-topology are compared to each other, the common graph that includes both networks is considered and its adjacency matrix is used. An example is shown in Fig. 1: one network represents a cyclic action scheme and a second network consists of a linear part of the cyclic reaction with an alternative route between two nodes of the network, the common network would refer to the cyclic reaction plus the alternative route.

[4] For a definition see Appendix

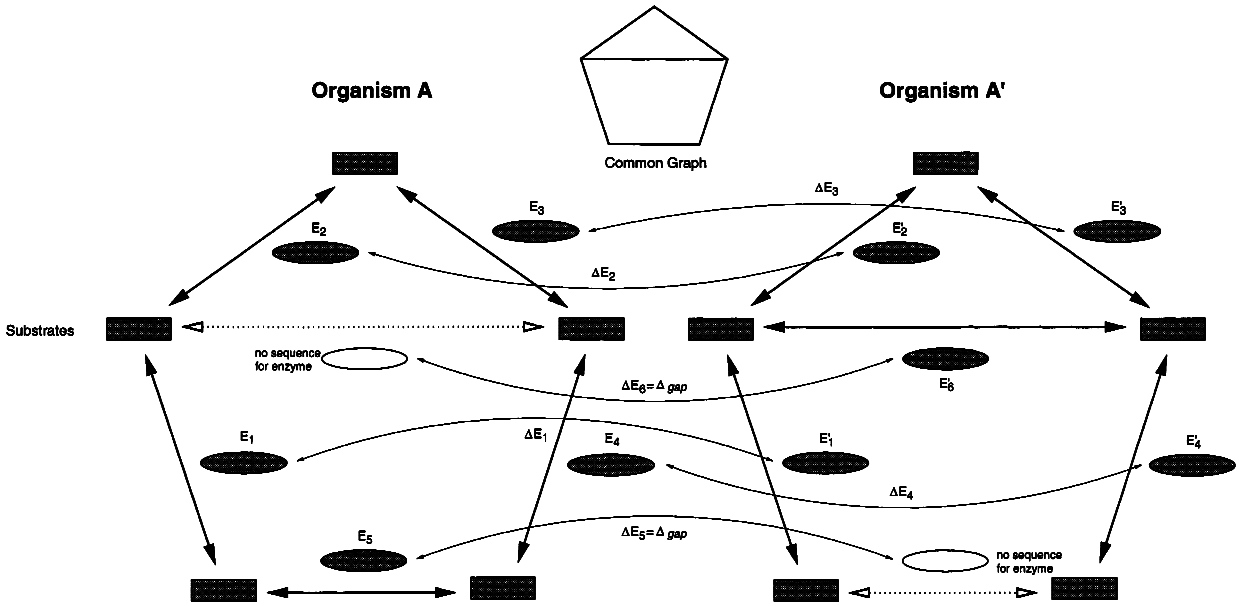

Fig. 1. Two networks and their common network. In this example two enzymes E'_5 and E'_6 are not present in both networks. From this results differences in graph-topology between networks of organism A and organism A′, one being a cyclic reaction scheme for A that becomes a linear scheme due to an absent enzyme E'_5 in A′, and another being a shortcut reaction via E'_6 that is not present in A. Gap penalty Δ_{gap} is assigned to the corresponding distances ΔE_5 and ΔE_6.

Based on the non-zero entries in the adjacency matrix the distance between both networks is calculated according to Eq. (2). If a functional role I_k is missing in a pathway Γ then the distance ΔX_k in Eq. (2) is not defined properly. In this case, to the otherwise undefined distance Δ_k a gap value Δ_{gap} is assigned; throughout the following, a value of $\Delta_{gap} = 0.9$ is assumed. This gap value must not be mistaken with the gap values used in sequence alignment.

A threshold t that corresponds to a confidence level for the minimum number of present functional roles is defined as follows: let Γ be a pathway with k functional roles; furthermore, let l be a lower bound of the number of functional roles that have to be present in a functioning pathway; then the confidence level is defined through

$$t = \frac{l}{k}, \qquad 0 \le t \le 1 \tag{3}$$

If the fraction of functional roles falls below t then a pathway is considered incomparable. In order to define a distance between any two pairs of pathways the distance Δ between an arbitrary pathway and an incomparable pathway is replaced by a penalty distance Δ_p.

Values for Δ_{gap} and Δ_p reflect estimates for expected distances. Δ_{gap} is chosen as the expected average distance between individual functional roles. Δ_p is an estimate for a distance between two pathways; it is typically of the order $n \cdot \Delta_{gap}$ where n has been defined above. Gap values are employed in examples discussed below.

Results and Discussion

We have investigated three different kinds of pathways, pathways that utilize ferredoxin in electron transfer reactions, terminal oxidation in quinol and cytochrome *c* oxidase complexes and the interconnected pathways of tryptophan and serine biosynthesis.

Ferredoxin–NADPH Reductase Pathway

We first apply our phylogenetic analysis to the ferredoxin–NADPH reductase pathway. This particular pathway has been chosen because it is of the most basic type, involving a single enzyme processing a single substrate, and because of the importance of ferredoxin in other pathways. The genomes which we considered in our analysis are collected in Table 1. Ferredoxins serve as electron acceptors and donors in many anabolic, catabolic, and electron transfer reactions, e.g., as redox partner in more than 50 known pathways. The pathway and a non-rooted phylogenetic tree generated by programs of the *PHYLIP* software suite are shown in Fig. 2. The phylogenetic tree was generated using a single functional role per ferredoxin and reductase. Each leaf or terminal node of the phylogenetic tree (Fig. 2b) displays the label that provides a unique pathway identification (pID); the labels refer to the name of the organism and the combination of functional roles used. For example, *M. tuberculosis*, with organism-code MT, according to Table 1, has three paralogs which code for ferredoxin (fdxA, fdxC, and fdxD) as well as two paralogs that code for ferredoxin reductase (fpr and fprA). Using all possible combinations between ferredoxin and ferredoxin reductase yields six representations of pathways[5] for *M. tuberculosis* that are listed in Table 2.

Table 3 shows pIDs as well as corresponding ORF-

[5] For a definition see Appendix

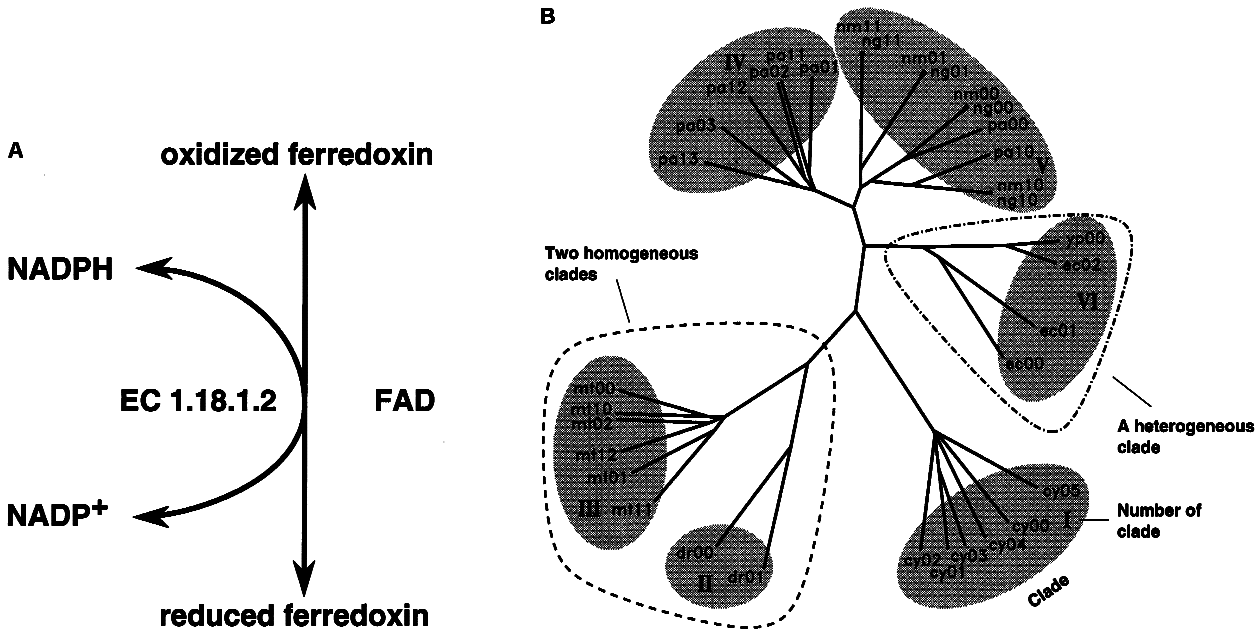

Fig. 2. Ferredoxin–NADPH reductase pathway. **(A)** Pathway shown with ferredoxin and ferredoxin reductase (EC 1.18.1.2). **(B)** Phylogenetic tree of the pathway established using the *PHYLIP* software suite with parameters $f = 0.5$, and $t = 1$. Clade I to VI are referred to in the text. The id-numbers (pIDs), which are uniquely assigned to each pathway representation, are a combination of two letters which code for the organism, as listed in Table 1, and a number. References from pIDs of the pathway representation to ORF-names of the corresponding functional roles per pathway used are listed in Table 3.

Table 2. Ferredoxin–NADPH reductase pathway in *M. tuberculosis*

#	pID	Ferredoxin	Reductase
1	mt00	fdxC_MYCTU	fpr_MYCTU
2	mt01	fdxA_MYCTU	fpr_MYCTU
3	mt02	fdxD_MYCTU	fpr_MYCTU
4	mt10	fdxC_MYCTU	fprA_MYCTU
5	mt11	fdxA_MYCTU	fprA_MYCTU
6	mt12	fdxD_MYCTU	fprA_MYCTU

names for all organisms of the phylogenetic tree in Fig. 2b; it also lists the so-called clades to which the pIDs belong. A clade is a set of closely related PRs that is presented as a subtree in the phylogeny as indicated in Fig. 2b, and can refer to one or more leafs in the phylogenetic tree. Examples for homogeneous clades[6] or paralog PRs (Fig. 2) are clade I (*Synechocystis sp.*), II (*D. radiodurans*), III (*M. tuberculosis*), and IV (*P. aeruginosa*). The cyanobacterium *Synechocystis sp.* (clade I), which carries a complete set of genes for oxygenic photosynthesis, is clearly separated from the non-autotroph bacteria. An interesting pair of homogeneous clades correspond to *D. radiodurans* (clade II) and *M. tuberculosis* (clade III). Not only the ferredoxin–NADPH reductase pathway but also other electron transfer pathways, such as the malate–aspartate shuttle, are closely related in case of the pathogen *Mycobacteria* and the ultrahigh-radiation tolerant *D. radiodurans* (phylogeny not shown). A special case arises for clade IV. The PRs of *P. aeruginosa* are present both in the homogeneous clade IV as well as in clade V.

Clade V and VI are examples of heterogeneous clades involving *E. coli*, *Y. pestis*, *Neisseria*, and *P. aeruginosa*. Clade VI is outlined in Fig. 2b. In clades V and VI distances between PRs of different organisms are shorter than a maximal distance which would include all paralog pathway representations of one organism. For example, the distance between *ec02* and *yp00* in clade VI is shorter than the distance between *ec02* and *ec00*. The close relationship between genera of the Enterobacteriaceae family such as *E. coli* and *Y. pestis* is evident in the observed clustering of *E. coli* and *Y. pestis* PRs. In contrast to a clustering of paralog PRs to form homogeneous clusters involving each *E. coli* and *Y. pestis*, as observed in clades I–IV in Fig. 2b, and as it would be in the case of common ancestry of the ferredoxin–NADPH reductase pathway, a heterogeneous cluster involving *E. coli* and *Y. pestis* is formed in clade VI. This finding suggests horizontal transfer of ferredoxin–NADPH reductase PRs between these organisms. Surprising similarities also exist between *Neisseria* and *P. aeruginosa*. In addition to the homogeneous clade of *P. aeruginosa* PRs (clade IV), a second, heterogeneous clade with *Neisseria* is observed (clade V). The relationship between *Neisseria* and *P. aeruginosa*, as seen in clade IV and V in Fig. 2b, is not as robust as in clades I–III, and VI. If one changes parameter *f*, clades I–III and VI preserve the graph-topology and, thus, their relationship between PRs. This is not the case for clades IV and V, for which the rela-

[6] For a definition see Appendix

Table 3. List of pathway identifications (pIDs) referring to ORF-names with corresponding functional roles

pID	Ferredoxin	Reductase	Cl.[a]	pID	Ferredoxin	Reductase	Cl.
cy00	slr1205	slr1643	I	ng00	RNG01106[b]	RNG00591	V
cy01	slr0150	slr1643	I	ng01	RNG00533	RNG00591	V
cy02	ssr3184	slr1643	I	ng10	RNG01106	RNG00984	V
cy03	ssl0020	slr1643	I	ng11	RNG00533	RNG00984	V
cy04	sll0662	slr1643	I	nm00	RNM00363	RNM01731	V
cy05	slr0148	slr1643	I	nm01	RNM00662	RNM01731	V
dr00	DR2330	DR0496	II	nm10	RNM00363	RNM00963	V
dr01	DR2075	DR0496	II	nm11	RNM00662	RNM00963	V
ec00	ykgJ_ECOLI	fpr_ECOLI	VI	pa00	RPA01015	RPA07749	V
ec01	yfhL_ECOLI	fpr_ECOLI	VI	pa01	RPA01568	RPA07749	IV
ec02	fdx_ECOLI	fpr_ECOLI	VI	pa02	RPA06033	RPA07749	IV
mt00	fdxC_MYCTU	fpr_MYCTU	III	pa03	RPA08046	RPA07749	IV
mt01	fdxA_MYCTU	fpr_MYCTU	III	pa10	RPA01015	RPA05251	V
mt02	fdxD_MYCTU	fpr_MYCTU	III	pa11	RPA01568	RPA05251	IV
mt10	fdxC_MYCTU	fprA_MYCTU	III	pa12	RPA06033	RPA05251	IV
mt11	fdxA_MYCTU	fprA_MYCTU	III	pa13	RPA08046	RPA05251	IV
mt12	fdxD_MYCTU	fprA_MYCTU	III	yp00	RYP01051	RYP02807	VI

[a] Clade

[b] RXXnnnnn refers to so-called reference identification numbers that code for ORFs in the WIT system; XX denotes the two letter codes shown in Table 1, nnnnn is a five-digit number that is uniquely assigned to an ORF per organiism.

tionship between PRs change with *f*. This weak confidence in clade IV suggests frequent and random gene replacements between *Neisseria* and *P. aeruginosa*.

Within the genus *Neisseria* we note that pathway representations always show up as ortholog pairs. Examples of such pairs of ortholog PRs in clade V are (*ng00, nm00*), (*ng01, nm01*), (*ng10, nm10*), and (*ng11, nm11*).

Prokaryotic organisms adapt to the need of commonly used proteins by establishing more than one representation in the genome. Often only a single gene codes for reductase but there may be as many as six ferredoxin genes present, e.g., in *Synechocystis sp.* (clade I). The abundance of genes coding for ferredoxin compared to genes which code for ferredoxin reductase stem from the universality of ferredoxin. Ferredoxins, as redox reagens, serve in many more biochemical redox-reactions than ferredoxin reductase.

For comparison of the phylogenetic analysis from the ferredoxin–NADPH reductase pathway with phylogenies of the individual sequences, phylogenetic trees of ferredoxins as well as ferredoxin–NADPH reductases are shown in Fig. 3. One of the major differences between pathway-phylogenies (Fig. 2b) and phylogenies of individual sequences in Fig. 3 is that in the former phylogenies paralog PRs are more closely related than ortholog PRs in most cases. In phylogenies of individual sequences, orthologs are more closely related than paralogs.

In contrast to the closely related paralog PRs of *M. tuberculosis* (Fig. 2b, clade III), paralog genes shown in Fig. 3 that code for ferredoxins (fdxA_MYCTU, fdxC_MYCTU, fdxD_MYCTU) and reductases (fpr_MYCTU and fprA_MYCTU) are more closely related to genes coding for ferredoxin and reductase, respectively, in different organisms. For example, ferredoxins fdxA_MYCTU and fdxC_MYCTU are more similar to ssr3184 and RPA08046 than to fdxD_MYCTU. Analogously, ferredoxin–reductase fprA_MYCTU is more closely related to DR0496 than to fpr_MYCTU. On the other hand, rather similar genes, such as *Neisseria* and *Pseudomonas* genes coding for ferredoxins (RNG01106, RNM00363, RPA01015, and RNG00533, RNM00662, RPA01568, in Fig. 3a) as well as reductases (RNG00591, RNM01731, RPA07749, and RNG00984, RNM00963, RPA05251, in Fig. 3b) contribute to PRs in the pathway-phylogeny with orthologous relationships (ng00, nm00, pa00, and ng10, nm10, pa10, in Fig. 2b, clade V). This comparison between phylogenies of individual genes on the one hand and the phylogeny of the pathway on the other hand suggests that, for a single organism, PRs are conserved, but single genes of their functional role do not share as close a relationship as PRs do.

The advantage of the presented approach of pathway-phylogenies is that it permits easy identification of conserved relationships between genes for each functional role (here for both ferredoxin and reductase). Thus, if both ferredoxins as well as NADPH-reductases are closely related to each other, as shown in Fig. 3 in the case of *M. tuberculosis* and *D. radiodurans*, then this conservation of relationships is reflected in a close relationship in the phylogeny of the ferredoxin–NADPH-reductase pathway (Fig. 2b).

Pathways with Ferredoxin as Functional Role. Out of about 50 known pathways in which ferredoxin plays a significant functional role, seven pathways were chosen for further investigations. The pathways are listed in Table 4. The remaining pathways are either found in one organism only or are incomparable due to uncharacter-

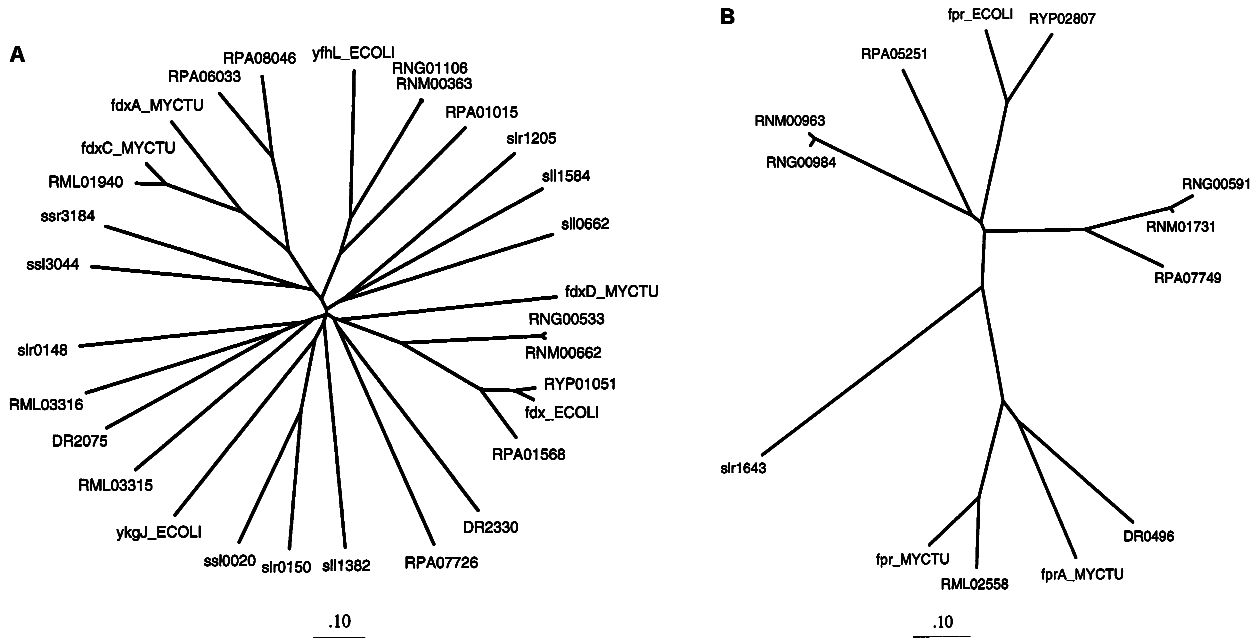

Fig. 3. Separate phylogenies of ferredoxin and NADPH reductase. Phylogenetic trees for **(A)** ferredoxin and **(B)** ferredoxin reductase (EC 1.18.1.2) were established using the *PHYLIP* software suite.

ized functional roles. The latter case occurs for rare pathways which were discovered for microbial organisms that are related to organisms in Table 1. For example, the pathway of toluene (or cyclohexanol) degradation to protocatechuate is only present in *Pseudomonadaceae.* A representative of this genus in Table 1 is *P. aeruginosa.* The toluene (or cyclohexanol) degradation pathway is counted as one out of 50 known pathways using ferredoxin as redox-reagens. Beside ferredoxin, there are three enzymes in this pathway. Information for all necessary enzymes is available only for *P. putida,* but not for *P. aeruginosa* sequences. Thus, even if the toluene (and cyclohexane) degradation pathway, using ferredoxin, exists in *Pseudomonadaceae,* it has not been considered for further investigations due to incomparable pathways in organisms of Table 1.

Figure 4a shows a phylogenetic graph of 418 representations of 7 pathways utilizing ferredoxin; this representation was determined for a confidence level $t = 0.5$. Pathways of the same organism are found in single clades. Comparing the number of PRs per organisms signifies both the importance of pathways involving ferredoxin for the distinct species as well as the completeness of the set of pathways. *Escherichia coli,* with the most complete set of pathways, leads the group of organisms with 200 PRs out of 418 shown. The highly adaptable clinical pathogen, *P. aeruginosa,* and the versatile solventogenic, *C. acetobutylicum,* follow with 72 and 42 PRs, respectively. The anaerobic submarine archaeon, *A. fulgidus,* is represented with 48 PRs. The capability of this organism to grow, using sulfate or thiosulfate as electron acceptors, and hydrogen or carbon dioxide, as well as complex organic material, for energy, explains the need of ferredoxin utilizing pathways.

Figure 4b depicts a reduced phylogeny of the same data. The phylogeny is simplified such that only one leaf per organism is shown. Similar to the observation reported in the previous section, PRs of *E. coli* and *Y. pestis* are closely related to each other (clade I). Close to clade I a second clade is formed by *M. tuberculosis, P. aeruginosa,* and *Synechocystis sp.* An interesting similarity exists between the hyper-thermophilic bacterium *A. aeolicus* and the archaeon *A. fulgidus* (clade III). *Aquifer,* with its representative *A. aeolicus,* is exceptional among bacteria in the way that it occupies the hyperthermophilic niche otherwise dominated by archaea (Pace 1997). Whether the observed close relationship of ferredoxin-utilizing pathways between *Aquifex* and *Archaeoglobus* is caused by continuous acquisition of thermotolerance genes from preadapted hyperthermophiles, or whether it is just a consequence of adaptation to an extreme thermophilic environment, cannot be decided with the present number of completely sequenced microbial genomes. More genomes of extremophilic archaea and thermophilic bacteria are necessary to detect possible horizontally transferred pathways.

Terminal Oxidase Complexes

Many similarities between the *E. coli* cytochrome bo_3 complex and the aa_3-type cytochrome *c* oxidase complexes led to the recognition of a superfamily of terminal oxidase complexes related in structure and function. This superfamily is divided into two families: the quinol (typically ubiquinol) oxidase complexes, and the cytochrome *c* oxidase (CcO) complexes. In general, all members of the superfamily of terminal oxidase complexes are

Table 4. Pathways with ferredoxin as functional role

	Pathway with functional roles	#[a]	Codes[b] of organisms for which pathways are assigned
1	2-Oxoglutarate, Glutamine–Glutamate Anabolism (Reduced Ferredoxin)	2	-- -- -- --
	EC 1.4.7.1: Glutamate Synthase		-- -- -- -- -- PA -- -- -- -- -- CY -- -- -- CA -- -- -- -- -- -- -- BS
	Ferredoxin		-- --
2	NADPH–Oxidized Ferredoxin Electron Transport	2	-- -- -- --
	EC 1.18.1.2: Ferredoxin–$NADP^+$ Reductase		-- DR EC YP -- PA NG NM -- -- -- CY -- -- -- -- -- MT -- -- -- -- -- --
	Ferredoxin		-- --
3	NADPH–Oxidized Ferredoxin Electron Transport (Plasma Membrane)	2	-- -- -- --
	EC 1.18.1.2: Ferredoxin–$NADP^+$ Reductase		-- DR EC YP -- PA NG NM -- -- -- CY -- -- -- -- -- MT -- -- -- -- -- --
	Ferredoxin		-- --
4	H_2–H^+ Catabolism (Oxidized Ferredoxin)	2	AG -- -- --
	EC 1.18.99.1: Hydrogenase		AA -- EC -- -- -- -- -- -- -- CJ CY -- -- -- CA -- -- -- -- -- -- -- --
	Ferredoxin		-- --
5	H^+–H_2 Anabolism (Reduced Ferredoxin) (Plasma Membrane)	2	AG -- -- --
	EC 1.18.99.1: Hydrogenase		AA -- EC -- -- -- -- -- -- -- CJ CY -- -- -- CA -- -- -- -- -- -- -- --
	Ferredoxin		-- --
6	Nitrate–NH_4^+, OH^- Catabolism (Ferrocytochrome 'c_{552}', Reduced Ferredoxin)	3	AG -- -- --
	EC 1.7.99.4: Nitrate Reductase		-- -- -- -- -- -- -- -- -- -- -- CY -- -- -- -- -- MT -- -- -- -- -- --
	EC 1.7.7.1: Ferredoxin–Nitrite Reductase		-- --
	Ferredoxin		
7	Phosphoadenylylsulfate–Sulfide Anabolism	4	-- -- -- --
	EC 1.8.99.4: Phosphoadenosine Phosphosulfate Reductase		-- -- -- -- -- PA -- -- -- -- -- CY -- -- -- -- -- -- -- -- -- ST -- -- --
	EC 1.8.7.1: Sulfite Reductase		SC --
	Ferredoxin		
	Thioredoxin		

[a] Number of functional roles per pathway
[b] The two letter code refers to organisms as listed in Table 1

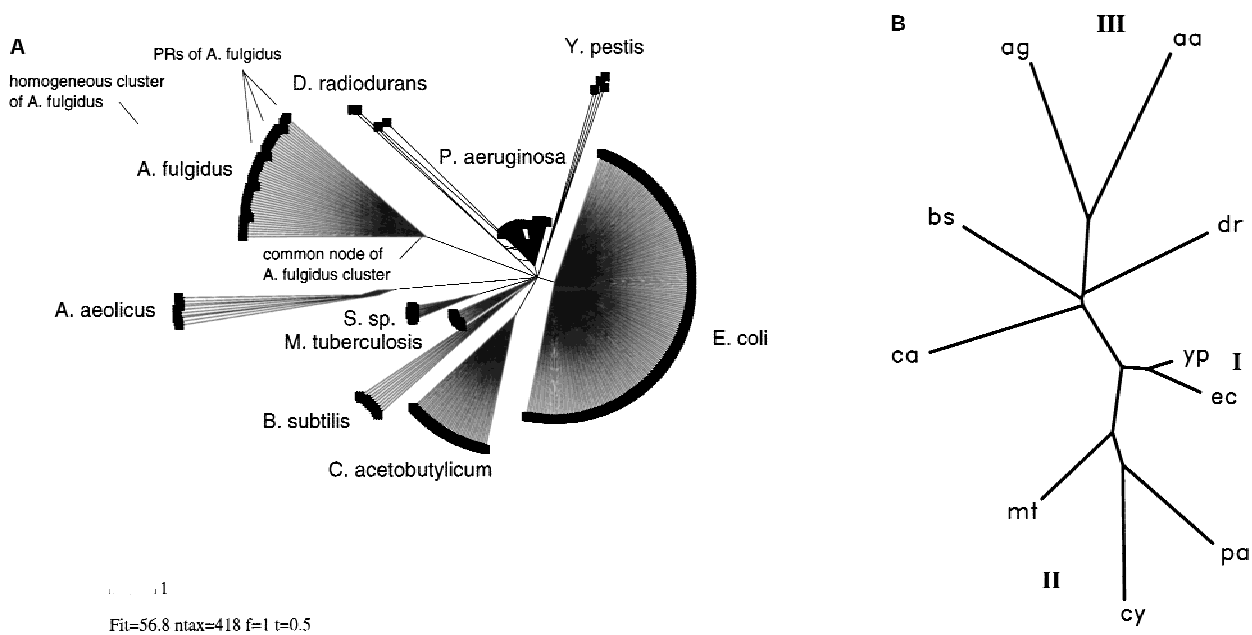

Fig. 4 Ferredoxin-related pathways. (**A**) 418 different representations of ferredoxin-related pathways are shown. 10 for *A. aeolicus*, 48 for *A. fulgidus*, 10 pathways for *B. subtilis*, 42 for *C. acetobutylicum*, 4 for *D. radiodurans*, 200 for *E. coli*, 12 for *M. tuberculosis*, 72 for *P. aeruginosa*, 16 for *Synechocystis sp.*, and 4 for *Y. pestis*. The non-rooted phylogenetic graph has been established using the program *SPLITSTREE2*. (**B**) A simplification of the phylogenetic tree is established using the *PHYLIP* software suite; each leaf represents the node of each of the homogeneous clades in (**A**). The distance matrix has been created with parameters $f = 1$ and $t = 0.5$.

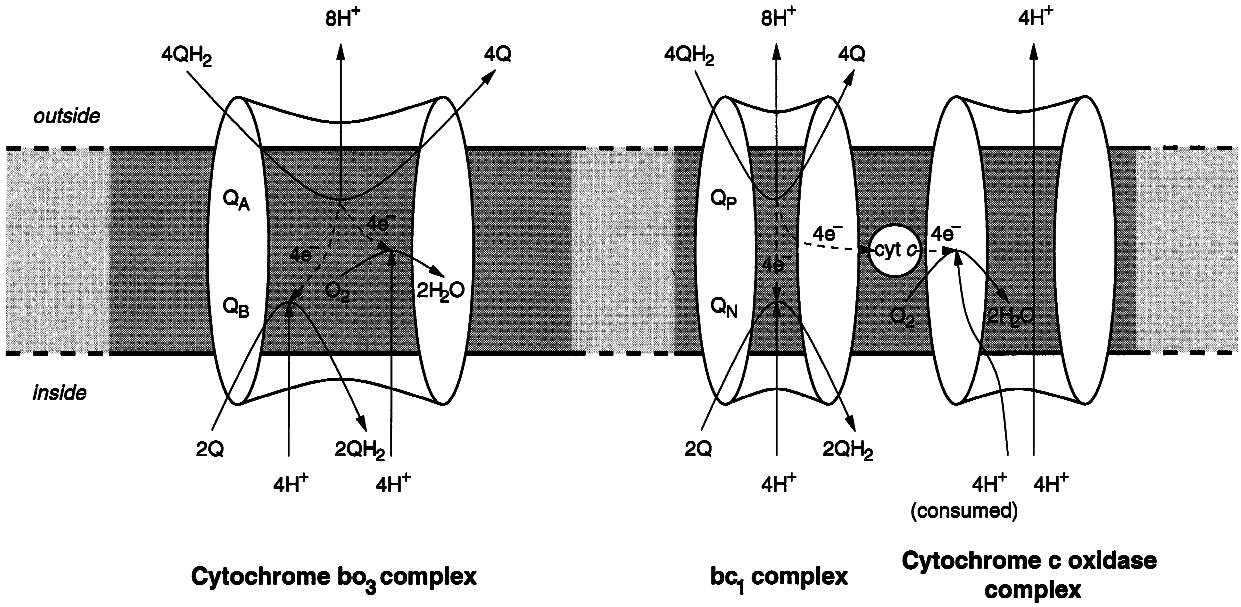

Fig. 5. Schema of the chemistry of cytochrome bo_3 and cytochrome bc_1/cytochrome *c* oxidase complexes. Both complexes perform the same overall reaction but with different efficiency. In addition to the less efficient cytochrome bo_3 complex that transallocates 8 H^+, the combined cytochrome bc_1/cytochrome *c* oxidase complex pumps four more protons through the complex.

thought to translocate protons against a transmembrane potential gradient, and all appear to catalyze dioxygen activation and reduction at a heme-copper binuclear center. Recently, Musser and Chan (1998) performed a detailed survey on the evolution of the cytochrome *c* oxidase proton pump. Their results serve as a reference to test the validity of the phylogenetic analysis presented here.

Musser and Chan suggest a primitive quinol oxidase complex as a common ancestor of all enzymes included in the scheme of cytochrome *c* oxidase complexes. As a member of the quinol oxidizing systems, cytochrome bo_3 is structurally and functionally the simplest complex found in present day organisms such as *E. coli*, *P. aeruginosa*, and *Y. pestis*. Stoichiometric and energetic considerations indicate that the cytochrome bo_3 complex carries out, in effect, the combined reactions catalyzed by the cytochrome bc_1, and cytochrome *c* oxidase complexes including a net translocation of 8 H^+:

$$\begin{aligned} &4\ QH_2\ (Q_A\ \text{site}) + 4\ H^+\ (\text{consumed, } Q_B\ \text{site}) + O_2 \\ &\qquad + 4\ H^+\ (Q_B\ \text{site}) + 2\ Q\ (Q_B\ \text{site}) \\ &\rightarrow 4\ Q\ (Q_A\ \text{site}) + 8\ H^+\ (Q_A\ \text{site}) + 2\ H_2O \\ &\qquad + 2\ QH_2\ (Q_B\ \text{site}) \end{aligned}$$

corresponding to a net reaction:

$$2\ QH_2 + 8\ H^+\ (Q_B\ \text{site}) + O_2 \rightarrow 2\ Q + 8\ H^+\ (Q_A\ \text{site}) + 2\ H_2O$$

The combination of the cytochrome bc_1 and cytochrome *c* oxidase complexes carry out the same overall reaction, but in two steps and with a net translocation of $12H^+$:

$$\begin{aligned} &4\ QH_2\ (Q_P\ \text{site}) + 4\ H^+\ (Q_N\ \text{site}) + 2\ Q\ (Q_N\ \text{site}) \\ &\qquad + 4\ \text{cyt}\ c^{3+} \quad \rightarrow 4\ Q\ (Q_P\ \text{site}) \\ &\qquad + 8\ H^+\ (Q_P\ \text{site}) \quad (\text{cyt}\ bc_1) \\ &\qquad + 2\ QH_2\ (Q_N\ \text{site}) + 4\ \text{cyt}\ c^{2+} \end{aligned}$$

$$\begin{aligned} &4\ \text{cyt}\ c^{2+} + 4\ H^+\ (\text{consumed, inside}) + O_2 \\ &\qquad + 4\ H\ (\text{inside}) \rightarrow 4\ \text{cyt}\ c^{3+} + 2\ H_2O \\ &\qquad + 4\ H^+\ (\text{outside}) \quad (\text{CcO}) \end{aligned}$$

The net reaction of these two steps is:

$$2\ QH_2 + 12\ H^+\ (\text{inside}) + O_2 \rightarrow 2\ Q + 12\ H^+\ (\text{outside}) + 2\ H_2O$$

The intriguing structural and functional relationship between the cytochrome bo_3 complex and cytochrome bc_1/cytochrome *c* oxidase suggests that during evolution a primitive quinol oxidase complex split into two separate enzyme complexes which catalyze the same net-chemistry, but with separate complexes linked together working more efficiently in terms of energy conservation (Musser and Chan 1998). A schematic comparison of the cytochrome bo_3 complex and the cytochrome bc_1/cytochrome *c* oxidase complex is provided in Fig. 5.

For our analysis, a hypothetical cytochrome *c* oxidase pathway with six functional roles has been constructed. The corresponding functional roles are: cytochrome *c* oxidase polypeptide, I to IV (1)–(4); bc_1 complex or its homologs, cytochrome bo_3, or quinol oxidase (5); and the Rieske protein (6). The remaining polypeptides, V to VIII, for cytochrome *c* oxidase are not included because sequences which code for these polypeptides have only been found in *S. cerevisiae*. Table 5 shows the functional

Table 5. Terminal oxidase complexes

pID	COX ppI[a]	COX ppII	COX ppIII	COX ppIV	bc_1/bo/qox[b]	Rieske protein
aa0	coxA1	coxB2	coxC	—	petB	petA
aa1	coxA1	coxB	coxC	—	petB	petA
ag0	—	AF0144	—	—	—	—
ag1	—	AF0142	—	—	—	—
bs qox	QOX1_BACSU	QOX2_BACSU	QOX3_BACSU	—	—	—
bs0	ctaD	ctaC	ctaE	ctaF	QCRC_BACSU	QCRA_BACSU
bs1	ctaD	ctaC	ctaE	ctaF	QCRB_BACSU	QCRA_BACSU
bs cytBO	—	—	—	—	yjdK	—
ce0	COX1_CAEEL	COX2_CAEEL	COX3_CAEEL	—	RCE02468[c]	—
ce1	COX1_CAEEL	COX2_CAEEL	COX3_CAEEL	—	RCE05944	—
cy0	slr2082	slr1136	slr2083	—	—	—
cy1	slr2082	slr1136	slr1138	—	—	—
cy2	slr2082	sll0813	slr2083	—	—	—
cy3	slr2082	sll0813	slr1138	—	—	—
cy4	slr1137	slr1136	slr2083	—	—	—
cy5	slr1137	slr1136	slr1138	—	—	—
cy6	slr1137	sll0813	slr2083	—	—	—
cy7	slr1137	sll0813	slr1138	—	—	—
dr	RDR03431	RDR03432	—	—	—	—
ec cytBO0	—	—	—	—	CYOB_ECOLI	—
ec cytBO1	—	—	—	—	CYOC_ECOLI	—
ec cytBO2	—	—	—	—	CYOD_ECOLI	—
ec cytBO3	—	—	—	—	CYOA_ECOLI	—
ml	RML00132	RML00125	—	—	—	—
mp	ctaD	—	—	—	—	—
mt	ctaD	ctaC	ctaE	—	—	—
ng	RNG00242	—	—	—	—	RNG01710
nm	RNM00498	—	—	—	—	RNM00680
pa0	RPA01502	RPA01501	RPA07870	—	—	RPA02313
pa1	RPA05019	RPA01501	RPA07870	—	—	RPA02313
pa2	RPA03570	RPA01501	RPA07870	—	—	RPA02313
pa3	RPA06826	RPA01501	RPA07870	—	—	RPA02313
pa4	RPA02545	RPA01501	RPA07870	—	—	RPA02313
pa cytBO0	—	—	—	—	RPA05051	—
pa cytBO1	—	—	—	—	RPA05049	—
pa cytBO2	—	—	—	—	RPA02450	—
pa cytBO3	—	—	—	—	RPA02451	—
rc	—	—	—	—	—	UCRI_RHOCA
sa qox	—	—	—	—	qox_SULAC	—
sa0	QOX1_SULAC	QOX2_SULAC	—	—	CYB_SULAC	—
sa1	QOX1_SULAC	QOX2_SULAC	—	—	cyb2_SULAC	—
sc1	COX1_YEAST	COX2_YEAST	COX3_YEAST	COX4_YEAST	UCRH_YEAST	UCRI_YEAST
sc2	COX1_YEAST	COX2_YEAST	COX3_YEAST	COX4_YEAST	UCR7_YEAST	UCRI_YEAST
sc3	COX1_YEAST	COX2_YEAST	COX3_YEAST	COX3_YEAST	UCRQ_YEAST	UCRI_YEAST
sc4	COX1_YEAST	COX2_YEAST	COX3_YEAST	COX4_YEAST	UCR2_YEAST	UCRI_YEAST
yp cytBO0	—	—	—	—	RYP03503	—
yp cytBO1	—	—	—	—	RYP03502	—
yp cytBO2	—	—	—	—	RYP03143	—
yp cytBO3	—	—	—	—	RYP03504	—

[a] Cytochrome *c* Oxidase Polypeptide I
[b] Quinol Oxidase
[c] See footnote[b] in Table 3

roles for each pathway representation. In addition to the sequence information of 30 genomes, following Musser and Chan, we also include in our analysis sequences of terminal oxidase complexes for *Sulfologus acidocaldarius* with organism code SA as listed in Table 5. The latter is exceptional among the studied organisms because it uses two different quinol oxidase complexes, *sa qox*, *sa0*, and *sa1*.

Figure 6 shows schematics of terminal oxidase complexes of *E. coli*, *B. subtilis*, and *S. cerevisiae* (Fig. 6a), as well as the phylogeny for the cytochrome *c* oxidase, and quinol oxidase complexes (Fig. 6b). A division between cytochrome bo_3/ubiquinol oxidase complexes (*cytBO*) and simple quinol ba_3 oxidase (*qox*) complexes, on the one hand, and cytochrome bc_1/cytochrome *c*/cytochrome *c* oxidase super-complexes, cytochrome

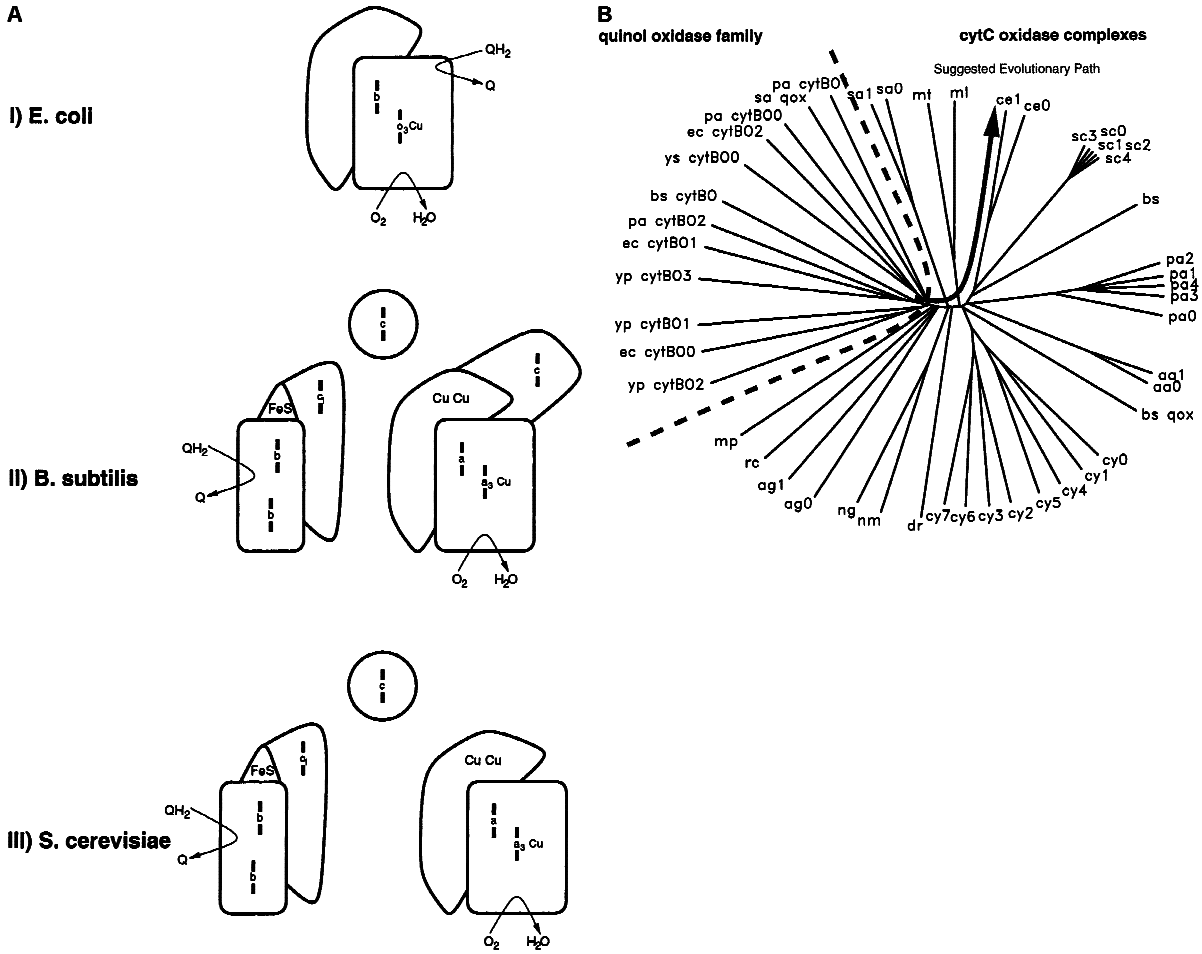

Fig. 6. Terminal oxidase complexes. **(A)** Schematics of three exemplary terminal oxidase complexes are shown: I) the *E. coli* cytochrome bo_3 ubiquinol oxidase complex; II) the cytochrome caa_3 complex of *B. subtilus;* III) the mitochondrial cytochrome bc_1 and cytochrome aa_3 (cytochrome *c* oxidase) complexes of *S. cerevisiae.* **(B)** The phylogenetic relationship of cytochrome *c* and quinol oxidase complexes are depicted. The thick dashed line indicates the division between quinol oxidase and cytochrome *c* oxidase complexes. The solid line denotes a suggested evolutionary path (see text). Parameters for constructing the phylogeny are $f = 0.25$ and $t = 0.01$.

caa_3 complexes, mitochondrial cytochrome bc_1/cytochrome aa_3 (cytochrome *c* oxidase) complexes, and aa_3 type quinol oxidase complexes, on the other hand, can be noticed. A close relationship of cytochrome bo_3 between *E. coli, P. aeruginosa,* and *Y. pestis* is recognized (left part of Fig. 6b); all three organisms use the cytochrome bo_3 complex as a terminal oxidase. Schematics of this complex is shown in Fig. 6a/I.

As in the analysis of Musser and Chan (1998), a progression from simple cytochrome *c* complexes (*S. acidocaldarius*) via *Mycobacteriaceae, Synechocystis, A. aeolicus,* and *P. aeruginosa,* to the complete set of the cytochrome caa_3 complex of *B. subtilis* (Fig. 6a/II), and the mitochondrial cytochrome bc_1, and cytochrome aa_3 (cytochrome *c* oxidase) complexes of *S. cerevisiae,* and *C. elegans* (Fig. 6a/III) is observed. Musser and Chan conclude that the common ancestor of cytochrome bc_1/cytochrome *c* oxidase complexes is a quinol oxidase complex. Figure 6b indicates the suggested evolutionary path from quinol terminal oxidase complexes and cytochrome bo_3 ubiquinol oxidase complexes to mitochondrial cytochrome bc_1/cytochrome *c* oxidase complexes, that resembles the evolutionary tree suggested by Musser and Chan and, thus, reconfirms their conclusion.

Tryptophan Biosynthesis Network

In the last step of tryptophan biosynthesis, serine combines with indoleglycerol phosphate to produce tryptophan and glyceraldehyde-3-phosphate. The two glycolytic enzymes that are present in almost all organisms, glyceraldehyde-3-phosphate dehydrogenase (*gapA*) and phosphoglycerate mutase (*pgk*), recycle the three-carbon glyceraldehyde-3-phosphate to 3-phosphoglycerate. The latter is then transformed via phosphoglycerate dehydrogenase (*serA*), phosphoserine transaminase (*serC*), and phosphoserine phosphatase (*serB*) to serine. Tryptophan itself is synthesized from chorismate via anthranilate synthase component α and β (*trpE,* and *trpG*), anthranilate phosphoribosyl transferase (*trpD*), N-(5′-phosphoribosyl)anthranilate isomerase (*trpF*), indole-3-

Table 6. Gene repertoire of the serine biosynthesis/salvage pathway

Code	*serA*	*serC*	*serB*	*gapA*	*pgk*
Ag	AF0813 AF1779	—	AF2031 AF2138	AF1732	AF1146
AP	APE1265 APE2507	—	APE0683	APE0171	APE0173 MTH1042 MTH1835
TH	MTH970	—	MTH1626	MTH1009	MTH1883
PO	PAB1008 PAB0514	—	PAB0081 PAB1207	PAB0257	PAB2253 PAB1679
PF	RPF00348 RPF01832	—	RPF00004 RPF01259	G3P_PYRWO	PGR_PYRWO tnG3288813
PH	PH0520 PH1387	—	PH0134 PH1885	PH1830	PH0149 PH1218 MJ0641
MJ	MJ1018	—	MJ1594	MJ1146	MJ1482
AA	ser A_AQUAE	—	—	gap_AQUAE	pgk_AQUAE
TM	TM0327 TM1401	—	—	TM0688	TM0689
DR	DR1291 DR1701	—	—	DR1343	DR1342
EC	b1033 serA_ECOLI	serC_ECOLI	serB_ECOLI	gapA_ECOLI gapC1_ECOLI gapC2_ECOLI	pgk_ECOLI
YP	RYP03500 RYP04138	RYP02528	RYP01285 RYP03514	RYP01996	RYP03963
HI	HI0465	HI1167	HI1033	RHI21861	HI0525
PA	serA_PSEAE	RPA05704	RPA01433 RPA02232 RPA05146 RPA06693 RPA07563 RPA08558	RPA01915 RPA08368	RPA00160
NG	—	serC_NEIGO	serB_NEIGO RNG00998	RNG01744 RNG01806	RNG00286
NM	—	serC_NEIME	serB_NEIME RNM01047	RNM01085 RNM01234	RNM01075
RC	RRC02911 RRC04234	RRC04065	—	RRC02235 RRC02393 RRC02393 RRC04298 RRC04318	RRC03491
HP	HP0397	—	HP0652	HP0921 HP1346	HP1345
CJ	serA_CAMJE RCJ02367	serC_CAMJE	serB_CAMJE RCJ00166 RCJ00710	RCJ00653	RCJ02950 RCJ02951
CY	slr2123 sll1908	—	—	slr0884 sll1342	slr0394
CQ	—	—	—	gapA_CHLPN	pgk_CHLPN
CT	—	—	—	gapA_CHLTR	pgk_CHLTR
BB	—	—	—	BB0057	BB0056
TP	—	—	—	TP0844	TP0538
CA	RCA00861 RCA00983	—	19937502_F2_3	24241712_F3_107	24667260_F1_35
ML	RML02403	RML02560	RML02458	RML02456	RML00090
MT	serA_MYCTU Rv0728c	serC_MYCTU	serB_MYCTU serB2_MYCTU Rv3661	gap_MYCTU	pgk_MYCTU
MG	—	—	—	MG301	MG300
MP	—	—	—	MP410	MP411
PN	—	—	—	RPN01440	RPN00835
ST	—	—	—	RST00038	RST00513
EF	REF00453 REF00824	—	—	REF01512 REF02761	REF02762

Table 6. Continued

Code	*serA*	*serC*	*serB*	*gapA*	*pgk*
BS	serA_BACSU yoaD_BACSU yvcT_BACSU	serC_BACSU	— —	gap_BACSU gapB_BACSU	pgk_BACSU
SC	sp|40510 YER081W	ser1_YEAST	ser2_YEAST	tdh1_YEAST tdh2_YEAST tdh3_YEAST	pgk1_YEAST
CE	C31C9.2	F26H9.5		gpd1_CAEEL gpd2_CAEEL gpd3_CAEEL gpd4_CAEEL	T03F1.3

glycerol phosphate synthase (*trpC*), and tryptophan synthase α and β chain (*trpA* and *trpB*).

The gene repertoire per organism for the serine biosynthesis and the serine salvage pathway is shown in Table 6. Genes functioning in the tryptophan biosynthesis pathway are presented in Table 7. The serine salvage pathway, with genes *gapA* and *pgk*, is present in almost all bacteria, in eukaryotes *C. elegans*, and yeast as well as in archaea. The universal presence of the salvage pathway in organisms of all three domains is due to the abundance of the substrate glyceraldehyde-3-phosphate, and cannot be explained by the degradation of serine alone. For example, none of the obligatory pathogens, such as *B. burgdorferi, T. pallidum,* and *Mycoplasmaceae*, possess the capability to synthesize serine, although, all such organisms are able to convert glyceraldehyde-3-phosphate to 3-phospho-glycerate. Glyceraldehyde-3-phosphate is product of a variety of pathways and has to be recycled by the cell metabolisms for further use.

The tryptophan biosynthetic pathway depends critically on the pathways of serine. Thus, organisms that are capable of synthesizing tryptophan also possess the capability to synthesize serine. Table 6 lists genes coding for the serine biosynthesis pathway (*serA, serC, serB*) and the serine salvage pathway (*gapA, pgk*). One of the observed exceptions is the nematode *C. elegans*, similar to all animals it cannot synthesize tryptophan although the serine biosynthesis is intact. On the other hand, *S. pneumoniae* possesses a complete set of *trp*-genes without sequence similarity evidence that a serine biosynthesis pathway is present. Experimental evidence exists that archaea use the standard phosphorylating pathway to synthesize serine (Stauffer 1983, Metcalf et al. 1996), although we were unable to identify genes by sequence similarity that code for *serC*. The first seven entries in Table 6 list archaea with unidentifiable *serC* genes. Neither serine nor tryptophan biosynthesis are performed in *B. burgdorferi, T. pallidum, Chlamydiae,* and *Mycoplasmaceae*.

Dandekar et al. (1998), as well as Xie and Jensen (personal communication), observed conserved operon organization of genes relevant for the tryptophan biosynthesis pathway, the chorismate and serine synthesis. For example, *B. subtilis* possesses a super-operon including *aroG-aroB-aroH-trpE-trpD-trpC-trpF-trpB-trpA-hisH*$_o$-*tyrA*$_p$-*aroF* (Xie et al. 1999). A co-analysis of operon organization and of the phylogeny of the tryptophan biosynthesis pathway reveals an intriguing similarity between operon conservation and close pathway distance, rather than close similarity between operon conservation and the phylogeny based on the 16S rRNA tree. Fig. 7 compares the phylogeny of the tryptophan biosynthesis pathway with a dendrogram based on 16S rRNA. We use the comparison between the dendrogram based on the highly conserved 16S rRNA with pathway phylogenies and operon organization to detect evolutionary events that are not conform with the overall organismic evolution suggested by the 16S rRNA phylogeny. Whenever differences between the 16S rRNA dendrogram and pathway phylogeny/operon organization are observed, independent adaption to a similar environment or horizontal transfer of genes, operon parts of complete operons may have occurred. For example, Crawford (1989) observed that tryptophan-pathway genes in *P. aeruginosa* are scattered into three widely spaced groups, rather than coexisting within one operon, as they are in the close relative, *E. coli*. Based on the 16S rRNA tree, *E. coli* is closely related to *Y. pestis, H. influenzae,* and *P. aeruginosa* as shown in Fig. 7b. On the other hand, in terms of pathways, *P. aeruginosa* is closely related to *R. capsulatus* as shown in Fig. 7a (clade I) with similar operon organization (*trpE-trpD-trpC . . . trpF-trpB*). Closely related *E. coli, H. influenzae,* and *Y. pestis,* based on the 16S rRNA tree, exhibit very similar pathways. *H. pylori,* distantly related to the former organisms based on the 16S rRNA tree, joins the group in the pathway phylogeny (Fig. 7a, clade III) with a common operon organization showing a gene-fusion between *trpC* and *trpF* (*trpE-trpG-trpD-trpC/F-trpB-trpA*).

Another example for a difference between pathway phylogeny and 16S rRNA tree is observed between archaea and bacteria. *M. thermoautotropicum* (operon: *trpE-trpG-trpC-trpF-trpB-trpA-trpD*) shows a pathway,

as well as operon structure, that is close to that of *T. maritima* and *C. acetobutylicum* (*trpE-trpG-trpD-trpC-trpF-trpB-trpA*) as shown in Fig. 7a (clade II). Only *trpD* changed place during evolution between *M. thermoautotropicum* on the one hand, and *C. acetobutylicum, T. maritima,* on the other hand. At comparison of operon structures of trp-genes for organisms in clade II, and of those in clade III, suggests a gene-fusion event between *trpC* and *trpF* genes. Non-fused *trpC* and *trpF* genes in clade II, involving *C. acetobutylicum, M. thermoautotropicum,* and *T. maritima,* have been fused during evolution, and are exhibited as fusion genes *trpC/F* in *E. coli, H. influenzae, H. pylori,* and *Y. pestis,* in clade III. The gene fusion occurs between the gram-positive bacterium *C. acetobutylicum,* the thermophile bacterium *T. maritima,* the archaeon *M. thermoautotropicum* (clade II), and gram-negative bacteria *E. coli, H. influenzae, H. pylori, Y. pestis* (clade III). Despite this gene-fusion, the overall operon organization for organisms in clades II and III is identical.

The evidence of operon conservation and close pathway similarity between thermophile archaea and *T. maritima* implies either a common thermophile ancestor, or adaption of the complete pathway by horizontal transfer of the complete operon. Within the archaean domain, operon conservation, as well as close pathway distance between *Archaeoglobus* and *P. furiosus,* is exhibited in Fig. 7a (clade IV), although, based on the 16S rRNA tree, *Archaeoglobus* is more closely related to *M. thermoautotropicum* than to *P. furiosus.* The conserved operon between *Archaeoglobus* and *P. furiosus* involves *trpC-trpD-trpE-trpG-trpF-trpB-trpA.* Genes *trpC* and *trpD* are fused in *Archaeoglobus.*

In contrast to the archaeon (*M. thermoautotropicum*), the thermophile bacterium (*T. maritima*), gram-positive (*C. acetobutylicum*) and gram-negative bacteria (*E. coli, H. influenzae, H. pylori,* and *Y. pestis*) that are distantly related, according to the 16S rRNA tree, the operon structure is highly conserved, and the pathways are very similar between these organisms. These findings indicate that these organisms share common ancestry of the complete pathway.

Exceptions of operon conservation are, e.g., *A. aeolicus,* cyanobacteria, and *Neisseria*. In these organisms, tryptophan-pathway genes are scattered all over the respective genomes (not shown). Based on pathway simi-

Table 7. Gene repertoire of the tryptophan biosynthesis pathway

pID	*trpE*	*trpG*	*trpD*	*trpF*	*trpC*	*trpA*	*trpB*
aa0	trpE_AQUAE	trpG_AQUAE	trpD1_AQUAE	trpF_AQUAE	trpC_AQUAE	trpA_AQUAE	trpB1_AQUAE
aa1	trpE_AQUAE	trpG_AQUAE	trpD2_AQUAE	trpF_AQUAE	trpC_AQUAE	trpA_AQUAE	trpB1_AQUAE
aa2	trpE_AQUAE	trpG_AQUAE	trpD1_AQUAE	trpF_AQUAE	trpC_AQUAE	trpA_AQUAE	trpB2_AQUAE
aa3	trpE_AQUAE	trpG_AQUAE	trpD2_AQUAE	trpF_AQUAE	trpC_AQUAE	trpA_AQUAE	trpB2_AQUAE
ag	AF1603	AF1602	AF1604	AF1601	AF1604	AF1599	AF1600
ap	APE2553	APE2555	APE2551	APE2547	APE2546	APE2550	APE2316
bs	trpE_BACSU	—	trpD_BACSU	trpF_BACSU	trpC_BACSU	trpA_BACSU	trpB_BACSU
ca	29563962_F1_2	34272142_C3_23	5866093_F2_4	34011567_F3_9	25412825_F2_5	34615937_F3_10	34267202_F2_6
cj	trpE_CAMJE	trpD_CAMJE	trpD_CAMJE	RCJ02329	trpC_CAMJE	trpA_CAMJE	RCJ02330
ct	—	—	—	CT327	—	CT171	CT170
cy	slr0738	sll1634	slr1867	sll0356	slr0546	slr0966	—
dr	DR1791	DR0196	DR1767	DR0123	DR1426	DR0942	DR0941
ec	trpE_ECOLI	trpD_ECOLI	ybiB_ECOLI	trpF_ECOLI	trpC_ECOLI	trpA_ECOLI	trpB_ECOLI
ef	—	REF02186	—	—	—	—	—
hi	HI1388	HI1387	HI1389	HI1389.1	HI1389.1	HI1432	HI1431
hp	HP1281	HP1282	trpD_HELPY	HP1279	HP1279	HP1277	HP1278
mj	MJ1075	MJ0238	MJ0234	MJ0451	MJ0918	MJ1038	MJ1037
ml	RML02925	—	RML00513	—	RML02171	RML03457	RML02172
mt	trpE_MYCTU	Rv2859c	trpD_MYCTU	—	RMT02112	trpA_MYCTU	trpB_MYCTU
ng0	trpE_NEIGO	—	trpD_NEIGO	RNG01697	trpC_NEIGO	trpA_NEIGO	trpB_NEIGO
ng1	trpE_NEIGO	—	RNG00967	RNG01697	trpC_NEIGO	RNG00643	trpB_NEIGO
nm	trpE_NEIME	—	trpD_NEIME	RNM01528	trpC_NEIME	trpA_NEINM	trpB_NEINM
pa	sp|P09785	RPA06418	sp|P20574	sp|Q59649	sp|P20577	trpA_PSEAE	trpB_PSEAE
pf	RPF01392	RPF01393	trpD_PYRFU	RPF01394	trpC_PYRFU	trpA_PYRFU	trpA_PYRFU
ph	—	—	—	—	—	—	PH1583
pn	trpE_STRPN	RPN00767	trpD_STRPN	RPN01507	trpC_STRPN	trpA_STRPN	trpB_STRPN
rc	RRC01412	RRC01414	RRC01415	RRC01787	RRC01416	RRC00192	RRC01786
rp	—	RRP00395	—	—	—	—	—
sc	trp2_YEAST	trp3_YEAST	trp4_YEAST	trp1_YEAST	trp3_YEAST	trp5_YEAST	trp5_YEAST
ss	trpE_SULSO	trpG_SULSO	trpD_SULSO	trpF_SULSO	trpC_SULSO	trpA_SULSO	trpB_SULSO
st	—	RST00504	—	—	—	—	—
th	MTH1655	MTH1656	MTH1661	MTH1658	RTH00813	MTH1660	MTH1659
tm	TM0142	TM0141	TM0141	TM0139	RTM01443	TM0137	TM0138
yp	RYP02572	RYP02570	RYP02266	RYP2569	RYP02569	RYP00647	RYP03957

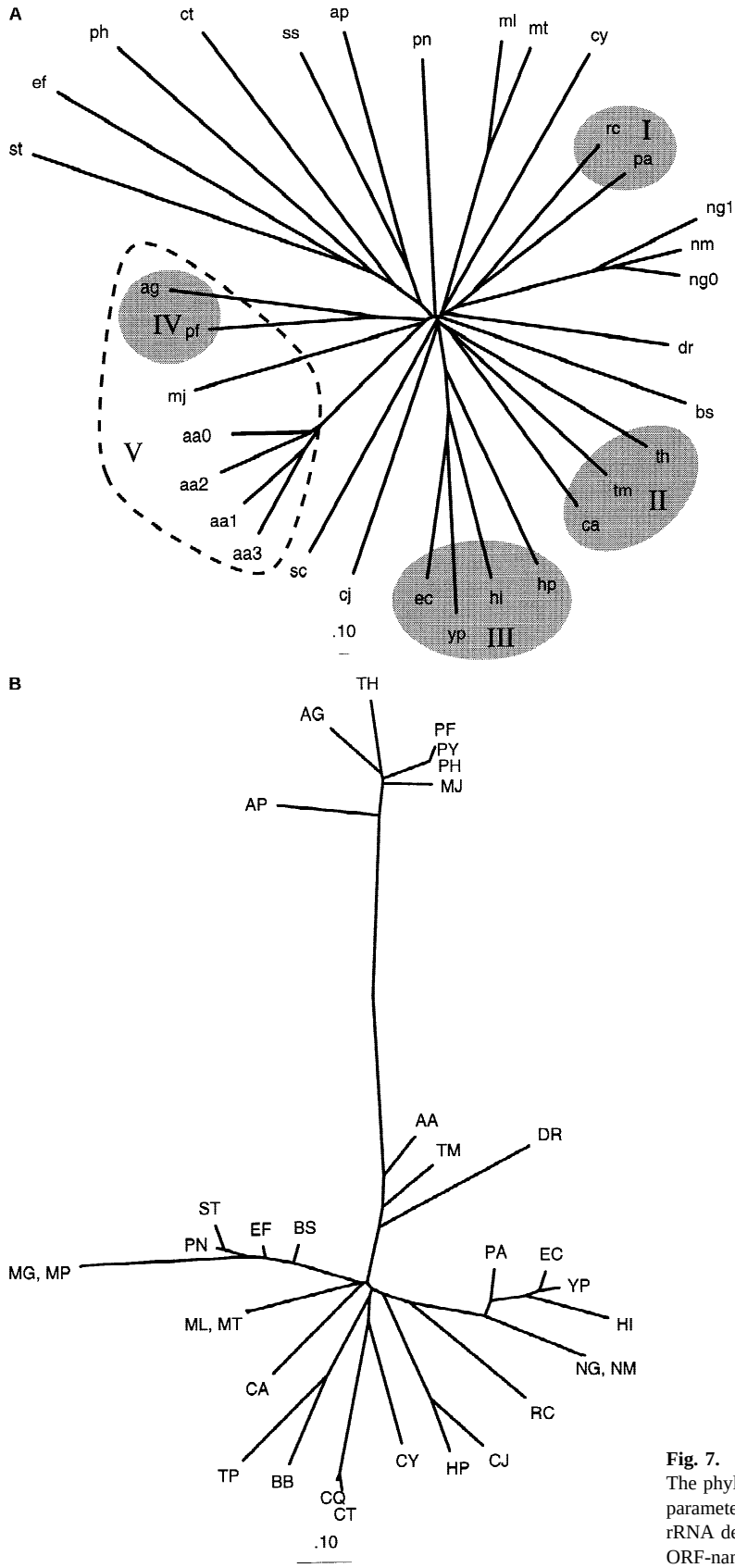

Fig. 7. **(A)** Tryptophan biosynthesis pathway. The phylogenetic tree is computed with parameters $f = 1$ and $t = 0.001$. **(B)** 16S rRNA dendrogram. References from pIDs to ORF-names are shown in Table 7.

larities, *A. aeolicus* is closely related to *Archaeoglobus, M. jannaschii,* and *P. furiosus,* rather than to *T. maritima.* The lack of operon conservation in the case of *A. aeolicus,* combined with the close relationship of the pathway to the archaean clade (clade V), suggests a convergent evolution of pathway genes adapting to a thermophilic environment, rather than a common ancestor with archaea or horizontal transfer of *trp*-genes from archaea. An equivalent solution that explains the lack of operon conservation in *A. aeolicus* would be the substitution of pathway regulation in *A. aeolicus* by operons with a co-regulatory mechanism based on the kinetics of the metabolic reaction (Bagheri-Chaichian and Wagner 2001). The relationships between *Neisseria* and other organisms reflect the relationship deduced from the 16S rRNA tree. The *trp*-pathway of *Neisseria* is similar to the trp-pathway of *Rhodobacter* and *Pseudomonas.*

A comparison of the phylogeny of the tryptophan/serine biosynthesis pathway with phylogenies of individual genes that contribute to this pathway (not shown) indicates a hierarchical mechanism of metabolic pathway evolution. For example, enzymes that use chorismate as substrate, such as anthranilate (*o*-amino-benzoate) synthase (trpE/G) in the tryptophan biosynthesis pathway and PABA (*p*-amino-benzoate) synthase, are paralogs, and thus have a common ancestral gene that possibly was able to synthesize both ortho- and para-amino-benzoate. On the other hand, the established tryptophan/serine biosynthesis pathway, with genes organized in a single operon, tends to be passed to offspring as a conserved unit.

Conclusion

We have developed a method for the comparison of metabolic pathways based on explicit sequence information. To illustrate the method, four metabolic networks have been analyzed: (1) the ferredoxin–NADPH reductase pathway, (2) pathways utilizing ferredoxin, (3) terminal oxidase complexes, (4) tryptophan/serine biosynthesis networks. Woese (1998b) states that metabolic genes are among the most modular in the cell, and that these genes are expected to travel laterally, even today. Evidence for both adaptations of single genes and horizontal transfer of complete pathways between organisms is seen in our phylogenetic analysis.

The analysis of the evolution of terminal oxidases, and the comparison of the results with a study performed by Musser and Chan (1998) serve as a validation for the method. Even with different numbers of functional roles due to differences in the organization of the terminal oxidase superclusters between organisms, the presented method constructs a phylogeny of quinol and cytochrome oxidase complexes in agreement with the results by Musser and Chan.

By computing distances of pathways with different weighting parameters, and by comparing the yielded phylogenies, one can draw conclusions about the robustness or versatility of relationships between pathway representations. Such a comparison has been provided in case of the ferredoxin–NADPH reductase pathway. Pathway representations of *Neisseria* and *P. aeruginosa* exhibit a non-conserved relationship between PRs and, thus, suggest a frequent and random exchange of pathways between both species.

We suggest a similarity between organisms of different domains in the special case of the thermophilic bacterium. *A. aeolicus* and archaea. Underlying scenarios are either continuous acquisition of thermotolerant genes from preadapted hyperthermophiles or convergent evolution by adaptation to an extreme environment. More genomes of both extremophilic archaea and thermophilic bacteria are necessary to provide evidence for either scenario.

The analysis of mixed-function supraoperons exhibit gene relationships with one another that are not as obvious as those encoding steps of linear pathways. The tryptophan-biosynthesis pathway together with the serine biosynthesis pathways and the serine salvage pathway represent a branched and interconnected metabolic network. This interconnectivity manifests itself in the observed mixed-function supraoperons that contain genes of serine, tryptophan, and aromatic amino acid biosynthesis. A majority of the studied organisms do possess such mixed-function supraoperons with conserved operon structures. The phylogenetic analysis of the interconnected tryptophan/serine-pathway displays relationships according to operon conservation that differ from the relationships revealed by 16S rRNA phylogeny. Although the coregulation of genes operating in a distinct pathway is a plausible explanation for operon conservation, it does not explain the dispersion and scattering of genes in groups as well as the total lack of operon conservation in *A. aeolicus,* cyanobacteria, and *Neisseria.* In these cases pathway-genes are suggested to be coregulated by the dynamics of regulatory networks between individual genes. Their evolutionary dominance may be caused by maximizing gene interactions of individual genes functioning in the pathway (Bagheri-Chaichian and Wagner 2001).

The phylogenetic analysis of individual genes shows that independent gene duplication is a plausible evolutionary process to initiate a metabolic pathway. The tryptophan biosynthesis pathway with trp E/G genes (anthranilate synthase) and its paralog PABA (p-amino-benzoate) synthase is an example for a hierarchical evolution of metabolic networks. After the functional pathway of tryptophan biosynthesis was established and organized in an operon, this operon was inherited by offspring as a conserved unit, or was reorganized by dynamic operon shuffling, gene fusion, or loss of genes by translocation.

The major advantage of the phylogenetic analysis of metabolic networks resides in the combined analysis of more than one functional role. The analysis is understood as an extension of the classic phylogenetic analysis of individual sequences towards a higher level of description. Pathway phylogenies classify relationships between genes, but also between pathways and multi-enzyme systems. With the advent of gene expression analysis, future studies will combine investigations of the relation and evolution of larger metabolic networks with gene regulatory networks.

The comparison and phylogenetic analysis of metabolic networks may also be useful for gene-diagnostics and gene-therapy that are currently based on comparative genomics. With the comparison of metabolic pathways, complex relationships between genes can be detected and more sophisticated directions for the cure of complex diseases may become feasible. For example, comparative genomics of metabolic networks may help researchers to find treatments for parasitic diseases such as gingivatis, gonorrhea, or malaria. By comparing the implied changes in the metabolisms of infected human cells as well as of parasites during infection on both the level of the metabolic network and the genome by monitoring changes in gene expression of enzymes, one may "decipher" the parasite–host system between human and the causative agent, *N. gonorrhoeae, P. gingivalis,* or *Plasmodium falciparum.* The knowledge gained from these studies will be used to disrupt key events in pathogenesis, in order to ameliorate the consequences of exposure to pathogenic organisms and to aid in the development of effective vaccines and small molecule therapeutics.

Acknowledgments. Fruitful discussions with Roy Jensen from the University of Florida, Ross Overbeek from the WIT-team at Integrated Genomics Inc. and Gary Xie from the Los Alamos National Laboratory are gratefully acknowledged. Special thanks are directed to Zaida Luthey-Schulten from the University of Illinois who initiated the analysis of the tryptophan/serine network by suggesting to study aromatic aminoacid biosynthesis pathways. We also thank Michael Wall for carefully reading the manuscript and the anonymous referees for suggestions. This work has been supported by grants from the National Institute of Health (NIH PHS 5 P41 RR05969) and the National Science Foundation (NSF BIR 94-23827 EQ).

Appendix

In the following we define concepts and expressions which are used throughout the manuscript.

Metabolic networks and pathways: A *metabolic network* is a directed reaction graph with substrates as vertices and directed, labeled edges denoting reactions between substrates catalyzed by enzymes (labels). A *metabolic pathway* is a special case of a metabolic network with distinct start and end points, initial, and terminal vertices, respectively, and a unique path between them.

Functional role. A *functional role* refers to a gene product and how this product is embedded in a metabolic network, i.e., what task it has to perform. Typical functional roles are *enzymes* which process substrates in a specific reaction or *substrates* which are processed by specific enzymes. A functional role also describes how a gene product functions in a protein complex.

Representation of a pathway. A representation of a pathway is a unique set of genes, one gene for each function in the corresponding pathway. For example, for a simple pathway with one enzyme processing one function and a genome of an organism that has two genes coding for the substrate (a, b) and two genes coding for the enzyme (E, F), a total of four representations exist for this hypothetical pathway (aE, bE, aF, bF). We refer to "representation of the pathway" as *pathway representation* or short *PR.*

Homogeneous and heterogeneous clades of pathway representation. A classification in phylogenetic trees is made according to distances between each of two pathway representations (PRs). *PR*s can be grouped into *clades.* A *clade* is defined as a set of representations with a minimal distance between any two members of this set with respect to the phylogenetic tree. Thus, on average, the distance between members of the clade is smaller than between members and non-members. A *homogeneous clade* is defined as a clade with pathway representations of a single organism. The maximum distance in a homogeneous clade is the maximal possible distance between *PR*s of the same organisms. By exceeding this maximum distance the next closest *PR* will be from an organism different from the organism in the clade. *A heterogeneous clade* is a clade with *PR*s from different organisms.

Ortholog and paralog pathway representations. Two pathway representations are defined *ortholog* to each other if all gene pairs (one gene in each *PR* which code for the same function) are orthologs. Two pathway representations are defined *paralog* to each other if at least one gene pair is paralog.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Andersson S, Zomordipour A, Andersson J, Sicheritz-Ponten T, et al. (1998) The genome sequence of *rickettsia provazekii* and the origin of mitochondria. Nature 396:133–140

Bagheri-Chaichian H, Wagner GP (2001) Mutational robustness of multi-enzyme pathways in regimes of maximal gene interaction. (submitted)

Blattner F, Plunkett G, Bloch C, Perna N, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1435–1474

Bult C, White O, Olsen G, Zhou L, et al. (1996) Complete genomic

sequence of the *methanogenic archaeon, Methanococcus jannaschii.* Science 273:1058–1073

Chatton E (1938) Titres et Travaux Scientifiques (1906–1937). Sottano, Sète, France

Cole S, Brosch R, Parkhill J, Garnier T, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537–544

Crawford IP (1989) Evolution of a biosynthetic pathway. Ann Rev Microbiol 43:567–600

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23(9):324–328

Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in proteins. In: Dayhoff M (ed), Atlas of Protein Sequence and Structure, Vol. 5. National Biomedical Research Foundation, Washington, D.C. pp 345–352

Deckert G, Warren P, Gaasterland T, Young W, et al. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus.* Nature 392:353–358

Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods in Enzymology 266:418–427

Fitch W (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113

Fitch W, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279–284

Fleischmann R, Adams M, White O, Clayton R, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser C, Casjens S, Huang W, Sutton G, et al. (1997) Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi.* Nature 390:580–586

Fraser C, Gocayne J, White O, Adams M, et al. (1995) The minimal gene complement of *Mycoplasma genitalium.* Science 270:397–403

Fraser C, Norris S, Weinstock G, White O, et al. (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281:375–388

Goffeau A, Aert R, Agostini-Carbone M, Ahmed A, et al. (1997) The yeast genome directory. Nature 387:5–105 (Suppl)

Haldane J (1928) The origin of life. Rationalist Ann 148:3–10

Hartman H (1975) Speculations on the origin and evolution. J Mol Evol 4:359–370

Henikoff S, Henikoff J (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89:10915–10919

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li B, Herrmann R (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.* Nucl Acids Res 24:4420–4449

Huson DH (1998) Splitstree: analyzing and visualizing evolutionary data. Bioinformatics 14:68–73

Kalman S, Mitchell W, Marathe R, Lammel C, et al. (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis.* Nat Genet 21(4):385–389

Kaneko T, Sato S, Kotani H, Tanaka A, et al. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp.* strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3: 109–136

Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, et al. (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res 6(2):83–101

Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, et al. (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res 5:55–76

Klenk H, Clayton R, Tomb J, White O, et al. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus.* Nature 390:365–370

Kunst F, Ogasawara N, Moszer I, Albertini A, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis.* Nature 390:249–256

Lipmann F (1965) THe origin of prebiological systems and of their molecular matrices. Academic Press, New York, NY pp 259–280

Mayr E (1998) Two empires or three? Proc Natl Acad Sci USA 95: 9720–9723

Metcalf WW, Zhang J-K, Shi X, Wolfe RS (1996) Molecular, genetic, and biochemical characterization of the *serc* gene of *methanosarcina barkeri* fusaro. J Bact 178(19)

Miller SL (1953) A production of amino acids under possible primitive earth conditions. Science 117:528–529

Musser SM, Chan SI (1998) Evolution of the cytochrome *c* oxidase proton pump. J Mol Evol 46:508–520

Oparin AI (1967) The origin of life. In: Bernal J (ed) The Origin of Life, World, Cleveland, OH. Originally published in: Proiskhozhdenie Zhiznyl (1924) IZD Moskovishii Rabochii, Moscow

Orgel LE (1968) Evolution of the genetic apparatus. J Mol Biol 38: 381–383

Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E, Jr., Kyrpides N, Fonstein M, Maltsev N, Selkov E (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nuc Acids Res 28(1):123–125

Overbeek R, Larsen N, Smith W, Maltsev N, Selkov E (1997) Representation of function: the next step. Gene 191:GC1–GC9

Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276:734–740

Parkhill J, Achtman M, James K, Bentley S, et al. (2000a) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature 404:502–506

Parkhill J, Wren B, Mungall K, Ketley J, et al. (2000b) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 403:665–668

Popper KR (1957) The Poverty of Historicism. Routledge and Kegan Paul, London

Popper KR (1963) Conjectures and Refutations. Routledge and Kegan Paul, London

Smith D, Doucette-Stamm LC, Deloughery CHL, et al. (1997) Complete genome sequence of *methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J Bacteriolog 179:7135–7155

Stauffer GV (1983) Regulation of serine, glycine, and one-carbon biosynthesis. In: Herrman K, Sommerville RL (eds), Amino Acids: biosynthesis and genetic regulation. Addison-Wesley Publishing Co, Reading, MA pp 103–113

Stephens R, Kalman S, Lammel C, Fan J, et al. (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis.* Science 282:754–759

Tatusov R, Koonin E, Lipman D (1997) A genomic perspective on protein families. Science 278:631–637

Tatusov R, Mushegian A, Bork P, Brown N, Hayes W, Borodovsky M, Rudd K, EV (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli.* Curr Biol 6:297–291

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282:2012–2018

Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. Nucl Acids Res 22:4673–4680

Tomb JF, White O, Kerlavage A, Clayton R, et al. (1997) The complete

genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388:539–547

Wächtershäuser G (1990) Evolution of the first metabolic cycles. Proc Natl Acad Sci USA 87:200–204

White O, Eisen J, Heidelberg J, Hickey E, et al. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiondurans* R1. Science 286:1571–1577

Woese C (1998a) Default taxonomy: Ernst Mayr's view of the microbial world. Proc Natl Acad Sci USA 95:11043–11046

Woese C (1998b) The universal ancestor. Proc Natl Acad Sci USA 95:6854–6859

Woese CR (1982) Archaebacteria and cellular origins: an overview. Zentralbl Bakteriolog Mikrobiol Hyg Ser C 3:1–17

Xie G, Bonner CA, Jensen RA (1999) A probable mixed-function supraoperon in *pseudomonas* exhibits gene organization features of both intergenomic conservation and gene shuffling. J Mol Evol 49:108–121

Xie G, Jensen R.A. Personal communication