Evolution of Metabolisms: A new Method for the Comparison of Metabolic Pathways

Christian V. Forst and Klaus Schulten
Theoretical Biophysics, Beckman Institute,
University of Illinois, Urbana-Champaign,
405 North Mathews Avenue, Urbana, IL, 61801,
USA

Abstract

The abundance of information provided by completely sequenced genomes defines a starting point for new insights in the multi-level 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 - substrates are processed by enzymes yielding other substrates.
A new method based on a combination of sequence information with graph-topology of the underlying pathway is presented. By this approach pathways of different organisms are related to each other by phylogenetic analysis: Metabolic pathways in organisms are assigned by the presence of genes with corresponding functions. A global distance between pathways is defined by using individual distances between sequences of the same functional role. Gap-penalties are introduced for existing pathways with missing functional roles.
As an example the method is applied to pathways related to the utilization of ferredoxin as redox-agent. First results report a high versatility and controversial phylogenies. Pathway which evolve early in evolution are present in organisms of all three domains (Archaea, Bacteria and Eucarya). Functional roles which are utilized in these very pathways are common in many pathways of the cell. Pathways which are only present in organisms of a single (or at maximum two) domains often perform "pathologic" tasks with rare and specialized functional roles. As example serves NADPH – Hg^{2+} Electron transport with mercury(II) reductase reaction as enzyme.
Beside a more comprehensive understanding of similarities and differences between organisms, this method already indicates different evolutionary rates between substrates and enzymes. It is also capable to hint towards variation and specialization of pathways in the sense of an evolution of metabolisms.

1 Introduction

Studying metabolisms of living systems and understanding their evolution is an old field of research. First studies to this area have been performed in the late 50s and early 60s by Karl Popper [30, 31] and Fritz Lipmann [23], and have been followed by other scientists. This extensive research is motivated by questions regarding the *Origin of Life and the Evolution of the Biosphere*. Seminal contributions by J.B.S. Haldane [14], Stanley Miller [25], Alexander Oparin [26], and Leslie Orgel [27] are mentioned in this context, discussing the (prebiotic) chemical environment suitable for a biotic evolution. Based on these discussions "speculations" on the origin and evolution of metabolism started [15] and questions regarding the emergence of the first metabolic cycles were addressed [50]. As a paradigm for a cyclic metabolic network serves the Krebs citric acid cycle in the respiratory chain to produce ATP by oxidizing acetate. Quite recently Meléndez-Hevia *et al.* [24] suggested the evolution of the Krebs cycle via a "horseshoe" structure, where one part of the cycle functions in the reversed order and the connection between succinate and succinyl-CoA by succinyl-CoA synthetase is missing.

All these studies address questions regarding energetics of chemical reaction and the possibility of their existence. Enzymes are abstract concepts to process chemical compound without further information of the catalysts themselves. In this contribution a synthesis between evolution of "abstract" metabolic pathways and "stand-alone" biomolecular enzymes and substrates is made. In chapter 2 a web-based information system suitable for this task is presented. Chapter 3 introduces a method for calculating distances between metabolic networks based on sequence-information of the involved biomolecules. Finally, in chapter 4 the method is applied to selected electron transfer pathways.

2 30 Genomes

2.1 WIT

A convenient system to obtain (almost) complete genomic, and organizational information about microbial organisms is the *WIT-system* by Overbeek *et al.*[29]. *W*hat *I*s *T*here, whose abbreviation is *WIT*, provides researchers with DNA and protein sequence information of complete or partially sequenced genomes. This information is associated with organizational information, gene- and operon-organization and especially information about metabolic networks. A major

goal of WIT is to perform a so called *metabolic reconstruction* of microbial genomes [28].

2.2 30 Genomes en détail

Following 29 genomes of microbial origin and of one multicellular organism (*C. elegans*) are accessible via the WIT-system. Currently 16 genomes are completely sequenced. The remaining 14 genomes are involved in ongoing sequencing projects and have to be accessed with care. In table 1 a summary is given.

3 "Distance" between Metabolic Networks

Aligning sequences to each other and measuring distances, e.g. BLOSUM [16] and PAM [4] similarity matrices, is a common approach to deduce a relationship between individual biopolymers.

In this paper an extension towards 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. Preliminary, we confine our approach to networks of the same topology to each other:

Given a small metabolic pathway of a substrate which is processed by an enzyme. The global distance Δ between such pathways is deduced by the individual distances between substrates ΔS and enzymes ΔE. In Fig. 1 an example for a simple pathway with two functional roles ($n = 2$) is shown. A general definition for n functional roles per

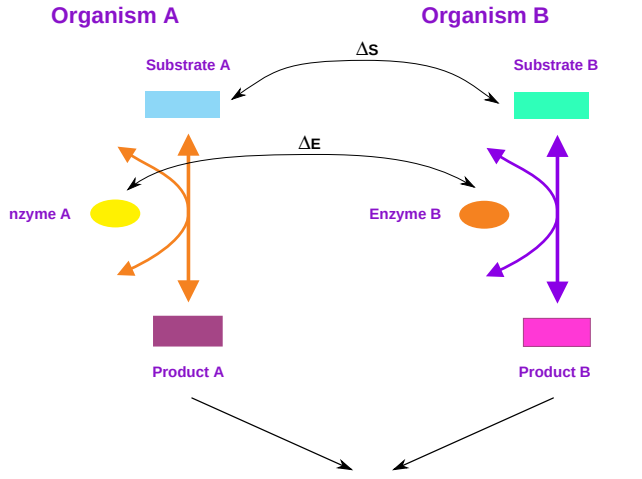

Figure 1: Calculating a distance between two pathways: individual distances $\Delta E, \Delta S$ between sequences of the same functional role are used to calculate a global distance Δ.

pathway is given in the following definition:

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

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

whereas $0 \le f \le 1$ and $p > 0$.

Paralog genes are discriminated against orthologs due to 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 [7]. Normally, orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if related to the original [33]. To reflect these differences for calculating distances between pathways parameter p is introduced in equ. 1. If $f = 1$ then 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.

Exponent p in 1 is introduced to tune the discrimination between long and short distances. Large (positive) values of p emphasizes large distances between individual functional roles.

Remark 1 *Due to the nonlinear property of equ. 1 the resulting distance Δ is no absolute measure of a relationship between two pathways. But the resulting phylogeny is valid for a qualitative analysis of a relationship between metabolisms of organisms.*

3.1 Gaps

It is often meaningful to consider pathways with missing functional roles. This is especially the case when there are strong hints by experiments that a certain pathway is present in the corresponding organism. A simple but nevertheless useful approach is to introduce so called *gap penalties* for these missing functional roles:

Assume that in a pathway Γ a distinct functional role I_k is missing. Thus according to Def. 1 the distance ΔX_k is not defined which yields following definition of a gap penalty:

Def. 2 *A* gap penalty *is an arbitrarily assigned value Δ_{gap} to the otherwise undefined distance $\Delta_{gap} \doteq \Delta_k = \delta(., I'_k)$.*

Throughout the paper a gap penalty of $\Delta_{gap} = 1.0$ is used. This *gap penalty* must not be mistaken with the gap penalties used in sequence alignment. Additionally a threshold t as *confidence level* for the maximum number of accepted gaps can be defined. For the number of gaps exceeding t a pathway is no longer considered as valid.

Remark 2 *Due to the nature of distance matrices, where distances between all possible combinations of members are present, it may happen that distances between invalid pathways (as defined above) are assigned. In this case these distances are reassigned by a* penalty distance $\Delta_p = 8.0$.

4 Electron Transfer

Electron transport pathways play a key role in the metabolism of a living cell. There are about 69[1] pathways known which are related to electron transfer. Table 2 previews a subset of 15 selected pathways out of the total 69. First let us define the following:

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

Table 1: 29 Microbial Genomes + One Multicellular Organism

Code	Organism	KD [a]	Size[kB] [b]	# ORFs [2]	cs [c]
AG	Archaeoglobus fulgidus	A	2178.40	2493	x[21]
TH	Methanobacterium thermoaut.	A	1751.38	1866	x[32]
PH	Pyrococcus horikoshii	A	1738.51	1825	x[20]
MJ	Methanococcus jannaschii	A	1739.93	1797	x[2]
AA	Aquifex aeolicus	B	1590.78	1744	x[5]
DR	Deinococcus radiodurans	B	3261.20	3771	o[35]
EC	Escherichia coli	B	4639.22	4289	x[1]
YP	Yersinia pestis	B	4501.71	4296	o[40]
HI	Haemophilus influenzae	B	1830.14	1846	x[8]
PA	Pseudomonas aeruginosa	B	6286.26	6477	o[49]
NG	Neisseria gonorrhoea	B	2063.17	1853	o[47]
NM	Neisseria meningitidis	B	2157.54	1838	o[36]
RC	Rhodobacter capsulatus SB1003	B	2024.62	2099	o[46]
HP	Helicobacter pylori	B	1667.88	1547	x[45]
CJ	Campylobacter jejuni	B	1644.03	2106	o[41]
CY	Synechocystis sp.	B	3573.47	3226	x[19]
PG	Porphyromonas gingivalis	B	2447.62	1832	o[39]
BB	Borrelia burgdorferi	B	1503.21	1688	x[9]
TP	Treponema pallidum	B	1138.82	946	x[11]
CA	Clostridium acetobutylicum	B	4030.73	3967	o[12]
ML	Mycobacterium leprae	B	2420.76	1568	o[42]
MT	Mycobacterium tuberculosis	B	4411.53	3924	x[3]
MG	Mycoplasma genitalium	B	580.07	532	x[10]
MP	Mycoplasma pneumoniae	B	816.39	674	x[17]
PN	Streptococcus pneumoniae	B	2104.82	1844	o[37]
ST	Streptococcus pyogenes	B	1799.24	1599	o[48]
EF	Enterococcus faecalis	B	3209.12	2967	o[38]
BS	Bacillus subtilis	B	4214.81	4093	x[22]
SC	Saccharomyces cerevisiae	E	12057.28	6125	x[13]
CE	Caenorhabditis elegans	E	63729.49	10064	o[43]

[a] Kingdom: A... Archaea, B... Bacteria, E... Eucarya
[b] These numbers has to be taken with care by ongoing sequence projects
[c] Completely sequenced genomes are marked by an x

Def. 3 *A pathway* Γ *is defined to be* present *in the organism if the fraction of present functional roles for* Γ *exceeds a certain confidence level* t*. This, of course, is a formal definition and has always be confirmed by experiments.*

The confidence level t for pathways in Table 2 is 100%. Four members of this family are present in organisms of all three domains:

Dihydrolipoamide – NAD^+ Electron Transport (pathway 2)
NADH – FAD Electron transport (plasma membrane) (pathway 4)
NADH – NAD^+ Electron Transport (malate, aspartate) (pathway 10)
NADPH – Oxidized thioredoxin Electron transport (pathway 15)

The substrates utilized in these pathways also play functional roles in other metabolisms. E.g. lipoamide dehydrogenase component (E3) (EC 1.8.1.4) (pathway 2) catalyzes reaction in carbohydrate metabolisms. NAD-dehydrogenase (EC 1.6.99.3) (pathway 4) is involved in many electron transport reactions. Aspartate aminotransferase (EC 2.6.1.1) and malate dehydrogenase (EC 1.1.1.37) (pathway 10) serve as reaction partners in aminoacid- and carbohydrate metabolisms. And last but not least thioredoxin and its reductase (EC 1.6.4.5) (pathway 15) is found in a variety of metabolism. Thioredoxin reductase catalyzes many electron transfer reactions (as two examples serve pathways 1, 15). Thioredoxin itself is involved in 29 pathways related to aminoacid-, 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). This pathway is only missing in *P. horikoshii*, *R. capsulatus*, and *C. elegans*. The incomplete genomes of *R. capsulatus* and *C. elegans* lack the sequence of thioredoxin reductase. Wherease *P. horikoshii* does not possess thioredoxin.

On the contrary, many electron transfer pathways are only present in organisms of one or two domains. Specialized pathways involving mitochondria are of course found in Eucarya only (pathways 5, 9). A large number of pathways are assigned to Bacteria only (pathways 2, 7, 8, 12, 13, 14). For example the "pathological" mercury(II) reductase reaction (pathway 12) is utilized by four Bacteria only. Surprisingly there are no electron transport pathways which are unique for Archaea. This observation confirms the know fact that entire pathways are acquired or displaced in the case of the Archaea. Actually, by investigating frequency

distribution 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[2].

Studying pathway-"distribution" within a domain (here: Bacteria) shows the following. In general one can observe a coarse correlation between genome size (or number of ORFs) and number of present pathways. As exception to this rule serves *B. burgdorferii* which genome is almost as long as the genome of *H. influenzae*: In contrast to *H. influenzae* which uses many electron transfer pathways, highly parasitic *B. burgdorferii* posesses only NADPH – oxidized thioredoxin electron transport (pathway 15). Without surprise, close related organisms use similar pathways: each pair of *Neisseria* NG, NM; *Mycobacteria* ML, MT; and *Mycoplasmae* MG, MP have the same set of pathways. Here the exception are *Streptococcae*, where *S. pyrogens* posesses a richer set of pathways than *S. pneumoniae*. From the family of *Mycobacteria* and *Mycoplasmae*, *M. tuberculosis* utilizes unproportionally many pathways. A special state manifests the quadruple *E. coli*, *Y. pestis*, *H. influenzae*, and *P. aeruginosa*. *E. coli* as one of the best studied microbial organism has the most complete set of pathways. Close related to this species the pathogens *H. influenzae*, *Y. pestis* and the metabolic very versatil *P. aeruginosa* possess a similar but less abundance of electron transfer pathways. In contrast to comparative genomics studies performed for *E. coli* and *H. influenzae* genomes [34] there are no pathways present in pathogens which are missing in free living relatives.

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. E.g. ferredoxin is redox-partner in more than 50 known pathways as part of amino acid metabolism, aromatic hydrocarbons metabolism, carbohydrate metabolism, electron transfer, hydrogen metabolism, lipid metabolism, nitrogen metabolism, and sulfur metabolism.

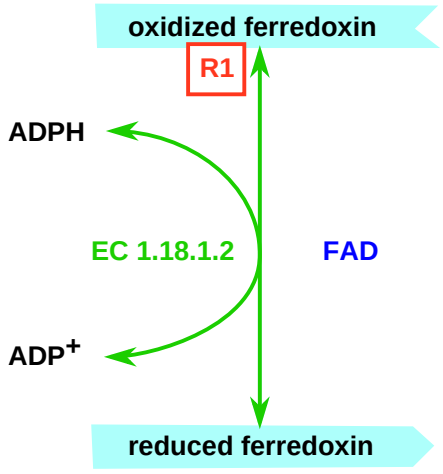

Figure 2: Ferredoxin – NADPH reductase pathway: The pathway itself is shown with ferredoxin and ferredoxin – NADPH reductase (EC 1.18.1.2) as involved biopolymers.

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

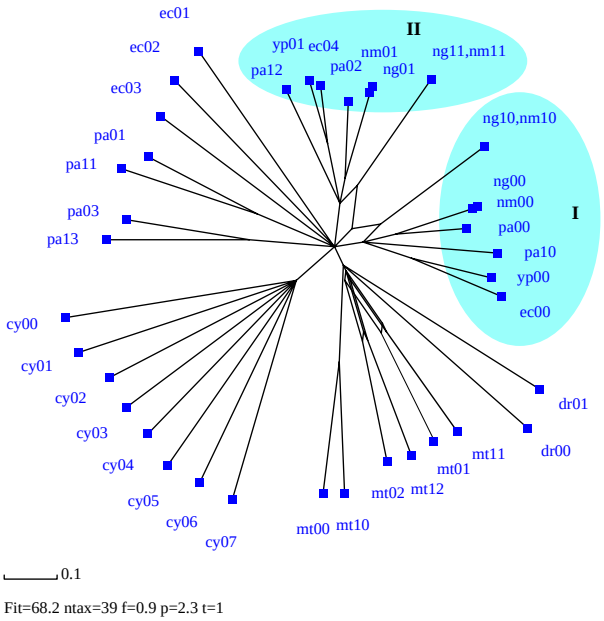

Figure 3: Ferredoxin – NADPH reductase pathway: Phylogenetic graph of the pathway drawn by *SPLITSTREE2* with parameters $f = 0.9$, $p = 2.3$, and $t = 1$. Cluster I and II are referred to in the text.

One quite important pathway in this class is the reversible *Ferredoxin – NADPH Reductase pathway* which can be found in *Bacillaceae*, *Cyanobacteria* and *Enterobacteriaceae*. In this redox-pathway ferredoxin is processed (either oxidized or reduced) by ferredoxin – NADPH reductase. Applying the method outlined in chapter 3 on this pathway yields the following. Sequences for ferredoxin and its reductase are obtained from the WIT-system. A multiple sequence alignment for each set of sequences is performed by ClustalW v1.74[44] with the BLOSUM62 similarity parameter. Alignment parameter are set to default values. The created Phylip distance matrices are then used for calculations of the pathway distance. Phylogenetic relationships are analyzed by phylogenetic graph reconstruction programs such as *SPLITSTREE2* [18] or the *PHYLIP* software suite [6]. Schematics of the pathway and a non-rooted phylogenetic graph drawn by *SPLITSTREE2* are shown in Fig. 2 and Fig. 3 respectively.

References to the ORF-names used in the pathways are listed in table 3. The nomenclature is according to the WIT-system. All pathways of the same organism are clustered in general. The cluster of *Synechocystis sp.* (*cyXX*) shows this property quite nicely. Similar clusters are observed for *D. radiodurans* (*drXX*) and *M. tuberculosis* (*mtXX*). On the contrary a more complex relationship between the *Neisseria*, *P. aeruginosa*, *E. coli* and *Y. pestis* can be noticed.

Independent of the distinction between orthologs and paralogs (parameter f in Equ. 1 — phylogeny for $f \to 0$ not shown) two diverse clusters are represented in the phylogenetic graph (Fig. 3 with almost identical topology: Cluster I consists of pathways *ec00, ng00, ng01, nm00, nm01, pa00, pa01* and *yp00*. Cluster II includes pathways *ec04, ng01, ng11, nm01, nm11, pa02, pa12* and *yp01*. *Neisseria* (*ngXX, nmXX*) are very close related and always show up as pairs in phylogenies. Quite surprising is the close similarity of *E. coli* (*ecXX*) to *Y. pestis* (*ypXX*). Studying corresponding

Table 2: Examples of Electron Transfer Pathways

	Pathway	#[a]	Organisms
1	2-Oxoglutarate – 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 -- NG NM -- -- -- CY PG -- -- -- ML MT MG MP ST PN EF BS
		•[b]	SC CE
3	NADH – Oxidized rubredoxin Electron Transport	2	AG TH -- MJ -- -- -- -- -- PA -- -- -- -- -- -- -- -- -- -- -- MT -- -- -- -- -- -- -- --
4	NADH – FAD Electron transport (plasma membrane)	1	AG TH -- -- AA DR EC YP HI PA NG NM RC HP CJ CY PG -- -- -- -- MT -- -- ST -- 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 -- -- PA -- -- RC -- -- CY -- -- -- -- -- MT -- -- -- -- -- -- --
8	NADH, H^+ – O_2, $H^+_{\text{periplasma}}$ Electron transport (ubiqinone, cytochrome *bo*) (plasma membrane)	19	-- -- -- -- AA -- EC -- -- PA -- -- RC -- -- CY -- -- -- -- -- MT -- -- -- -- -- -- --
9	$NADH_{\text{mitochond. matrix}}$ – O_2, mitochond. matrix Electron transport (mitochondrial inner membrane, mitochondrial intermembrane space)	47	-- SC CE
10	NADH – NAD^+ Electron Transport (malate, aspartate)	2	AG TH PH MJ AA DR EC YP HI PA -- -- -- -- CJ CY -- -- TP CA ML MT -- -- -- -- EF BS
		•	SC CE
11	NADH – O_2 Electron transport (plasma membrane)	1	AG TH PH MJ AA DR EC YP HI PA NG NM RC HP CJ CY PG -- -- -- -- MT -- -- ST -- EF BS -- --
12	NADPH – Hg^{2+} Electron transport	1	-- -- -- -- -- -- EC -- -- -- -- -- -- -- -- CY -- -- -- -- -- MT -- -- -- PN -- -- -- --
13	NADPH – Oxidized flavodoxin Electron transport (plasma membrane)	2	-- -- -- -- AA -- EC YP HI PA -- -- RC HP CJ CY PG -- -- -- -- -- -- -- -- -- EF BS -- --
14	NADPH – Oxidized ferredoxin Electron transport	2	-- -- -- -- -- DR EC YP -- PA NG NM -- -- -- CY -- -- -- -- -- MT -- -- -- -- EF BS -- --
15	NADPH – Oxidized thioredoxin Electron transport	2	AG TH -- MJ AA DR EC YP HI PA NG NM -- HP CJ CY PG BB TP CA ML MT MG MP ST PN EF BS SC --

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

pairs between cluster I and II indicates diversity in ferredoxins rather that in reductases. This gives hint to a higher demand of utilizing ferredoxins for different tasks by nature than ferredoxin–NADPH reductases. This issue will be discussed in the next section.

A similar scenario can be observed in the case of the *mtXX* cluster. Pathways *mt01, mt11* and *mt02, mt12* cluster and differ significantly from *mt00, mt10*. This might gives hint to a specialization of the pathway within the organism.

4.2 Pathways with Ferredoxin as Functional Role

Out of these approx. 50 pathways where ferredoxin plays a significant functional role 7 pathways are chosen for further investigations — the remaining non-used pathways are either assigned in one organism only or are absent at all due to missing functional roles in all organisms:

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

Fig. 4 shows a phylogenetic graph of 418 representations of these 8 pathways with a confidence level of $t = 0.5$. Thus 50% of functional roles have to be present. Otherwise the distinct pathway is not considered. Pathways of the same organism are found in one cluster, although a detailed analysis (not shown) discloses a richer relationship. Similar to the

Table 3: List of Codes referring to ORF-names with corresponding Functional Roles

Code	Reductase	Ferredoxin
cy00	RCY25830	RCY45187
cy01	RCY25830	RCY08865
cy02	RCY25830	RCY02872
cy03	RCY25830	RCY15530
cy04	RCY25830	RCY13929
cy05	RCY25830	RCY13110
cy06	RCY25830	RCY46991
cy07	RCY25830	RCY42273
dr00	RDR02099	RDR01803
dr01	RDR02099	RDR01783
ec00	REC06248	REC02502
ec01	REC06248	REC06625
ec02	REC06248	REC01352
ec03	REC06248	REC00044
ec04	REC06248	REC05525
mt00	RMT05972	RMT02674
mt01	RMT05972	RMT00975
mt02	RMT05972	RMT04402
mt10	RMT00705	RMT02674
mt11	RMT00705	RMT00975
mt12	RMT00705	RMT04402
ng00	RNG00591	RNG01106
ng01	RNG00591	RNG00533
ng10	RNG00984	RNG01106
ng11	RNG00984	RNG00533
nm00	RNM01731	RNM00363
nm01	RNM01731	RNM00662
nm10	RNM00963	RNM00363
nm11	RNM00963	RNM00662
pa00	RPA07749	RPA01015
pa01	RPA07749	RPA08046
pa02	RPA07749	RPA01568
pa03	RPA07749	RPA07726
pa10	RPA05251	RPA01015
pa11	RPA05251	RPA08046
pa12	RPA05251	RPA01568
pa13	RPA05251	RPA07726
yp00	RYP02807	RYP00405
yp01	RYP02807	RYP01051

observation reported in the previous sections pathways of *E. coli* and *Y. pestis* are close related to each other. In the same "super-cluster" but weaker related to the former species a second cluster is formed by *M. tuberculosis*, *P. aeruginosa* and *Synechocystis sp.*. For the remaining organisms (*A. aeolicus*, *A. fulgidus*, *B. subtilis*, *C. acetobutylicum*, and *D. radiodurans*) the significance of similarity is speculative.

5 Discussion

The accessibility of numerous complete genomes as phylogenetic diverse representatives of all three known kingdoms will have and already has dramatic effects on the strategy on analyzing completely sequenced genomes. Not only "making the next step" towards a representation of higher-level functional components per organism as performed by Overbeek *et al.*[28] but also developing methods to compare different representations between organisms (and within) to each other are new goals to aim at.

The method presented here is a first approach to relate pathways to each other by using explicit sequence information. It finds application in[3] the important class of metabolic networks related to electron transport. Studying this evolutionary old network expose some mechanisms how nature evolves functionality on the level of metabolic networks. According to Woese [51] metabolic genes are among the most modular in the cell, and so, their genes are expected to travel laterally, even today. Thus especially single-cellular organisms adapt quite easily their metabolisms to their needs. Regarding electron transport this means a diverse presence of member-pathways across organisms. Essential pathways important for respiratory purposes are observed in all three kingdoms whereas, e.g., NADPH–Hg^{2+} electron transport is only found in *E. coli*, *S. pneumoniae*, and *Synechocystis sp.* Pathways utilizing ferredoxin confirm these observations. Conserved pathways are well-defined per organism and present in all three kingdoms (Fig. 4). Whereas pathways assigned for Bacteria only, such as ferredoxin – NADPH reductase pathways, show high versatility and controversial phylogenies (Fig. 3).

Extending the approach of biological evolution Carl Woese asked about the minimal requirement for a living cell. On the level of genes *Mycoplasmae* are close to this minimal assumption. Regarding a autopoiesic closure of a free living organism, providing all metabolites by itself parasitic *Mycoplasmae* are far away from these requirements. Thus "constructing" such a free living organism which presumably was a member of the universal ancestral cells [51] of all known living creatures (Archaea, Bacteria and Eucarya) is a very attractive question to address. Indeed, such a hypothesis has to be verified by experiments.

[3] but is not restricted to

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 Institute of Health (NIH PHS 5 P41 RR05969), the National Science Foundation (NSF BIR 94-23827 EQ), and the Beckman Institute.

References

[1] F. Blattner, G. P. 3rd, C. Bloch, N. Perna, et al. *Science*, 277:1435–1474, 1997.

[2] C. Bult, O. White, G. Olsen, L. Zhou, et al. *Science*, 273:1058–1073, 1996.

[3] S. Cole, R. Brosch, J. Parkhill, T. Garnier, et al. *Nature*, 393:537–544, 1998.

[4] M. Dayhoff, R. Schwartz, and B. Orcutt. volume 5 of *Atlas of Protein Sequence and Structure*, pages 345–352, Washington. National Biomedical Research Foundation.

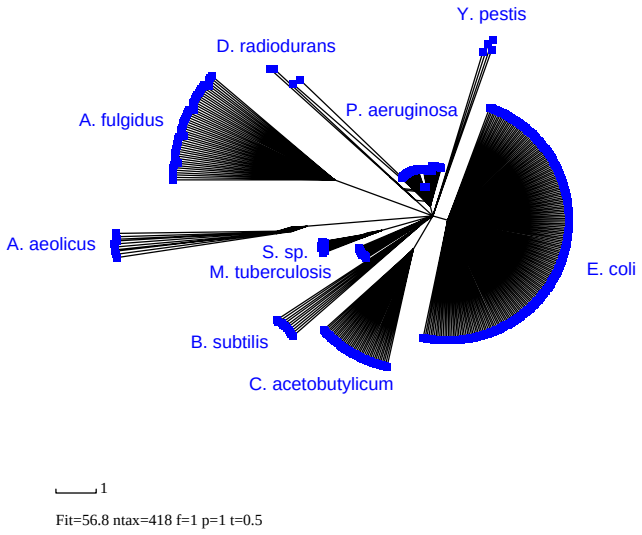

Figure 4: Ferredoxin-related pathways: 418 different representations of ferredoxin-related pathways are shown: 10 for *A. aeolicus*, 48 for *A. fulgidus*, 10 for *B. subtilis*, 42 for *C. acetobutylicum*, 4 for *D. radiodurans*, 200 for *E. coli*, 12 for *M. tuberculosis*, 72 for *P. aeruginosa*, 16 for *Synechocystis sp.*, and 4 for *Y. pestis*. The distance matrix is created with parameters $f = 1$, $p = 1$, and $t = 0.5$ and drawn by the program *SPLITSTREE2*.

[5] G. Deckert, P. Warren, T. Gaasterland, W. Young, et al. *Nature*, 392:353–358, 1998.

[6] J. Felsenstein. http://evolution.genetics.washington.-edu/phylip.html, 1997.

[7] W. Fitch. Distinguishing homologous from analogous proteins. 19:99–113, 1970.

[8] R. Fleischmann, M. Adams, O. White, R. Clayton, et al. *Science*, 269:496–512, 1995.

[9] C. Fraser, S. Casjens, W. Huang, G. Sutton, et al. *Nature*, 390:580–586, 1997.

[10] C. Fraser, J. Gocayne, O. White, M. Adams, et al. *Science*, 270:397–403, 1995.

[11] C. Fraser, S. Norris, G. Weinstock, O. White, et al. *Science*, 281:375–388, 1998.

[12] Genome Therapeutics Corporation. http://www.genomecorp.-com/genesequences/clostridium/clospage.html.

[13] A. Goffeau et al. *Nature*, 387:5–105, 1997. (Suppl.).

[14] J. Haldane. *Rationalist Ann.*, 148:3, 1928.

[15] H. Hartman. *J. Mol. Evol.*, 4:359–370, 1975.

[16] S. Henikoff and J. Henikoff. *Proc. Natl. Acad. Sci.*, 89:10915–10919, 1992.

[17] R. Himmelreich, H. Hilbert, H. Plagens, E. Pirkl, B. Li, and R. Herrmann. *Nucl. Acids Res.*, 24:4420–4449, 1996.

[18] D. H. Huson. *Bioinformatics*, 14:68–73, 1998.

[19] T. Kaneko, S. Sato, H. Kotani, A. Tanaka, et al. *DNA Res.*, 3:109–136, 1996.

[20] Y. Kawarabayasi et al. *DNA Res.*, 5:55–76, 1998.

[21] H. Klenk, R. Clayton, J. Tomb, O. White, et al. *Nature*, 390:365–370, 1997.

[22] F. Kunst, N. Ogasawara, I. Moszer, A. Albertini, et al. *Nature*, 390:249–256, 1997.

[23] F. Lipmann. pages 259–280. Academic, New York, 1965.

[24] E. Meléndez-Hevia, T. G. Waddell, and M. Cascante. *J. Mol. Evol.*, 43:293–303, 1996.

[25] S. L. Miller. *Science*, 117:528–529, 1953.

[26] A. I. Oparin. World, Cleveland, 1967. also published in: Proiskhozhdenie Zhizny. IZD Moskovishii Rabochii, Mowcow, 1924.

[27] L. E. Orgel. *J. Mol. Biol.*, 38:381–383, 1968.

[28] R. Overbeek, N. Larsen, W. Smith, N. Maltsev, and E. Selkov. *Gene*, 191:GC1–GC9, 1997.

[29] R. Overbeek, G. Pusch, M. Dsouza, N. Larsen, E. S. jr., E. Selkov, and N. Maltsev. http://wit.mcs.anl.gov/WIT2/wit.html, 1998. WIT2.

[30] K. R. Popper. Routledge & Kegan Paul, London, 1957.

[31] K. R. Popper. Routledge & Kegan Paul, London, 1963.

[32] D. Smith, L. Doucette-Stamm, H. L. C. Deloughery C, et al. *J. Bacteriolog.*, 179:7135–7155, 1997.

[33] R. Tatusov, E. Koonin, and D. Lipman. *Science*, 278:631–637, 1997.

[34] R. Tatusov, A. Mushegian, P. Bork, N. Brown, W. Hayes, M. Borodovsky, K. Rudd, and E.V. *Curr. Biol.*, 6:297–291, 1996.

[35] The Institute of Genome Research. ftp://ftp.tigr.org/-pub/data/d_radiodurans/.

[36] The Institute of Genome Research. ftp://ftp.tigr.org/-pub/data/n_meningitidis/.

[37] The Institute of Genome Research. ftp://ftp.tigr.org/pub/data/s_pneumoniae/.

[38] The Institute of Genome Research. ftp://ftp.tigr.org/-pub/data/e_faecalis/.

[39] The Institute of Genome Research and Forsyth Dental Center. http://www.forsyth.org/pggp/.

[40] The Sanger Centre. http://www.sanger.ac.uk/-Projects/Y_pestis/.

[41] The Sanger Centre. ftp://ftp.sanger.ac.uk/pub/-pathogens/cj/.

[42] The Sanger Centre. ftp://ftp.tigr.org/pub/data/-m_tuberculosis/.

[43] The Sanger Centre. http://www.sanger.ac.uk/-Projects/C_elegans/.

[44] J. Thompson, D. Higgins, and T. Gibson.

[45] J.-F. Tomb, O. White, A. Kerlavage, R. Clayton, et al. *Nature*, 388:539–547, 1997.

[46] University of Chicago and Institute of Molecular Genetics, CSFR. http://capsulapedia.uchicago.edu/.

[47] University of Oklahoma. ftp://ftp.genome.ou.edu/-pub/gono.

[48] University of Oklahoma. http://dna1.chem.uoknor.-edu/strep.html.

[49] University of Washington AND PathoGenesis. http://www.pseudomonas.com/.

[50] G. Wächtershäuser. *Proc. Natl. Acad. Sci. USA*, 87:200–204, 1990.

[51] C. Woese. *Proc. Natl. Acad. Sci. USA*, 95:6854–6859, 1998.