JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 6, Numbers 3/4, 1999
Mary Ann Liebert, Inc.
Pp. 343–360

Evolution of Metabolisms: A New Method for the Comparison of Metabolic Pathways Using Genomics Information

CHRISTIAN V. FORST and KLAUS SCHULTEN

ABSTRACT

The abundance of information provided by completely sequenced genomes defines a starting point for new insights in the multilevel organization of organisms and their evolution. At the lowest level enzymes and other protein complexes are formed by aggregating multiple polypeptides. At a higher level enzymes group conceptually into metabolic pathways as part of a dynamic information-processing system, and substrates are processed by enzymes yielding other substrates. A method based on a combination of sequence information with graph topology of the underlying pathway is presented. With this approach pathways of different organisms are related to each other by phylogenetic analysis, extending conventional phylogenetic analysis of individual enzymes. The new method is applied to pathways related to electron transfer and to the Krebs citric acid cycle. In addition to providing a more comprehensive understanding of similarities and differences between organisms, this method indicates different evolutionary rates between substrates and enzymes.

Key words: metabolic networks, phylogeny, electron transfer, citric acid cycle, microbial genomes.

1. INTRODUCTION

METABOLISMS OF LIVING SYSTEMS and their evolution have been investigated for long time. First studies had been performed in the late 1950s and early 1960s by Popper [38, 39] and Lipmann [29], and had been followed by other scientists. This extensive research was motivated by questions regarding the origin of life and the evolution of the biosphere. Seminal contributions by Haldane [18], Miller [31], Oparin [33], and Orgel [34] are mentioned in this context, discussing the (prebiotic) chemical environment suitable for a biotic evolution. Based on these discussions hypotheses on the origin and evolution of metabolism started [19] and questions regarding the emergence of the first metabolic cycles were formulated [57]. An example for a cyclic metabolic network[1] is the Krebs citric acid cycle in the respiratory chain. The above studies address questions regarding energetics of chemical reaction and the implied possible existence of such reactions in living systems.

We suggest a synthesis between representation and comparison of "abstract" metabolic networks and individual enzymes and substrates. This approach has already been applied to the malate–aspartate shuttle pathway and to terminal oxidase supercomplexes [11]. In this study, web-based information systems are

Theoretical Biophysics, Beckman Institute, University of Illinois, Urbana-Champaign, Urbana, Illinois.

[1]For a definition see appendix.

presented in Section 2. Section 3 introduces a method for calculating distances between metabolic networks by using genomics information. Section 4 describes applications of the method to some electron transfer pathways and to the Krebs citric acid cycle.

2. DATABASES

The analysis of metabolic networks based on sequence information of enzymes and substrates requires access to adequate databases. One example of such databases is the Kyoto Encyclopedia of Genes and Genomes (KEGG) [32] which provides both an online map of metabolic pathways and the ability to focus on metabolic reactions in specific organisms. Karp's and Riley's EcoCyc [24, 25] is essentially specific to *E. coli*, and includes detailed information about known metabolic networks of *E. coli* and the reactions they embody.

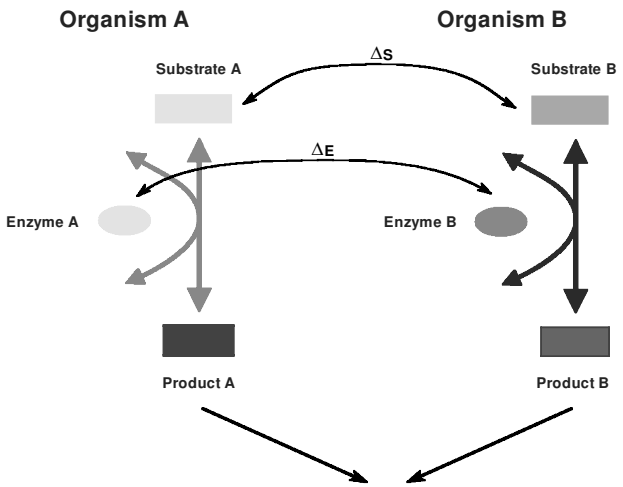
A convenient database to obtain genomic and organizational information of completely sequenced microbial organisms is the *WIT-system* by Overbeek *et al.* [36]. This system provides researchers with DNA and protein sequence informations of complete or partially sequenced genomes. The information is annotated with organizational information, gene and operon organization, and information on metabolic networks. Using WIT, researchers are able to perform a so-called metabolic reconstruction of microbial genomes [35].

Currently, 38 genomes of microbial origin and one of multicellular origin (*C. elegans*) are accessible via the WIT-system. Of these genomes, 17 are completely sequenced; the remaining ones are involved in ongoing sequencing projects and have to be accessed with care. An overview is shown in Table 1.

3. DISTANCES BETWEEN METABOLIC NETWORKS

A common approach to deduce a relationship between individual biopolymers is to align sequences to each other and measure distances, e.g., by using BLOSUM [20] and PAM [5] similarity matrices. In this paper an extension toward a relationship based on metabolic networks present in a living organism is made. For this purpose we combine sequence information of involved genes with topological information of the corresponding network.

A first example involves the simplest type of metabolic pathway, a substrate processed by an enzyme. The global distance Δ between such pathways is deduced by the individual distances between substrates ΔS and enzymes ΔE. Figure 1 illustrates schematically the comparison of such a simple pathway with two functional roles[1] ($n = 2$). A *global distance* Δ for n functional roles per pathway is defined as follows.



$$\Delta = (\Delta S^p + \Delta E^p) \cdot \left(\frac{1+f}{2}\right)^p,\ 0 \le f \le 1,\ 0 < p$$

FIG. 1. Calculating a distance between two pathways: individual distances ΔE, ΔS between sequences of the same functional role are used to calculate a global distance Δ.

TABLE 1. GENOMES STUDIED

Code	*Organism*	*D*[a]	*Size (kb)*[b]	*Number of ORFs*[c]	*Sequencing*[d]	*References*
AG	*Archaeoglobus fulgidus*	A	2178.40	2493	x	[27]
TH	*Methanobacterium thermoautotrophicum*	A	1751.38	1866	x	[41]
PH	*Pyrococcus horikoshii*	A	1738.51	1825	x	[26]
MJ	*Methanococcus jannaschii*	A	1739.93	1811	x	[3]
AA	*Aquifex aeolicus*	B	1590.78	1744	x	[6]
DR	*Deinococcus radiodurans*	B	3261.20	3771	o	[43]
EC	*Escherichia coli*	B	4639.22	4289	x	[2]
YP	*Yersinia pestis*	B	4501.71	4296	o	[50]
HI	*Haemophilus influenzae*	B	1830.14	1846	x	[10]
PA	*Pseudomonas aeruginosa*	B	6286.26	5642	o	[56]
NG	*Neisseria gonorrhoea*	B	2063.17	1853	o	[54]
NM	*Neisseria meningitidis*	B	2157.54	1838	o	[45]
RC	*Rhodobacter capsulatus* SB1003	B	2079.41	1989	o	[53]
HP	*Helicobacter pylori*	B	1667.88	1547	x	[52]
CJ	*Campylobacter jejuni*	B	1644.03	2106	o	[48]
CY	*Synechocystis* sp.	B	3573.47	3226	x	[23]
PG	*Porphyromonas gingivalis*	B	2447.62	1832	o	[47]
BB	*Borrelia burgdorferi*	B	1519.86	1666	x	[12]
TP	*Treponema pallidum*	B	1138.82	1031	x	[14]
CA	*Clostridium acetobutylicum*	B	4030.73	3967	o	[15]
ML	*Mycobacterium leprae*	B	2420.76	1568	o	[49]
MT	*Mycobacterium tuberculosis*	B	4411.53	3924	x	[4]
MG	*Mycoplasma genitalium*	B	580.07	532	x	[13]
MP	*Mycoplasma pneumoniae*	B	816.39	674	x	[21]
PN	*Streptococcus pneumoniae*	B	2104.82	1844	o	[46]
ST	*Streptococcus pyogenes*	B	1799.24	1599	o	[55]
EF	*Enterococcus faecalis*	B	3209.12	2967	o	[44]
BS	*Bacillus subtilis*	B	4214.81	4093	x	[28]
SC	*Saccharomyces cerevisiae*	E	12057.28	6125	x	[17]
CE	*Caenorhabditis elegans*	E	165227.99	16332	x	[7]

[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.

Definition 1. *Let* Γ, Γ' *be metabolic paths of identical topology involving n functional roles* $I_i, I_i', i = 1, \ldots, n$. *Furthermore let* $\Delta X_i = \delta(I_i, I_i')$ *be a distance calculated by an alignment* δ. *Then a distance* Δ *between* Γ *and* Γ' *is defined as*

$$\Delta = \Phi \cdot \sum_{i=1}^{n} \Delta X_i^p, \qquad \Phi = \begin{cases} 1 & \text{for orthologs} \\ \left[\frac{1+f(n-1)}{n}\right]^p & \text{for paralogs} \end{cases} \tag{1}$$

where $0 \le f \le 1$ *and* $p > 0$.

Two parameters f and p are introduced with values $0 \le f \le 1$ and $p > 0$: altering f changes the distinction between paralog and ortholog pathways,[1] and p provides different weightings of long and short distances.

Paralog genes are differentiated from orthologs based on their (proposed) functional role in the genome, which differs from functions of orthologs. Orthologs are genes in different species that evolved from a common ancestral gene by specification; by contrast, paralogs are genes related by duplication within a genome [9]. Normally, orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if related to the original [42]. Parameter f in Equation (1) reflects this difference in homologous genes. For $f = 1$, orthologs and paralogs are treated the same; for $f = 0$, the total distance Δ between "paralog pathways" is actually the mean value of the individual distances ΔX_i.

3.1. Gaps

Addressing missing functional roles is a feasible approach to the investigation of pairs of pathways with different graph topology. If a distinct functional role I_k is missing in a pathway Γ then a gap penalty Δ_{gap} is assigned to the otherwise undefined distance $\Delta X_k = \Delta_{gap}$ [Equation (1)].

Definition 2. *A* confidence level *t for the minimum number of present functional roles is defined. For the number of functional roles that fall below this threshold value t a pathway is considered as incomparable.*

To define a distance between any two pairs of pathways the global distance Δ between an arbitrary pathway and an incomparable pathway is defined by a *penalty distance* $\Delta_p = \Delta$. Values for Δ_{gap} and Δ_p are chosen to reflect estimates for expected distances. Thus, Δ_{gap} is similar to average distances between individual functional roles. Δ_p is an estimate for a global distance between two pathways. It is typically of the order of magnitude of $n \times \Delta_{gap}$. In the present study the gap penalty has been set to $\Delta_{gap} = 1$ and Δ_p has been assigned a value according to the number of functional roles involved per investigated pathway.

4. ELECTRON TRANSFER

Electron transport pathways play a key role in the metabolism of a living cell. There are about 69[2] pathways known that are related to electron transfer. Table 2 previews a subset of 15 selected pathways out of the total 69. The confidence level (Definition 2) for pathways in Table 2 is $t = 1$. Four members of the set of 15 selected electron transfer pathways are present in organisms of all three domains:

Dihydrolipoamide–NAD^+ electron transport (pathway 2 in Table 2)
NADH–FAD electron transport (plasma membrane) (pathway 4 in Table 2)
NADH–NAD^+ electron transport (malate, aspartate) (pathway 10 in Table 2)
NADPH–oxidized thioredoxin electron transport (pathway 15 in Table 2)

The substrates utilized in these pathways also play functional roles in other metabolisms. For example, lipoamide dehydrogenase component (E3) (EC 1.8.1.4) in pathway 2 catalyzes reactions in carbohydrate metabolisms. NAD–dehydrogenase (EC 1.6.99.3) in pathway 4 is involved in many electron transport reactions. Aspartate aminotransferase (EC 2.6.1.1) and malate dehydrogenase (EC 1.1.1.37) in pathway 10 serve as reaction partners in amino acid and carbohydrate metabolisms. And last but not least, thioredoxin and its reductase (EC 1.6.4.5) in pathway 15 is found in a variety of metabolisms. Thioredoxin reductase catalyzes many electron transfer reactions (as two examples serve pathways 1 and 15). Thioredoxin itself is involved in 29 pathways related to amino acid electron transfer, protein, purine, pyrimidine, and sulfur metabolisms as almost universal redox reagens. This universality of thioredoxin and thioredoxin reductase is also reflected in its presence in almost all studied genomes (Table 1), except in *P. horikoshii*. This organism does not possess thioredoxin.

Other electron transfer pathways are present only in organisms of one or two domains. Specialized pathways involving mitochondria, of course, are found in eukaryotes only (pathways 5 and 9). A large number of pathways are assigned to bacteria only (pathways 1, 7, 8, 12, and 14). For example, the "pathological" mercury (II) reductase reaction (pathway 12) is utilized by four bacteria only. Surprisingly there are no electron transport pathways that are unique for archaea. This observation confirms the known fact that entire pathways are acquired or displaced in the case of the archaea. Actually, by investigating frequency distributions and conserved operons of functional roles between all three domains a tendency of horizontal gene transfer from bacteria to archaea can be detected. The reverse direction (from archaea to bacteria) is less probable.[3]

Pathways within the archaeal and bacterial domain show the following general properties. Genome size is related neither to the number of all pathways nor to the number of pathways related to electron transfer. Depending on the individual properties and habitat of a microorganism different pathways are utilized

[2]The exact number depends on the definition of the corresponding pathways. Typically, pathways that perform similar tasks but that reside in different parts of the cell—e.g., in the periplasma or in the mitochondria—are each counted as an individual pathway.

[3]N. Kyrpides, and C.V.F., unpublished data.

by the cell, e.g., *B. burgdorferi* has a genome that codes for almost as many genes as *H. influenzae*. In contrast to *H. influenzae*, which uses many electron transfer pathways, the highly parasitic *B. burgdorferi* possesses only NADH–FAD electron transport (pathway 4) and NADPH–oxidized thioredoxin electron transport (pathway 15).

Closely related organisms use similar pathways: each pair of *Neisseria* NG, NM, Mycobacteria (ML, MT), and Mycoplasmae (MG, MP) has the same set of pathways. Exceptions are Streptococcae (PN, ST) that are involved in ongoing sequencing projects. Thus, accessible genome data about both organisms are still preliminary and information regarding new genes may be missing. With respect to these preliminary data, *S. pyogenes* possesses a richer set of pathways than *S. pneumoniae*. A major cause of the lack of pathways for *S. pneumoniae* lies in missing glutathione reductase (EC 1.6.4.2) in *S. pneumoniae*; glutathione reductase is involved in seven electron transfer pathways. Compared to other species in the geni Mycobacteria and Mycoplasmae, *M. tuberculosis* utilizes unproportionally many pathways. This reflects the ability of the bacillus to adapt to environmental changes. Not only is *M. tuberculosis* capable of competing with the lung for oxygen, it also adapts to the microaerophilic/anaerobic environment at the heart of the burgeoning granuloma. *Escherichia coli*, one of the best studied microbial organisms has the most complete set of pathways. Closely related to this species are the pathogens *H. influenzae, Y. pestis*, and the metabolic very versatil *P. aeruginosa*. They possess a similar, but somewhat less abundance of electron transfer pathways compared to *E. coli*.

4.1. Ferredoxin, an important coenzyme

Ferredoxins, besides thioredoxin, flavodoxin, and rubredoxin, are important coenzymes in metabolic pathways. They serve as electron acceptors and donors in many anabolic, catabolic, and electron transfer reactions. For example, ferredoxin is a redox partner in more than 50 known pathways.

An example of ferredoxin-utilizing pathways is the reversible *ferredoxin–NADPH reductase pathway*, which can be found in Bacillaceae, Cyanobacteria, and Enterobacteriaceae. In this pathway ferredoxin is processed (either oxidized or reduced) by ferredoxin–NADPH reductase. Sequences for both functional roles are obtained from the WIT-system. A multiple sequence alignment for each set of sequences has been performed by ClustalW v1.74 [51] with the BLOSUM62 similarity parameter. Alignment parameter had been set to default values. The created Phylip distance matrices were then used for calculations of the pathway distance according to Equation (1). Parameters f and p were chosen such that the calculated global distance matrix (which yields the phylogenetic tree) has a minimum number of distance triples that violate the triangle inequality. Phylogenetic relationships were analyzed by phylogenetic graph reconstruction programs such as *SPLITSTREE2* [22] or the *PHYLIP* software suite [8]. The pathway and a nonrooted phylogenetic tree drawn by the *PHYLIP* software suite are shown in Figs. 2 and 3 respectively.

Each leaf of the phylogenetic tree (Fig. 3) displays the label that provides a unique pathway identification (pID); the labels refer to the name of the organism (defined in Table 1) and the combination of functional roles used. For example, *M. tuberculosis* (according to Table 1 with organism code MT) has three paralogs that code for ferredoxin (fdxA, fdxC, and fdxC) as well as two paralogs that code for ferredoxin reductase (fprA and YZ14). Using all possible combinations between ferredoxin and ferredoxin reductase yields six representations of pathways[1] (PRs) for *M. tuberculosis*. Table 3 shows pIDs as well as corresponding ORF

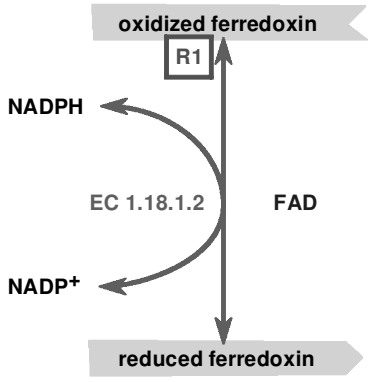


FIG. 2. Ferredoxin–NADPH reductase pathway. The pathway is shown with functional roles ferredoxin and ferredoxin–NADPH reductase (EC 1.18.1.2).

TABLE 2. EXAMPLES OF ELECTRON TRANSFER PATHWAYS

	Pathway	*Number*[a]	*Codes of organisms for which pathways are assigned*
1	2-Oxyglutarate–oxidized thioredoxin electron transport (via EC 1.2.4.2)	5	-- -- -- -- – DR EC YP HI PA NG NM -- -- -- -- -- -- -- -- -- -- MT -- -- -- -- -- BS -- --
2	Dihydrolipoamide–NAD^+ electron transport	1	AG TH -- MJ AA DR EC YP HI PA NG NM -- -- -- CY PG -- -- -- ML MT MG MP ST PN EF BS SC CE
		•[b]	
3	NADH–oxidized rubredoxin electron transport	2	AG TH -- MJ -- -- -- -- -- PA -- -- -- -- -- -- -- -- -- -- -- -- MT -- -- -- -- -- -- -- --
4	NADH–FAD electron transport (plasma membrane)	1	AG TH PH MJ AA DR EC YP HI PA NG NM RC HP CJ CY PG BB TP CA ML MT MG MP ST PA EF BS -- CE
5	$NADH–FAD_{\text{mitochondrial inner membrane}}$ electron transport (glycerol 3-phosphate)	2	-- SC CE
6	NADH–oxidized glutathione electron transport	1	-- -- -- -- AA -- EC YP HI PA -- -- RC -- -- -- -- -- -- -- -- MT -- -- ST -- -- EF SC CE
7	NADH, $H^+–O_2$, $H^+_{\text{periplasma}}$ electron transport (ubiqinone, cytochrome *bd*) (plasma membrane)	16	-- -- -- -- AA -- EC YP -- PA -- -- RC -- -- CY -- -- -- -- -- -- -- MT -- -- -- -- -- --
8	NADH, $H^+–O_2$, $H^+_{\text{periplasma}}$ electron transport (ubiqinone, cytochrome *bo*) (plasma membrane)	19	-- -- -- -- AA -- EC YP -- PA -- -- -- -- -- -- -- -- -- -- -- MT -- -- -- -- -- -- -- --

9	$NADH_{mitochond.\ matrix}$–$O_2$, mitochondrial matrix electron transport (mitochondrial inner membrane, mitochondrial intermembrane space)	47	-- SC CE
10	NADH–NAD^+ electron transport (malate, aspartate)	2 •[b]	AG TH PH MJ AA DR EC YP HI PA -- -- -- -- CJ CY PG -- -- CA ML MT -- -- -- -- EF BS SC CE
11	NADH–O_2 electron transport (plasma membrane)	1	AG TH -- -- AA DR EC YP HI PA NG NM RC HP CJ CY PG -- -- -- -- MT -- -- ST PN EF BS -- --
12	NADPH–Hg^{2+} electron transport	1	-- -- -- -- -- -- EC -- -- -- -- -- -- -- -- CY -- -- -- -- -- MT -- -- -- PN -- -- -- --
13	NADPH–oxidized flavodoxin electron transport (plasma membrane)	2	AG TH PH MJ AA -- EC YP HI PA -- -- RC HP CJ CY PG -- TP -- -- -- -- -- -- -- EF BS -- --
14	NADPH–oxidized ferredoxin electron transport	2	-- -- -- -- -- DR EC YP -- PA NG NM -- -- -- CY -- -- -- -- -- MT -- -- -- -- -- -- -- --
15	NADPH–oxidized thioredoxin electron transport	2	AG TH -- MJ AA DR EC YP HI PA NG NM RC HP CJ CY PG BB TP CA ML MT MG MP ST PN EF BS SC CE

[a]Number of functional roles per pathway.
[b]This pathway is located in mitochondria for eukaryotes.

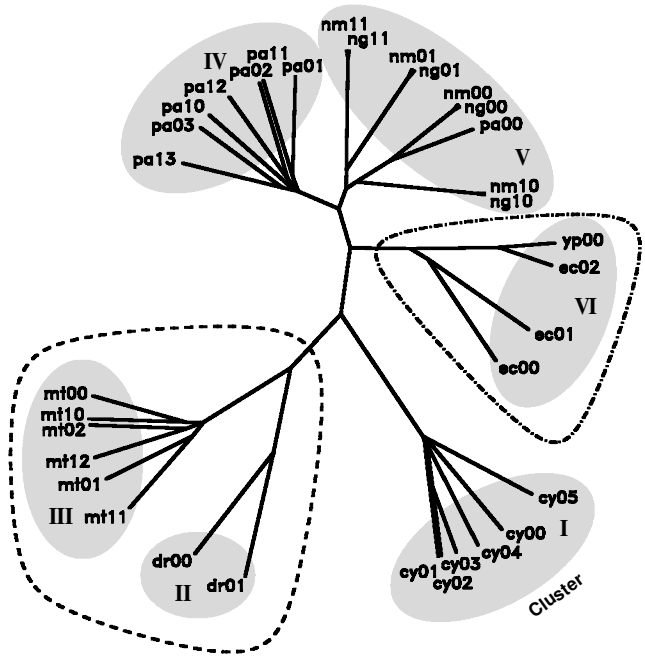


FIG. 3. Ferredoxin–NADPH reductase pathway. Phylogenetic tree of the pathway drawn by *PHYLIP* with parameters $f = 0$, $p = 1$, and $t = 1$. Clusters I to VI are referred to in the text. The id numbers (pIDs), which are uniquely assigned to each pathway representation, are a combination of two letters that code for the organism (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. Closed dashed and dot–dashed lines refer to subtrees of the phylogenetic tree that define homogeneous and heterogeneous clusters, respectively.

TABLE 3. LIST OF PIDS REFERRING TO ORF NAMES WITH CORRESPONDING FUNCTIONAL ROLES

pID	*Ferredoxin*	*Reductase*	*Cluster*	*pID*	*Ferredoxin*	*Reductase*	*Cluster*
cy00	slr1205	slr1643	I	ng00	RNG01106	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	RDR01803	RDR02099	II	nm10	RNM00363	RNM00963	V
dr01	RDR01783	RDR02099	II	nm11	RNM00662	RNM00963	V
ec00	YKGJ_ECOLI	FENR_ECOLI	VI	pa00	RPA01015	RPA07749	V
ec01	YFHL_ECOLI	FENR_ECOLI	VI	pa01	RPA08046	RPA07749	IV
ec02	FER_ECOLI	FENR_ECOLI	VI	pa02	FER_PSEAE	RPA07749	IV
mt00	fdxC	fprA	III	pa03	RPA07726	RPA07749	IV
mt01	fdxA	fprA	III	pa10	RPA01015	RPA05251	IV
mt02	fdxD	fprA	III	pa11	RPA08046	RPA05251	IV
mt10	fdxC	YZ14_MYCTU	III	pa12	FER_PSEAE	RPA05251	IV
mt11	fdxA	YZ14_MYCTU	III	pa13	RPA07726	RPA05251	IV
mt12	fdxD	YZ14_MYCTU	III	yp00	RYP01051	RYP02807	VI

names for all organisms of the phylogenetic tree. The induced subtrees of the phylogeny (Fig. 3) for *PR*s of *M. tuberculosis* and *D. radiodurans* are highlighted (cluster II and III).

Examples for homogeneous clusters[1] of paralog pathways are cluster I (*Synechocystis* sp.), II (*D. radiodurans*), III (*M. tuberculosis*), and IV (*P. aeruginosa*). The cyanobacterium *Synechocystis* sp. (cluster I), which carries a complete set of genes for oxygenic photosynthesis, is clearly separated from the nonautotroph bacteria. An interesting pair of homogeneous clusters corresponds to *D. radiodurans* (cluster II) and *M. tuberculosis* (cluster III). Not only in the case of the ferredoxin–NADPH reductase pathway, but also in the case of other electron transfer pathways, such as the malate–aspartate shuttle [11], PR cluster of the pathogen

Mycobacteria, and the ultrahigh radiation-tolerant *D. radiodurans* are closely related to each other. A special case represents cluster IV. The *PR*s of *P. aeruginosa* are present both in the homogeneous cluster IV as well as in cluster V.

Clusters V and VI are examples of heterogeneous clusters for *E. coli*, *Y. pestis, Neisseria*, and *P. aeruginosa*. The subtree that induces cluster VI is outlined in Fig. 3. In clusters V and VI distances between *PR*s of different organisms are shorter than a maximal distance that would include all paralog pathway representations of one organism. For example, the distance between *ec02* and *yp00* in cluster IV is shorter than the distance between *ec02* and *ec00*. The close relationship between genera of Enterobacteriaceae such as *E. coli* and *Y. pestis* is evident in the observed clustering of *E. coli* and *Y. pestis PR*s. Due to the heterogeneous nature of cluster VI in contrast to two homogeneous clusters for each *E. coli* and *Y. pestis*, horizontal transfer of ferredoxin–NADPH reductase *PR*s between these organisms suggests itself. Surprising similarities exist between *Neisseria* and *P. aeruginosa*. In addition to the homogeneous cluster of *P. aeruginosa PR*s (cluster IV) a second, heterogeneous cluster with *Neisseria* is observed (cluster V). The relationship between *Neisseria* and *P. aeruginosa* as shown in clusters IV and V is not as robust as in clusters I–III, and VI. By changing parameter f clusters I–III, and VI preserve the graph topology, and thus their relationship between *PR*s of corresponding subtrees. This is not the case for the corresponding subtree of clusters IV and V. Here, the relationship between *PR*s changes with f. This result suggests frequent and random gene replacements between *Neisseria* and *P. aeruginosa*. Within the genus *Neisseria* a very close relationship can be reported. Pathway representations of *Neisseria* always show up as ortholog pairs in phylogenies. For example, such pairs of ortholog *PR*s in cluster V are (*ng00, nm00*), (*ng01*, *nm01*), (*ng10, nm10*), and (*ng11, nm11*).

The number of genes that code for ferredoxin and ferredoxin reductase differs per organism. Often only a single gene codes for reductase but there may be as many as six ferredoxin genes present, e.g., in *Synechocystis* sp. (cluster I). The abundance of genes coding for ferredoxin compared to genes that code for ferredoxin reductase originate in the universality of ferredoxin. Ferredoxins, as redox reagens, serve in many more biochemical redox reactions than ferredoxin reductase.

4.2. Pathways with ferredoxin as functional role

Out of approximately 50 pathways in which ferredoxin plays a significant functional role, seven pathways were chosen for further investigations. The remaining nonused pathways are either found in one organism only or are absent at all due to missing functional roles:

2-Oxoglutarate, glutamine–glutamate anabolism (reduced ferredoxin)
NADPH–oxidized ferredoxin electron transport
NADPH–oxidized ferredoxin electron transport (plasma membrane)
H_2–H^+ catabolism (oxidized ferredoxin)
H^+–H_2 anabolism (reduced ferredoxin) (plasma membrane)
Nitrate–NH_4^+, OH^- catabolism (ferrocytochrome "c_{552}," reduced ferredoxin) (plasma membrane, cytosol)
Phosphoadenylylsulfate–sulfide anabolism

Pathways that are completely absent in the considered organisms are rare pathways that 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 present only 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. In addition to 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, is an observed pathway in Pseudomonadaceae, it has not been considered for further investigation due to missing functional roles in organisms of Table 1.

Figure 4 shows a simplified phylogenetic tree of 418 representations of 8 pathways with a confidence level of $t = 0.5$. Only one leaf per organism is drawn. An interesting similarity exists between the hyperthermophilic bacteria *A. aeolicus* and the archaeon *A. fulgidus* (cluster I). *Aquifex* with its representative *A. aeolicus* is exceptional among bacteria in the way that it occupies the hyperthermophilic niche otherwise dominated by archaea [37]. 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 adaption to the existence in an extreme thermophilic environment cannot be decided with the present number of completely sequenced microbial genomes. More genomes of both extremophilic archaea and thermophilic bacteria are necessary to detect possible horizontally

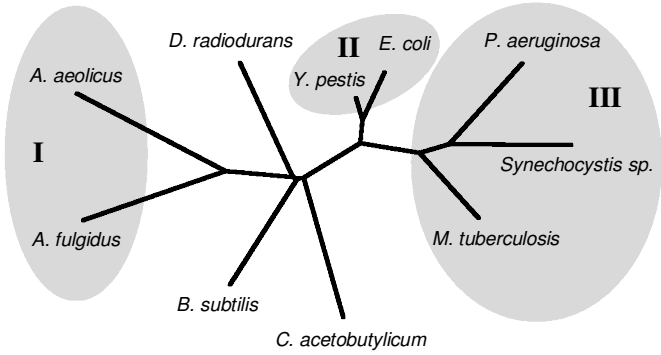


FIG. 4. Ferredoxin-related pathways: 418 different representations of ferredoxin-related pathways have been used to calculate the phylogeny (10 for *A. aeolicus*, 48 for *A. fulgidus*, 10 for *B. subtilis*, 42 for *C. acetobutylicum*, 4 for *D. radiodurans*, 200 for *E. coli*, 12 for *M. tuberculosis*, 72 for *P. aeruginosa*, 16 for *Synechocystis* sp., and 4 for *Y. pestis*). Only the common node per paralog pathway is shown. The distance matrix has been created with parameters $f = 1$, $p = 1$, and $t = 0.5$ and drawn using the program *PHYLIP* software suite.

transferred pathways. Similar to the observation reported in the previous section, *PR*s of *E. coli* and *Y. pestis* are closely related to each other (cluster II). Close to cluster I a second cluster is formed by *M. tuberculosis, P. aeruginosa*, and *Synechocystis* sp. (cluster III).

5. KREBS CITRIC ACID CYCLE

The evolutionary origin of the Krebs citric acid cycle (Krebs cycle) has long been a model case in the understanding of the origin and evolution of metabolic pathways. Although the chemical steps of the cycle are preserved intact throughout nature, diverse organisms make diverse use of its chemistry. In some cases organisms use only selected portions of the cycle.

For our analysis the Krebs cycle with feeder reactions via phosphoenolpyruvate, pyruvate, and acetyl-CoA, and the shortcut via glyoxylate, has been used as shown in Fig. 5. Only a single sequence per functional role and organism (with best hit against sequences with identical functional roles but of different organismic origin) has been chosen in our present study. Tables 4–6 show the functional roles for each pathway representation. Table 4 refers to the leg of the Krebs cycle in the oxidative direction (clockwise in Fig. 5), from oxaloacetate to 2-oxoglutarate (leg I). It also includes the 2-oxoglutarate dehydrogenase step from 2-oxoglutarate to succinyl-CoA. Table 5 lists the leg from oxaloacetate to succinyl-CoA in the reductive direction of the Krebs cycle (counterclockwise in Fig. 5, leg II). Table 5 shows the shortcut from isocitrate to succinate and malate, respectively, via glyoxylate. It also includes the feeding reactions from phosphoenolpyruvate and pyruvate, respectively, to oxaloacetate and citrate.

Figure 6 depicts two phylogenies of the Krebs cycle. In Figure 6a the phylogeny has been constructed with $\Delta_{\text{gap}} = 0.9$ and $\Delta_p = 17.1$. Thus, it assumes a nonnecessarily intact Krebs cycle and reveals relationships between organisms. Organisms within a genus, such as Neisseriaceae (*ng*, *nm*, cluster II), Mycoplasmae (*mg*, *mp*, cluster VIII), and Streptococci (*pn*, *st*, cluster VII), are closely related. Similar to observations made in previous sections, Enterobacteriaceae, such as *E. coli* and *Y. pestis*, as well as *H. influenzae* cluster together (cluster III). Likewise, close relationships between *D. radiodurans* and *M. tuberculosis* as well as *B. subtilis* (cluster I), *C. elegans* and *yeast* (cluster IV), *A. aeolicus* and *A. fulgidus* (cluster V), and *M. janaschii* and *M. thermoautotrophicus* (cluster VI) are found.

The distance matrix that is used to construct the phylogenetic tree shown in Figure 6b has been calculated with gap penalties $\Delta_{\text{gap}} = 1.8$ and $\Delta_p = 34.2$. Thus, missing functional roles influence the phylogeny strongly. Clearly, organisms are classified in the following groups:

1. *D. radiodurans, E. coli, M. tuberculosis, P. aeruginosa*, and yeast possess almost all functional roles shown in Tables 4–6 and thus utilize the complete Krebs cycle in their metabolism. *D. radiodurans, E. coli*, and *P. aeruginosa* lack pyruvate carboxylase (EC. 6.4.1.1). Organisms with complete Krebs cycle but with missing shortcut reactions via glyoxylate (isocitrate lyase, EC 4.1.3.1 and malate synthase, EC 4.1.3.2) are *C. elegans* and *B. subtilis*. The incomplete genomes of *Neisseriae* and *Y. pestis* may only presently lack single functional roles to close the Krebs cycle (malate dehydrogenase, EC 1.1.1.37 and ATP citrate synthase, EC 4.1.3.7, respectively), but may actually possess a complete Krebs cycle.

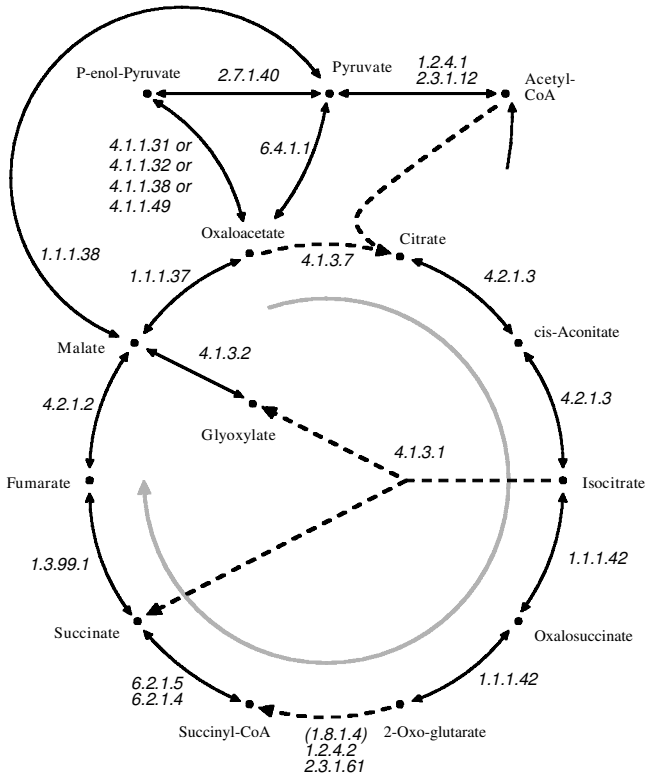


FIG. 5. Krebs citric acid cycle. Solid lines refer to reversible reactions. Dashed lines denote irreversible reactions. The arrowed, gray open circle indicates the direction of the reactions in the oxidative Krebs cycle. Enzymes are coded with their EC numbers. References to enzyme names are listed in Tables 4–6.

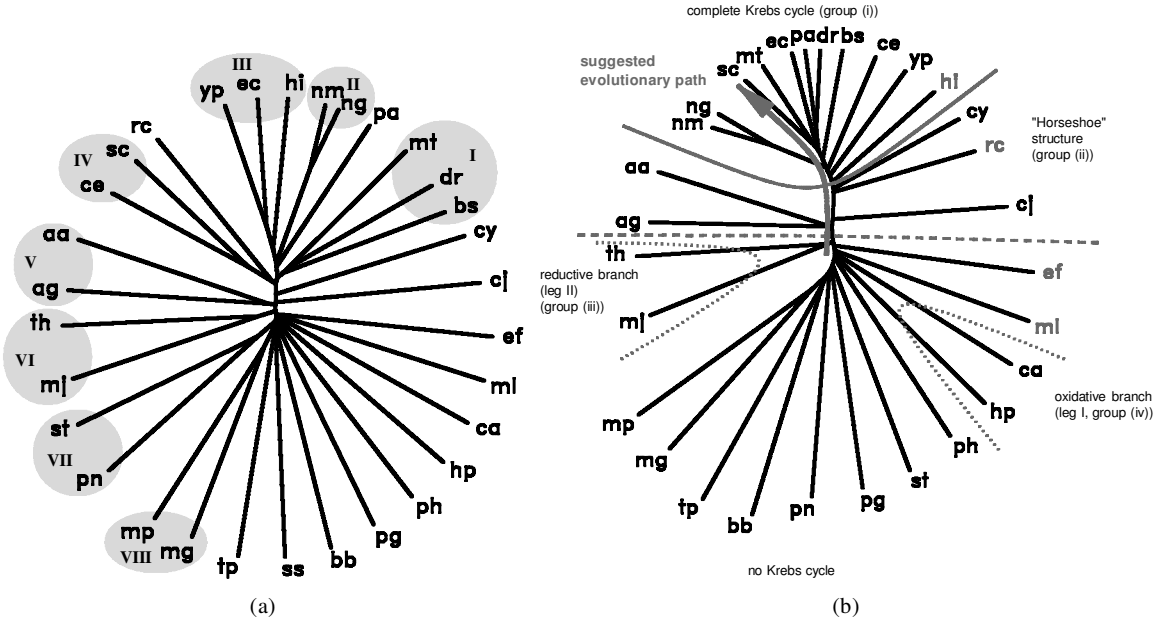


FIG. 6. Phylogenetic tree of the Krebs cycle drawn by the *PHYLIP* software suite. Nineteen functional roles are used according to Tables 4–6. Phylogenetic trees generated with gap penalties (a) $\Delta_{\text{gap}} = 0.9$, $\Delta_{p} = 17.1$ and (b) $\Delta_{\text{gap}} = 1.8$, $\Delta_{p} = 34.2$ are used as gap penalties. The gray solid, dashed, and dotted lines separate organisms with complete Krebs cycle, "horseshoe" structure, reductive and oxidative branch, and no Krebs cycle. Organisms with gray pID are not considered to be classfied into groups (see text). The solid line denotes a suggested evolutionary path. Parameters are set to $p = 1$ and $t = 0.01$.

TABLE 4. KREBS CITRIC ACID CYCLE

	Oxaloacetate → 2-Oxoglutarate			*2-Oxoglutarate → Succinyl-CoA*	
pID	*EC 4.1.3.7*[a]	*EC 4.2.1.3*[b]	*EC 1.1.1.42*[c]	*EC 1.2.4.2*[d]	*EC 2.3.1.61*[e]
ag	AF1340	—	AF0647	—	—
th	MTH1726	—	—	—	—
ph	—	—	PH1722	—	—
mj	—	—	—	—	—
aa	gitA	aco	icd	—	—
dr	RDR01411	RDR03579	RDR00757	RDR02407	RDR01815
ec	gltA	acnA	icdA	sucA	sucB
yp	—	RYP00454	RYP00837	RYP01676	RYP01675
hi	—	—	—	HI1662	HI1661
pa	RPA08103	RPA06672	RPA04040	RPA05385	RPA08107
ng	RNG00841	RNG01614	RNG00003	RNG00842	RNG00843
nm	RNM00765	RNM01689	RNM00931	RNM00766	RNM00767
rc	RRC01423	—	RRC01533	RRC02789	RRC02788
hp	HP0026	HP0779	HP0027	—	—
cj	RCJ00502	RCJ00379	RCJ01489	—	—
cy	sll0401	slr0665	slr1289	—	—
pg	—	—	—	—	—
bb	—	—	—	—	—
tp	—	—	—	—	—
ca	—	RCA01679	RCA01678	—	—
ml	—	—	—	—	—
mt	citA	acn	icd1	sucA	sucB
mg	—	—	—	—	—
mp	—	—	—	—	—
pn	—	—	—	—	—
st	—	—	—	—	—
ef	—	—	—	—	—
bs	citA	citB	citC	odhA	odhB
ce	CE00513	CE00516	F59B8.2	CE17244	CE14480
sc	CIT1	ACO1	IDP1	KGD1	KGD2

[a]ATP citrate synthase.
[b]Aconitate hydratase.
[c]Isocitrate dehydrogenase (NADP).
[d]2-Oxoglutarate decarboxylase.
[e]Dihydrolipoamide succinyltransferase.

2. In the next group of organisms the Krebs cycle is composed of two parts (oxidative, leg I, and reductive, leg II) and misses the 2-oxoglutarate dehydrogenase system. In *A. aeolicus, A. fulgidus, C. jejuni*, and *Synechocystis* sp. two partial versions of the cycle can be observed. Actually, all the enzymes of the CO_2-fixing, reductive citric acid cycle have been found in *Aquifex pyrophilus* [1], but not so in *Aquifex aeolicus*.
3. Autotrophic, anaerobic archaea such as *M. jannaschii* and *M. thermoautotrophicum* possess only the reductive part (leg II) of the Krebs cycle.
4. The oxidative part (leg I) can be observed in the strictly anaerobic *C. acetobutylicum* and microaerobic *H. pylori*.
5. Strict parasites with a reduced gene set, such as Mycoplasmae, *B. burgdorferi*, and *T. pallidum*, but also Streptococci, do not possess functional roles of the Krebs cycle.

Enterococcus faecalis, H. influenzae, R. capsulatus, and *M. leprae* are not considered to be classified in above five groups. The incomplete genomes of *E. faecalis, R. capsulatus*, and *M. leprae* may only presently lack functional roles of the Krebs cycle, whereas the facultative anaerobic *H. influenzae* may only possess

TABLE 5. KREBS CITRIC ACID CYCLE
Oxaloacetate ↔ Succinyl-CoA

pID	*EC 6.2.1.4–5a/b*[a]		*EC 1.3.99.1f*[b]*/ i*[c]		*EC 4.2.1.2*[d]	*EC 1.1.1.37/ EC 1.1.1.82*[e]
ag	AF2185	AF1540	AF0681	AF1773	AF1098	AF0855
th	MTH563	MTH1036	MTH1502	MTH1850	MTH1910	MTH1205
ph	—	—	—	—	PH1684	PH1277
mj	MJ1246	MJ0210	MJ0033	MJ0092	MJ0617	MJ1425
aa	sucD2	sucC2	frdA	frdB1	fumX	RAA00600
dr	RDR00373	RDR00374	RDR02612	RDR01584	RDR02593	RDR03472
ec	sucD	sucC	sdhA	sdhB	fumC	mdh
yp	RYP04284	RYP04283	RYP04292	RYP04293	RYP01155	RYP01708
hi	HI1197	HI1196	HI0835	HI0834	HI1398	HI1210
pa	RPA01727	RPA08108	RPA06227	RPA05961	Q51404	RPA04569
ng	RNG00847	RNG00846	RNG00012	RNG00840	RNG01401	—
nm	RNM00772	RNM00771	RNM00762	RNM00763	RNM01514	—
rc	RRC02792	RRC02793	RRC00663	RRC02776	—	—
hp	—	—	HP0192	HP0191	HP1325	—
cj	RCJ01484	RCJ01485	—	RCJ02431	—	RCJ01486
cy	sll1557	sll1023	slr1233	sll1625	slr0018	sll0891
pg	—	—	RPG01822	RPG01821	RPG01372	—
bb	—	—	—	—	—	—
tp	—	—	—	—	—	—
ca	—	—	—	—	RCA01368	RCA02427
ml	—	—	Q49920	Q49916	—	P50917
mt	sucD	sucC	sdhA	sdhB	fum	mdh
mg	—	—	—	—	—	—
mp	—	—	—	—	—	—
pn	—	—	—	—	—	—
st	—	—	—	—	—	—
ef	REF02444	—	REF00652	—	—	REF01552
bs	sucD	sucC	sdhA	sdhB	citG	citH
ce	C05G5.4	F47B10.1	F48E8.3	F42A8.2	CE11580	F36A2.3
sc	P53598	P53312	SDH1	DHSB	FUM1	MDH1

[a]Succinyl-CoA ligases, α- and β-chain, respectively.
[b]Succinate dehydrogenase, flavoprotein subunit.
[c]Succinate dehydrogenase, iron-sulfur protein.
[d]Fumarate hydratase.
[e]Malate dehydrogenase.

functional roles needed for growth in heme and glutamate-rich media [10]. Glutamate can be directed into the Krebs cycle by conversion to 2-oxoglutarate by glutamate dehydrogenase.

As already suggested by Weitzman [58], the origins of the Krebs cycle may be found in the more primitive anaerobic organisms of the past. He proposed that the evolution of pyruvate–ferredoxin oxidoreductase, a widely distributed system in archaea that yields $acetyl\text{-}CoA + CO_2 + H_2$, led to the appearance of a citrate synthase and the reaction sequence $oxaloacetate + acetyl\text{-}CoA \rightarrow citrate \rightarrow cis\text{-}aconitate \rightarrow isocitrate \rightarrow 2\text{-}oxoglutarate$. Already 4 years earlier Gest [16] suggested that the reductive sequence $oxaloacetate \rightarrow malate \rightarrow fumarate \rightarrow succinate$ arose first as a mechanism to accept electrons generated in sugar fermentation and, thus, regenerate electron carriers. Gest's and Weitzman's sequences would have provided two legs of an interrupted cycle [40]. A similar "horseshoe" structure of the Krebs cycle has been recently considered by Meléndez-Hevia *et al.* [30], who suggested the final linkage to form a complete Krebs cycle between succinyl-CoA and succinate, which cannot be confirmed by our studies. The phylogenetic analysis (Fig. 6b) emphasizes the evolution from individual legs (leg I and II, organisms of group 4 and 3, respectively) of the Krebs cycle via the horseshoe structure (group 2) toward the full cycle (group 1). The suggested evolutionary path from *H. pylory* to *yeast* is shown in the phylogeny.

TABLE 6. KREBS CITRIC ACID CYCLE: SHORTCUT VIA GLYOXYLATE AND FEEDING REACTION

pID	*EC 4.1.3.1*[a]	*EC 4.1.3.2*[b]	*EC 2.7.1.40*[c]	*EC 6.4.1.1.1*[d]	*EC 4.1.1.al*[e]	*EC 1.2.4.1*[f]	*EC 2.3.1.12*[g]	*EC 1.1.1.38*[h]
ag	—	—	—	—	—	—	—	AF1727
th	—	—	—	—	—	—	—	—
ph	—	—	PH0570	—	PH0312	—	—	PH1275
mj	—	—	MJ0108	—	—	—	—	—
aa	—	—	—	—	—	—	—	—
dr	RDR02937	RDR00925	RDR02583	—	RDR02659	RDR00099	RDR01507	RDR00927
ec	aceA	glcB	pykF	—	ppc	aceE	aceF	b2463
yp	RYP03554	—	RYP01135	—	RYP03962	—	—	RYP01714
hi	—	—	HI1573	—	HI1636	HI1233	HI1232	HI1245
pa	RPA04897	RPA00593	RPA04577	—	RPA06249	RPA01374	RPA05361	RPA00241
ng	—	—	RNG01302	—	RNG00495	RNG01168	RNG01167	RNG00636
nm	—	—	RNM01327	—	RNM00284	RNM00293	RNM00295	RNM01547
rc	RRC01711	RRC01957	RRC00307	RRC02720	—	RRC03495	—	RRC03007
hp	—	—	—	—	—	—	—	—
cj	—	—	RCJ02399	—	RCJ01897	—	—	RCJ02660
cy	—	—	sll1275	—	sll0920	sll1721	sll1841	slr0721
pg	—	—	—	—	—	—	—	RPG01305
bb	—	—	BB0348	—	—	—	—	—
tp	—	—	—	—	TP0122	—	—	—
ca	—	—	RCA02616	RCA02904	—	—	—	23634675_F2_6
ml	P46831	—	—	—	CAPP	RML00805	—	RML00065
mt	aceA	glcB	pykA	pca	pckA	aceE	pdhC	mez
mg	—	—	MG216	—	—	MG274	MG272	—
mp	—	—	MP534	—	—	MP446	MP448	—
pn	—	—	RPN00842	—	RPN01541	RPN00131	RPN00130	—
st	—	—	RST00804	—	RST00317	RST00962	RST00961	RST00041
ef	—	—	REF00443	REF00815	—	REF00473	REF00475	REF01680
bs	—	—	pykA	pycA	pckA	pdhA	acoC	malS
ce	—	—	CE15899	D2023.2	R11A5.4	T05H10.6	F23B12.5	—
sc	ICL1	MLS1	PYK2	PYC1	PCK1	PDA1	LAT1	sp p36013

[a]Isocitrate lyase.
[b]Malate synthase.
[c]Pyruvate kinase.
[d]Pyruvate carboxylase.
[e]Phosphoenolpyruvate carboxylases.
[f]Pyruvate decarboxylase.
[g]Dihydrolipoamide dehydrogenase.
[h]Malate oxidoreductase (NAD).

6. DISCUSSION

Our method represents a new approach for the comparison of metabolic pathways based on explicit sequence information. To illustrate the method, two electron transport pathways have been analyzed: (1) the ferredoxin–NADPH reductase pathway and (2) pathways utilizing ferredoxin. The analysis reveals a close relationship between pathways of organisms within the same genus. According to Woese [59], metabolic genes are among the most modular in the cell, and their genes are expected to travel laterally, even today. Such adaptions of single genes as well as horizontal transfer of complete pathways between organisms are confirmed by our phylogenetic analysis.

The analysis of the Krebs citric acid cycle confirms earlier results on the design of this metabolic pathway during evolution [16, 40, 58]. The origins of key reactions are found in the more primitive anaerobic organisms of the past. Two branches of the Krebs cycle, the oxidative branch via *oxaloacetate* + *acetyl-CoA* → *citrate* → *cis-aconitate* → *isocitrate* → *2-oxoglutarate* and the reductive branch via *oxaloacetate* → *malate* → *fumarate* → *succinate*, can be linked by the 2-oxoglutarate oxidoreductase system.

A statistical analysis of electron transfer pathways for each organism has been performed. The analysis shows an overrepresentation of the fraction of assigned electron transfer pathways for *M. tuberculosis* and a lack for *B. burgdorferi*, respectively. These results suggest different evolutionary forces of the environment on the organisms and the adaption of the organisms to these forces. The high fraction of assigned electron transfer pathways for *M. tuberculosis* reflects the ability of the bacillus to adapt to environmental changes. In contrast, the obligatory parasitic *B. burgdorferi*, which lacks a respiratory electron transport chain due to missing cytochromes, possesses only three electron transfer pathways.

A surprising relationship between ferredoxin–NADPH reductase pathway representations of *Neisseria* and *P. aerugiosa* has been observed. In contrast to clusters I–III, and VI of the phylogenetic tree (Fig. 3), the corresponding subtrees of clusters IV and V, and, thus, the relationships between *PR*s in clusters IV and V change when the discrimination between paralog and ortholog genes [parameter f in Equation (1)] is changed. This nonconserved relationship between pathway representations suggests a frequent and random exchange of pathways between *Neisseria* and *P. aerugiosa*.

We propose a similarity between organisms of different domains in the special case of the thermophilic bacteria *A. aeolicus* and archaea. Regarding ferredoxin-utilizing pathways, *A. aeolicus* is closely related to *A. fulgidus*. One possible explanation of the close relationship between these two organisms is the continuous acquisition of thermotolerance genes from preadapted hyperthermophiles by *A. aeolicus*. It might also be just a consequence of adaption to the existence in an extreme thermophilic environment. More genomes of both extremophilic archaea and thermophilic bacteria are necessary for a decision in this case.

With our approach, we have not only classified relationships between genes, but also between pathways. Ongoing studies combine investigations of relation and evolution of larger metabolic networks that cover more than one class of intermediate metabolic and bioenergetic networks, for example, information-processing networks, electron transport, transmembrane transport, and signal transduction. For this purpose a general graph-theoretical approach is pursued that can take into account differences between the topologies of reaction networks and compare interacting genes embedded in a metabolic network.

7. APPENDIX

In the following we define concepts and expressions that are used throughout this paper.

Metabolic Networks and Pathways. A *metabolic network* is a directed reaction graph with substrates as vertices and directed, labeled edges denoting reations 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* that process substrates in a specific reaction or *substrates* that 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 (according to Fig. 1) 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 (PR)*.

Homogeneous and Heterogeneous Clusters of Pathway Representation ***PR*****.** A classification in phylogenetic trees is made according to distances between each of two *PR*s. *PR*s can be grouped in *clusters*. A *cluster* 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 cluster is smaller than between members and nonmembers. A *homogeneous cluster* is defined as a cluster with pathway representations of a single organism. The maximum distance in a homogeneous cluster 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 cluster. A *heterogeneous cluster* is a cluster 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* that codes for the same function) are orthologs. Two pathway representations are defined *paralog* to each other if at least one gene pair is paralog.

ACKNOWLEDGMENTS

Fruitful discussions with Ross Overbeek from the WIT-team at Argonne National Laboratory is gratefully acknowledged. This work is supported by grants from the National Institutes of Health (NIH PHS 5 P41 RR05969), the National Science Foundation (NSF BIR 94-23827 EQ), and the Beckman Institute.

REFERENCES

Beth, M., Strauss, G., Huber, R., Stetter, K., and Fuchs, G. 1993. Enzymes of the reductive citric acid cycle in the autotrophic eubacterium *Aquifex pyrophilus* and in the archaebacterium *Thermoproteus neutrophilus*. *Arch. Microbiol*. 160, 306–311.

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

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.

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

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

T. *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* 282, 2012–2018.

Felsenstein, J. 1997. Phylip software. http://evolution.geneti cs.washington.edu/ phylip.html.

Fitch, W. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool*. 19, 99–113.

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.

Forst, C.V., and Schulten, K. 1999. Phylogenetic analysis of metabolic pathways. *J. Mol. Evol*. Submitted.

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

Genome Therapeutics Corporation. *Clostridium acetobutylicum* genome, strain ATCC824. http://www.genomecorp. com/genesequence s/clostridium/closp age.html.

Gest, H. 1981. Evolution of the citric acid cycle and respiratory energy conversion in prokaryotes. *FEMS Microbiol. Lett*. 12, 209–215.

Goffeau, A., *et al*. 1997. The yeast genome directory. *Nature (London)* 387, 5–105.

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., and Henikoff, J. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* 89, 10915–10919.

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

Huson, D.H. 1998. Splitstree: Analyzing and visualizing evolutionary data. *Bioinformatics* 14, 68–73.

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.

Karp, P., and Riley, M. 1993. Representations of metabolic knowledge, 207–215. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, New York.

Karp, P., Riley, M., and Pellegrini-Toole, A. 1996. EcoCyc: Electronic encyclopedia of *E. coli* genes and metabolism. *Nucleic Acids Res.* 24, 32–40, 1996. URL: http://ecocyc.Pangea Systems.com/ecocyc/ecocyc.html.

Kawarabayasi, 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 (London)* 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 (London)* 390, 249–256.

Lipmann, F. 1965. Fox, S.W., ed. *The Origin of Prebiological Systems and of Their Molecular Matrices*, 259–280. Academic Press, New York.

Meléndez-Hevia, E., Waddell, T.G., and Cascante, M. 1996. The puzzle of the Krebs citric acid cycle: Assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. *J. Mol. Evol.* 43, 293–303.

Miller, S.L. 1953. A production of amino acids under possible primitive earth conditions. *Science* 117, 528–529.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 27, 29–34.

Oparin, A.I. 1967. The origin of life. In Bernal, J., ed. *The Origin of Life*. World, Cleveland. Also published in Proiskhozhdenie Zhizny. IZD Moskovishii Rabochii, Mowcow, 1924.

Orgel, L.E. 1968. Evolution of the genetic apparatus. *J. Mol. Biol.* 38, 381–383.

Overbeek, R., Larsen, N., Smith, W., Maltsev, N., and Selkov, E. 1997. Representation of function: The next step. *Gene* 191, GC1–GC9.

Overbeek, R., Pusch, G., Dsouza, M., Larsen, N., Selkov, E., Jr., Selkov, E., and Maltsev, N. 1998. What is there—interactive metabolic reconstruction on the web. http://wit.mcs.anl.gov/WIT2/wit.html, WIT2.

Pace, N.R. 1997. A molecular view of microbial diversity and the biosphere. *Science* 276, 734–740.

Popper, K.R. 1957. *The Poverty of Historicism*. Routledge & Kegan Paul, London.

Popper, K.R. 1963. *Conjectures & Refutations*. Routledge & Kegan Paul, London.

Romano, A., and Conway, T. 1996. Evolution of carbohydrate metabolic pathways. *Res. Microbiol.* 147, 448–455.

Smith, D., Doucette-Stamm, L. Deloughery, C, H.L.C., *et al.* 1997. Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: Functional analysis and comparative genomics. *J. Bacteriol.* 179, 7135–7155.

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

The Institute of Genome Research. *Deinococcus radiodurans* genome. ftp://ftp.tigr.org/pub/data /d_radiodurans/.

The Institute of Genome Research. *Enterococcus faecalis* genome. ftp://ftp.tigr.org/pub/data /e_faecalis/.

The Institute of Genome Research. *Neisseria meningitidis*. ftp://ftp.tigr.org/pub/dat a/n_meningitidis/.

The Institute of Genome Research. *Streptococcus pneumoniae* genome. ftp://ftp.tigr.org/pub/data /s_pneumoniae/.

The Institute of Genome Research and Forsyth Dental Center. *Porphyromonas gingivalis* genome http://www.forsyth.org/pggp/.

The Sanger Centre. *Campylobacter jejuni* genome. ftp://ftp.sanger.ac.uk/pu b/pathogens/cj/.

The Sanger Centre. *Mycobacterium leprae* genome. ftp://ftp.tigr.org/pub/data /m_tuberculosis/.

The Sanger Centre. *Yersinia pestis* genome. http://www.sanger.ac.uk/Projects/Y_pestis/.

Thompson, J., Higgins, D., and Gibson, T. 1994. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680.

Tomb, J.-F., White, O., Kerlavage, A., Clayton, R., *et al.* 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature (London)* 388, 539–547.

University of Chicago and Institute of Molecular Genetics, CSFR. *Rhodobacter capsulatus*. http://capsulapedia.uchicago.edu/.

University of Oklahoma. *Neisseria gonorrhoeae* genome. ftp://ftp.genome.ou.edu/pub/gono.

University of Oklahoma. *Streptococcus pyogenes* genome. http://dnal.chem .uoknor.edu/strep.html .

Univeristy of Washington and PathoGenesis. *Pseudomonas aeruginosa* genome. http://www.pseudomonas.com/.

Wächtershäuser, G. 1990. Evolution of the first metabolic cycles. *Proc. Natl. Acad. Sci. U.S.A.* 87, 200–204.

Weitzman, P. 1985. Evolution in the citric acid cycle, 253–275. In Schleifer, K.H., and Stackebrandt, E., eds., *Evolution of Prokaryotes*, Volume 29 of *FEMS Symposium*. Academic Press, Orlando, FL.
Woese, C. 1998. The universal ancestor. *Proc. Natl. Acad. Sci. U.S.A.* 95, 6854–6859.

Address reprint requests to:
Christian V. Forst
Theoretical Biophysics
Beckman Institute
University of Illinois, Urbana-Champaign
Urbana, IL 61801

E-mail: chrisf@ks.uiuc.edu