CO-EVOLUTION OF METABOLISM AND PROTEIN SEQUENCES

MORITZ SCHÜTTE[1] NIELS KLITGORD[2]
schuette@mpimp-golm.mpg.de niels@bu.edu

DANIEL SEGRÈ[2,3] OLIVER EBENHÖH[1,4,5,6]
dsegre@bu.edu ebenhoeh@abdn.ac.uk

[1] *Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam–Golm, Germany*
[2] *Boston University, Bioinformatics Program, 24 Cummington Street, Boston, MA 02215, USA*
[3] *Boston University, Departments of Biology and Biomedical Engineering, 24 Cummington Street, Boston, MA 02215, USA*
[4] *Potsdam University, Institute of Biochemistry and Biology, Karl–Liebknecht–Straße 24–25, 14476 Potsdam–Golm, Germany*
[5] *Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen, AB24 3UE, UK*
[6] *Institute of Medical Sciences, Foresterhill, University of Aberdeen, Aberdeen, AB25 2ZD, UK*

The set of chemicals producible and usable by metabolic pathways must have evolved in parallel with the enzymes that catalyze them. One implication of this common historical path should be a correspondence between the innovation steps that gradually added new metabolic reactions to the biosphere-level biochemical toolkit, and the gradual sequence changes that must have slowly shaped the corresponding enzyme structures. However, global signatures of a long-term co-evolution have not been identified. Here we search for such signatures by computing correlations between inter-reaction distances on a metabolic network, and sequence distances of the corresponding enzyme proteins. We perform our calculations using the set of all known metabolic reactions, available from the KEGG database. Reaction-reaction distance on the metabolic network is computed as the length of the shortest path on a projection of the metabolic network, in which nodes are reactions and edges indicate whether two reactions share a common metabolite, after removal of cofactors. Estimating the distance between enzyme sequences in a meaningful way requires some special care: for each enzyme commission (EC) number, we select from KEGG a consensus set of protein sequences using the cluster of orthologous groups of proteins (COG) database. We define the evolutionary distance between protein sequences as an asymmetric transition probability between two enzymes, derived from the corresponding pair-wise BLAST scores. By comparing the distances between sequences to the minimal distances on the metabolic reaction graph, we find a small but statistically significant correlation between the two measures. This suggests that the evolutionary walk in enzyme sequence space has locally mirrored, to some extent, the gradual expansion of metabolism.

Keywords: metabolism; networks; evolution; protein sequence; enzyme; EC number; KEGG; COG.

1. Introduction

The evolutionary walk from an early proto-metabolism to the current biochemical pathways must have been shaped by innovations concurrently involving enzymes and chemical compounds [13]. While it is generally assumed that todays enzymes have evolved from a few ancestors that were able to catalyze the first reactions, a clear correspondence between the evolution of metabolic functions and their catalyzing enzymes remains to be established. The evolution of metabolic pathways has been addressed by several competing models, including the patchwork model [6, 27], and models of forward [3] and retrograde [4] evolution. In addition, several studies have addressed the relation between sequence homology and protein function. These studies have been widely used for the prediction of protein functions [19, 24, 26] associated with newly sequenced genes and for the analysis of relations between sequences and functions in Gene Ontology terms [7].

To date, there exists only few studies of evolutionary relation between enzyme sequence homology and distance on the metabolic reaction network. Most of these studies are restricted to networks of single organisms. These works show a possible link between homology and metabolic network distance. For example in *E.coli* [15, 18], it was found that homologous protein sequences are more likely to be found in close vicinity on the metabolic network than what expected by chance and the same trend has been confirmed for protein-protein interactions [5]. Similarly, in studies of yeast [25] a link has been found between the metabolic network structure and enzymatic evolution. Since single enzymes or even entire operons have been copied and changed to fulfill new functions, high promiscuity in the locality of catalyzing enzymes complicates the search for evolutionary relations [11, 20, 21].

Here, we test the hypothesis that the global scope of metabolism has evolved in parallel with the enzymes that make up the network. To investigate this hypothesis, we have taken a large-scale approach using the entire set of reactions from the KEGG [8–10] reaction database. If our hypothesis holds true, then we expect to find that a similarity between protein sequences should be reflected by a closeness in the metabolic reaction network. In order to test our hypothesis we define different measures of sequence distances, both symmetric and asymmetric, based on a reciprocal pair-wise BLAST analysis. Asymmetry becomes important if two sequences of different lengths are compared, because it is more likely that a shorter sequence is the evolutionary child of a longer sequence, than the other way around. These distances are then compared with the distances on an enzyme-enzyme network of chemical reactions that we construct from the KEGG database. As KEGG provides multiple protein sequences per reaction, we first choose a consensus sequence set based on the cluster of orthologous groups of proteins (COG) database [22, 23] that greatly reduces the sequence space we must analyze. Our analysis supports our hypothesis, showing that enzymes that are close in the reaction network are enriched for sequence similarity. The observed trend is small but significant against a simulated control.

2. Protein Sequence Distances and Enzyme Distances

To address our hypothesis on a large scale, we use the entirety of reactions from the KEGG database. We construct a reaction-centric network where reactions are nodes and two reactions are linked if they involve a common metabolite. Some metabolites, such as cofactors, participate in a variety of reactions and thus produce short-cuts that do not carry actual fluxes. To account for this, we extract the cofactor pairs ATP/ADP, NAD^+/NADH, NADP^+/NADPH, and CoA/Acetyl-CoA from the reactions where they appear as pairs [5]. Furthermore, we delete highly abundant molecules, H_2O, H^+, and O_2, which appear in more than 500 reactions, from the reaction set. Every reaction is catalyzed by one or more enzymes given in terms of enzyme commission (EC) numbers, exceptions in the reaction set are spontaneous reactions or those for which the catalyst is not known. We link the reactions to the enzyme sequence using the EC numbers. Such, we transform the reaction network to an enzyme-enzyme network. As we only know sequences for roughly half of all EC numbers we still use all reaction links but only calculate shortest paths between enzymes for which we know the sequences of starting and ending enzymes. For this purpose we use the Dijkstra algorithm [1]. Intermediate reactions without catalyzing enzyme, like spontaneous reactions, or with an enzyme without known sequence, are still counted as a step in the distance.

The distance in protein sequence space requires more elaboration. We use the Blastp in the *bl2seq* program to obtain pair-wise comparisons between sequences [28]. A rather intuitive measure obtained from blast is the *identity* measure counting how many residues on a certain aligned fragment are identical between the two sequences. To get a comparable scoring measure we need to normalize the identities by the sequence lengths. It is important to note that we are interested in interpreting such a measure as an estimate of the probability that one sequence has evolved from the other. Since asymmetry can play a crucial role in the transition from one sequence to the other [16], we will take into account not only the total length, but also the difference between the two lengths. We define $\mathcal{I} = \sum_i I_i$ as the sum of the identities of every found alignment. Based on this measure, we define three different estimates of the probability Γ_{AB} that a sequence B (with length L_B) has evolved from a sequence A (with length L_A):

$$\Gamma_{AB}^{(\mathrm{id})} = 1 - \frac{2 \cdot \mathcal{I}}{L_A + L_B}\,, \tag{1}$$

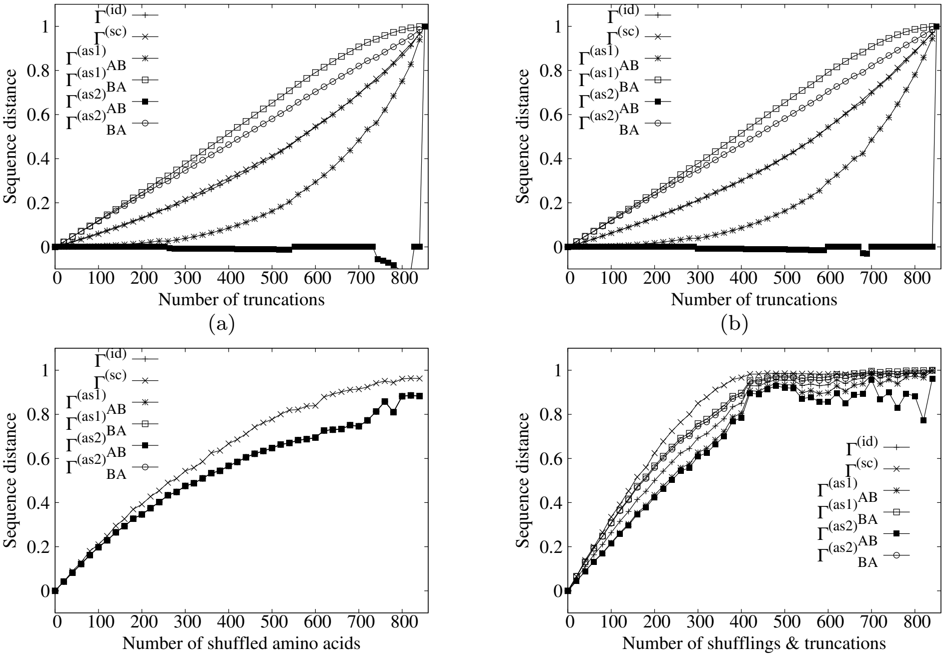
$$\Gamma_{AB}^{(\mathrm{as1})} = 1 - \frac{2 \cdot \mathcal{I}}{L_A + L_B}\left[1 + \frac{(L_A - L_B)}{(L_A + L_B)}\right]\,, \tag{2}$$

$$\Gamma_{AB}^{(\mathrm{as2})} = 1 - \frac{\mathcal{I}}{L_B}\,. \tag{3}$$

The second measure follows from a Taylor expansion of Eq. (1). We consider the length difference as a small correction to the mean of the lengths: $(L_A + L_B)/2 \Longrightarrow \bar{L} - \Delta L/2$: $(\bar{L} - \Delta L/2)^{-1} \approx 1/\bar{L} + \Delta L/2\bar{L}^2$. If the directionality of a linear pathway is known, an asymmetric distance like Eqs. (2)–(3) could in principle be used to test for retrograde or forward evolution [3, 4]. In addition to Eqs. (1)–(3) we use the *score* values of the best hit obtained from BLAST and normalize it by the score of the sequence with itself:

$$\Gamma^{(\mathrm{sc})}_{AB} = 1 - \frac{2 \cdot score(A,B)}{score(A,A) + score(B,B)} \,. \tag{4}$$

To evaluate the utility of these proposed measures, we tested how they perform on simple pairs of sequences, computed so as to simulate a gradual transition from exact identity to large differences, see Fig. 1. For the purposes of our simulation, we chose one random but rather long (839 amino acids) test sequence (*glo:Glov_1829 integral membrane sensor signal transduction histidine kinase (EC:2.7.13.3)*). These experiments simulate some of the possible scenarios of sequence changes during enzyme evolution. In the first two experiments, we took a copy of the sequence and iteratively cut off amino acids from either start or end of the copy. This reduced sequence is then blasted against the original one, Figs. 1(a) and (b). In experiment three we shuffled increasingly more amino acids in the copy and blasted against the



(a) (b) (c) (d)

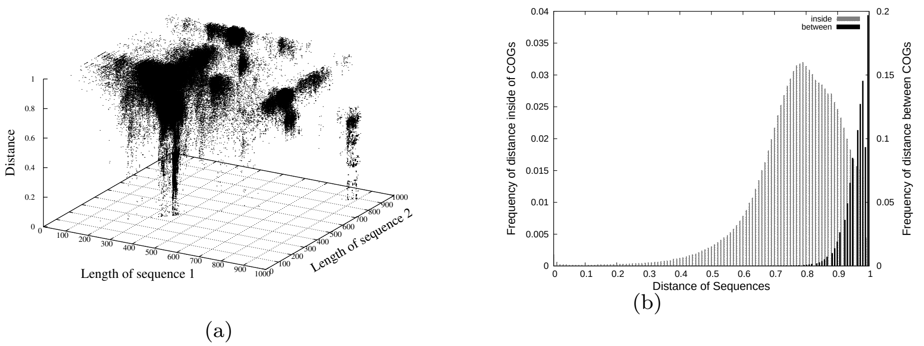
Fig. 1: Simulation of four evolutionary scenarios. We take a test sequence, change it and blast it against an original copy. In (a) and (b) we iteratively delete amino acids from the start (a) or end (b) of the test sequence. (c): The amino acids are increasingly shuffled. (d) combines (a) and (c): we shuffle and truncate the sequence.

original sequence, (c). The last experiment, (d), is a combination of shuffling and length reduction. Here, we observe an edge at around 0.9 distance. Below this value the behavior seems random.

These experiments show that $\Gamma_{AB}^{(\mathrm{as2})}$, Eq. (3), is not an appropriate measure. Specifically, in the first two experiments this measure reaches negative values, resulting from the fact that we sum all identities. Because the aligned fragments become very small, it is likely that the same fragment is found twice or that an overlap between two matches is found in the original copy. Since the measure is only normalized by its own length, this may result in negative values. A similar artifact is also observed in experiment four, where the measure shows a false positive agreement when sequences have been highly shuffled, and greatly truncated. Thus this measure will not be used for further investigations.

3. Consensus Sequence Set

We characterize an enzyme by two features: its sequence, and its function. These features can, to some degree, vary independently: One enzyme may have multiple functions, or conversely, one specific function can be performed by multiple sequences [2]. Our goal is to investigate whether the evolutionary distance of sequences relates to their functional distance as determined by the metabolic reaction network where the reactions are defined in terms of EC numbers. Since each EC number can be associated with a rather large number of sequences, we define a consensus set of sequences serving as representatives of the specific function. In total, the KEGG database contains approximately 750000 sequences that contain one or more EC numbers in their description. As can be seen in Fig. 2(a), the distribution



(a)

(b)

Fig. 2: (a) All pair-wise blasts of the alcohol dehydrogenase (EC 1.1.1.1) sequences scored by the distance Eq. (1). Two apparent clusters of small distance are observable around length 300 and 900. (b) Benchmark of COG inner distance versus the distances between representatives of each COG. The cut-off of 0.9 to differ seems reasonable.

of the number of sequences per EC number can vary quite a bit in KEGG. This has been greatly reduced in our consensus sequence set. For example, the sequence by sequence distances for alcohol dehydrogenase, EC 1.1.1.1, are shown in Fig. 2(a). The lengths vary by a factor three and even the sequences of similar lengths need not at all be similar. Even the set of 188 alcohol dehydrogenase sequences that are all 350 amino acids in length varies from completely identical to completely different with a mean of $\Gamma^{(\mathrm{id})} = 0.72 \pm 0.12$. One can observe several distinct clusters of high similarity in Fig. 2(a).

To simplify our task of selecting representative sequences, we utilize the COG database [22, 23] that clusters proteins by their function based on homology. Every COG contains between a few and a few hundred sequences that are related by a duplication or speciation event. We pick the longest sequence of every COG as the representative of this COG [12]. In order to cover as many as possible EC numbers we cluster the remaining KEGG sequences by a very simple procedure. We group the sequences by EC number and perform all pairwise blasts dropping those sequences that have a 0.9 or higher distance according to Eq. (1). This loose cut-off is justifiable using the COGs as a benchmark. We calculated all inner-COG distances and compared them with the distances between the representatives of every COG, see Fig. 2(b). A second argument for this cutoff comes through the

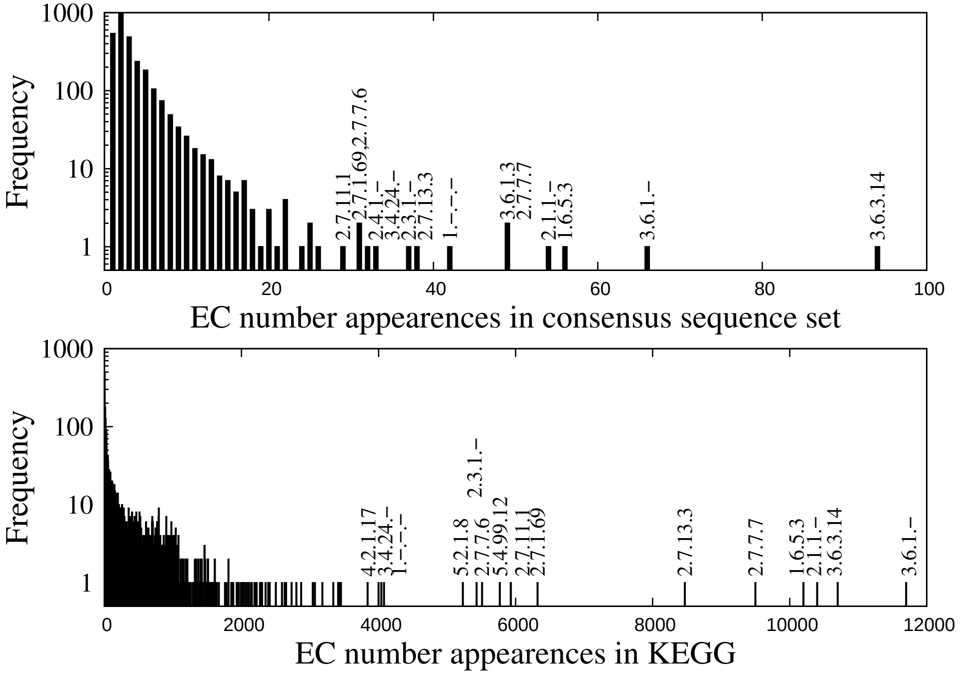


Fig. 3: Distribution of EC numbers how often they appear in descriptions of different protein sequences. Top in the consensus set, bottom in the KEGG sequences.

result of Fig. 1 (d) where all curves show a somewhat random behavior for distances larger than 0.9.

Following this procedure we obtain a consensus set of 8123 sequences coding for 2821 EC numbers. Fig. 3 shows a histogram of the frequency of a certain EC number with that it appears in descriptions of different sequences. The top bar graph shows the distribution in the final consensus set and the bottom one in the starting set from all KEGG sequences. There is a good agreement between the ranking by EC number with Pearson correlation 0.81 and Spearman Rank correlation 0.53.

4. Correlation of Network and Sequence Distances

We use the previously defined consensus sequence set to analyze a relationship between distance on the enzyme-enzyme graph and the sequences of the enzymes. We use a sample of 4.8 million shortest paths which all start and end with a sequenced enzyme.

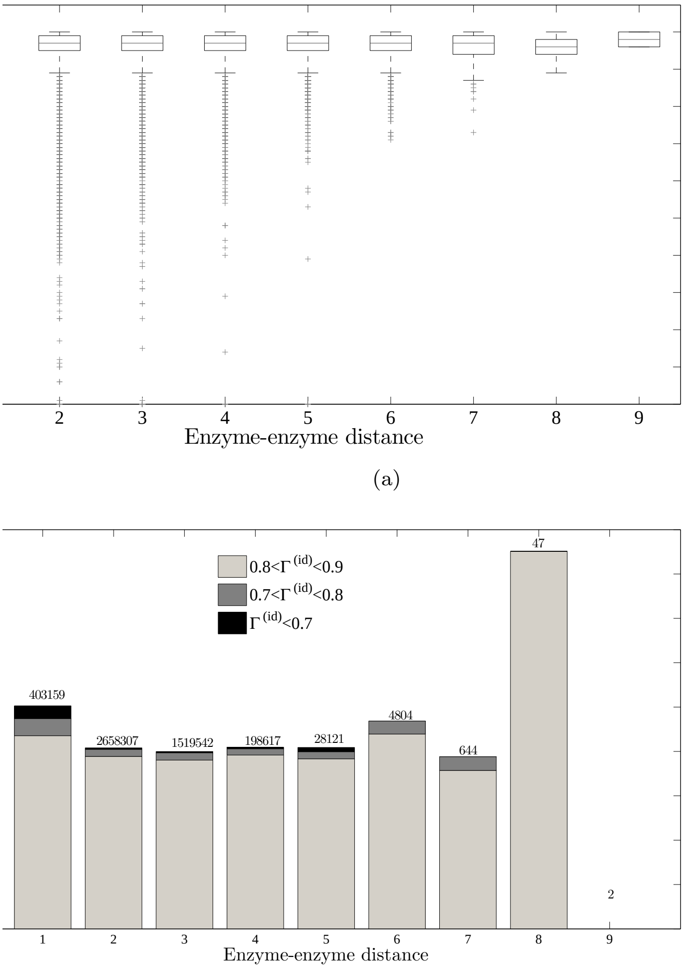
Figure 4(a) shows a boxplot of the correlation using the measure $\Gamma^{(\mathrm{id})}$, Eq. (1). We see a highly significant but small correlation that is mirrored by a trend seen in the outliers where similar enzymes tend to be closer in the network. In order to test the results against the null hypothesis that they appeared by chance, we calculated the p-value which is based on the sample size. We performed a second control calculation utilizing a permutation test where we shuffled the sequence distances to generate a random set. For this control simulation the correlation is completely lost and we observe high p-values, see Tab. 1.

To further quantify the observation, we analyze the results sorting them by particular path-lengths, Fig. 4(b). Neighboring enzymes tend to show higher similarity on the sequence level. For enzyme-enzyme distance 1, the relative proportion of distances below 0.8 and 0.7 is enriched. The bar on distance 8 is the highest but it represents only a sample of four similar enzymes of 47.

Table 1: Comparison of correlations between enzyme sequence distances and distances obtained form the enzyme-enzyme graph. In the control measurement we shuffle the enzyme distance matrix and repeat the simulation. The distances in the sequence space were calculated with the measures described in section 2 (sample size: 4.8 million shortest paths). Although the correlations are very low, they are highly significant in comparison with the control data.

measure	$\Gamma^{(\mathrm{id})}_{AB}$	control to $\Gamma^{(\mathrm{id})}_{AB}$	$\Gamma^{(\mathrm{as1})}_{AB}$	control to $\Gamma^{(\mathrm{as1})}_{AB}$	$\Gamma^{(\mathrm{sc})}_{AB}$	control to $\Gamma^{(\mathrm{sc})}_{AB}$
correlation	0.0127	0.0002	0.0170	-0.0003	0.0035	0.0004
p-value	10^{-171}	0.7270	10^{-304}	0.4931	10^{-14}	0.4189

Table 1 compares the results for different measures of sequence similarity. The asymmetric measure Eq. (2) yields the highest correlation and significance in calculating the sequence distance. For every enzyme pair we chose the more similar value of the asymmetric sequence distance $\Gamma^{(\mathrm{as1})}$.



(a)

(b)

Fig. 4: (a) Boxplot for the correlation between distances on the enzyme-enzyme graph and the measure $\Gamma^{(\mathrm{id})}$ for the sequences. The correlation is very low, 0.0127, but significant, see Tab. 1. (b) Sequence distances sorted by particular enzyme-enzyme distances. The plot shows the fraction of enzyme sequence pairs within a certain distance of $\Gamma^{(\mathrm{id})}$ compared to all enzymes found in the distance on the network. The small numbers on top of the bars represent the total number of enzymes found in the particular distance. For neighboring enzymes we observe a higher fraction of enzymes with similar sequences.

5. Conclusion

We have investigated the evolutionary relation of enzyme sequences and their distance in the metabolic network on a large-scale using a consensus sequence set from the entire KEGG database. However, the choice of the consensus set is strongly biased by two aspects of the used database: the choice of sequenced organisms and the accuracy in investigating proteins. A large number of redundant sequences is due to the variety of organisms whose proteomes are sequenced, and whose function was assigned via homology. The sequences in the consensus set come from 27 animals, 6 plants, 21 fungi, 535 bacteria, 51 archea and 17 protists and these result in 1395, 274, 585, 4974, 598, 297 sequences from the particular kingdoms. The variability in plant-specific enzymes might be underestimated as only a few model plants are well investigated. For the carbon-fixating enzyme RuBisCO (EC 4.1.1.39) we obtain only two different sequences from bacteria *Synechocystis* and *Anabaena*. The majority of organisms are bacteria for which lateral gene transfer is an important factor [14, 17]. By the use of the COG database we might neglect this possibility of sequence change. The second bias appears through the way proteins are investigated. As an example we examine ATP synthase, EC 3.6.3.14. This protein is the second most abundant in KEGG and the most abundant in the consensus set, Fig. 3. It catalyzes only one reaction, $\mathrm{ATP} + \mathrm{H_2O} + \mathrm{H}^+_{in} \rightleftharpoons \mathrm{ADP} + \mathrm{phosphate} + \mathrm{H}^+_{out}$. This reaction is essential in most organisms and frequently investigated. We thus capture the variability of sequences for known enzymes but do not grasp it for less known ones.

We have observed a weak correlation between enzymes that are neighbors in the graph representing metabolism, and the corresponding sequences. Our finding extends to an all-organism level results previously obtained for single organisms [15] and using protein-protein interactions [5]. The correlation detected indicates a certain degree of co-evolution between the topology of metabolism and its enzyme capabilities. We envisage that future simulations of the evolutionary expansion of metabolism, possibly employing our proposed asymmetric measure of sequence distance, could shed more insight into the nature of this correlation.

6. Acknowledgments

We acknowledge financial support from the International Research Training Group *Genomics and Systems Biology of Molecular Networks* IRTG 1360 (MS), the German Federal Ministry of Education and Research, Systems Biology Research Initiative *Go*FORSYS, the Scottish University Life Science Alliance SULSA (OE), the NASA Astrobiology Institute, and the US Department of Energy (NK, DS).

References

[1] Dijkstra, E.W., A note on two problems in connexion with graphs, *Numerische Mathematik*, 1: 269–271, 1959.

[2] Galperin, M.Y., Walker, D.R., Koonin, E.V., Analogous enzymes: independent inventions in enzyme evolution, *Genome Res.*, 8: 779–790, 1998.

[3] Granick, S., Speculations on the origins and evolution of photosynthesis, *Ann N Y Acad. Sci.*, 69: 292–308, 1957.

[4] Horowitz, N.H., On the evolution of biochemical syntheses, *PNAS*, 31(6): 153–157, 1945.

[5] Huthmacher, C., Gille, C., Holzhütter, H.G., A computational analysis of protein interactions in metabolic networks reveals novel enzyme pairs potentially involved in metabolic channeling, *J. Theor. Biol.*, 252: 456–464, 2008.

[6] Jensen, R.A., Enzyme recruitment in evolution of new function, *Annu. Rev. Microbiol.*, 30: 409–435, 1976.

[7] Joshi, T., Xu, D., Quantitative assessment of relationship between sequence similarity and function similarity, *BMC Genomics*, 8: 222, 2007.

[8] Kanehisa, M., Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res.*, 28: 27–30, 2000.

[9] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M., From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Res.*, 34: 354–357, 2006.

[10] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y., KEGG for linking genomes to life and the environment, *Nucleic Acids Res.*, 36: 480–484, 2008.

[11] Khersonsky, O., Roodveldt, C., Tawfik, D.S., Enzyme promiscuity: evolutionary and mechanistic aspects, *Curr Opin Chem Biol.*, 5: 498–508, 2006.

[12] Krause, A., Stoye, J., Vingron, M., Large scale hierarchical clustering of protein sequences, *BMC Bioinformatics*, 6: 15, 2005.

[13] Lazcano, A., Miller, S.L., On the origin of metabolic pathways, *J. Mol. Evol.*, 49: 424–431, 1999.

[14] Lercher, M.J., Pàl, C., Integration of horizontally transferred genes into regulatory interaction networks takes many million years, *Mol. Biol. Evol.*, 25(3): 559–567, 2008.

[15] Light, S., Kraulis, P., Network analysis of metabolic enzyme evolution in *Escherichia coli*, *BMC Bioinformatics*, 5: 15, 2004.

[16] Notebaart, R.A., Kensche, P.R., Huynen, M.A., Dutilh, B.E., Asymmetric relationships between proteins shape genome evolution, *Genome Biol.*, 10: R19, 2009.

[17] Pàl, C., Papp, B., Lercher, M.J., Adaptive evolution of bacterial metabolic networks by horizontal gene transfer, *Nat Genet.*, 17(12): 1372–1375, 2005.

[18] Rison, S.C., Teichmann, S.A., Thornton, J.M., Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*, *J. Mol. Biol.*, 318(3): 911–932, 2002.

[19] Rost, B., Enzyme function less conserved than anticipated, *J. Mol. Biol.*, 26: 595–608, 2002.

[20] Schmidt, S., Sunyaev, S., Bork, P., Dandekar, T., Metabolites: a helping hand for pathway evolution?, *Trends Biochem Sci.*, 6: 336–341, 2003.

[21] Spirin, V., Gelfand, M.S., Mironov, A.A., Mirny, L.A., A metabolic network in the evolutionary context: multiscale structure and modularity, *Proc Natl Acad Sci*, 103(23): 8774–8779, 2006.

[22] Tatusov, R.L., Koonin, E.V., Lipman, D.J., A genomic perspective on protein families, *Science*, 278: 631–637, 1997.

[23] Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A., The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, 4: 41, 2003.

[24] Tian, W., Skolnick, J., How well is enzyme function conserved as a function of pairwise sequence identity?, *J. Mol. Biol.*, 31: 863–882, 2003.

[25] Vitkup, D., Kharchenko, P., Wagner, A., Influence of metabolic network structure and function on enzyme evolution, *Genome Biology*, 7: R39, 2006.

[26] Weinhold, N., Sander, O., Domingues, F.S., Lengauer, T., Sommer, I., Local function conservation in sequence and structure space, *PLoS Comput Biol*, 4(7): e1000105, 2008.

[27] Ycas, M., On earlier states of the biochemical system, *J. Theor. Biol.*, 44: 145–160, 1974.

[28] `http://blast.ncbi.nlm.nih.gov/`