doi:10.1016/S0022-2836(02)00546-6 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 320, 751–770

JMB

Evolution of Enzymes in Metabolism: A Network Perspective

Rui Alves[1,2], Raphael A. G. Chaleil[2] and Michael J. E. Sternberg[1,2]*

[1]*Department of Biological Sciences, Structural Bioinformatics Group Biochemistry Building Imperial College of Science Technology and Medicine London SW7 2AZ, UK*

[2]*Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX UK*

**Corresponding author*

Several models have been proposed to explain the origin and evolution of enzymes in metabolic pathways. Initially, the retro-evolution model proposed that, as enzymes at the end of pathways depleted their substrates in the primordial soup, there was a pressure for earlier enzymes in pathways to be created, using the later ones as initial template, in order to replenish the pools of depleted metabolites. Later, the recruitment model proposed that initial templates from other pathways could be used as long as those enzymes were similar in chemistry or substrate specificity. These two models have dominated recent studies of enzyme evolution. These studies are constrained by either the small scale of the study or the artificial restrictions imposed by pathway definitions. Here, a network approach is used to study enzyme evolution in fully sequenced genomes, thus removing both constraints. We find that homologous pairs of enzymes are roughly twice as likely to have evolved from enzymes that are less than three steps away from each other in the reaction network than pairs of non-homologous enzymes. These results, together with the conservation of the type of chemical reaction catalyzed by evolutionarily related enzymes, suggest that functional blocks of similar chemistry have evolved within metabolic networks. One possible explanation for these observations is that this local evolution phenomenon is likely to cause less global physiological disruptions in metabolism than evolution of enzymes from other enzymes that are distant from them in the metabolic network.

Keywords: sequence analysis; protein structure; enzyme classification; metabolic databases; comparative genomics

Introduction

Cellular metabolism is a complex network of physico-chemical processes, most of them catalyzed by enzymes, that allows the survival and reproduction of cells. In the early stages of evolution of metabolism, it is likely that a small number of enzymes with low effectiveness and broad specificity existed. Under natural selection, these enzymes were duplicated, becoming increasingly specialized and effective. The broad specificity of many of these enzymes must have been lost. A "second wave" of enzyme evolution would then have started with even more specialized enzymes evolving from pre-existing enzymes that already had high specificity and efficiency. There are two prevalent models currently used to explain enzyme evolution. Retro-evolution,[1] originally suggested by Horowitz, proposes that enzymes at the beginning of pathways evolve from the enzymes at the end of pathways, by duplication and mutation of the latter. As substrates from the end of the pathway were depleted in the primordial soup, there was a selective pressure for new enzymes to produce these substrates from other pre-existing compounds. Later, Jensen proposed the recruitment model of enzyme evolution,[2] suggesting that enzymes are likely to have evolved by duplication and mutations of similar enzymes from other pathways. One way to quantify this similarity is by using the enzyme commission (EC) classification scheme, based on a number hierarchy of four digits.[3] The first digit

Present address: R. Alves, Departament Ciencies Mediques Basiques, Universidad de Lleida, Av Rouvira Roure 44, 25198 Lleida, Spain.

Abbreviations used: EC, enzyme commission; PSIBLAST, position-specific iterated basic local alignment search tool; SCOP, structural classification of protein structures.

E-mail address of the corresponding author: m.sternberg@ic.ac.uk

0022-2836/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved

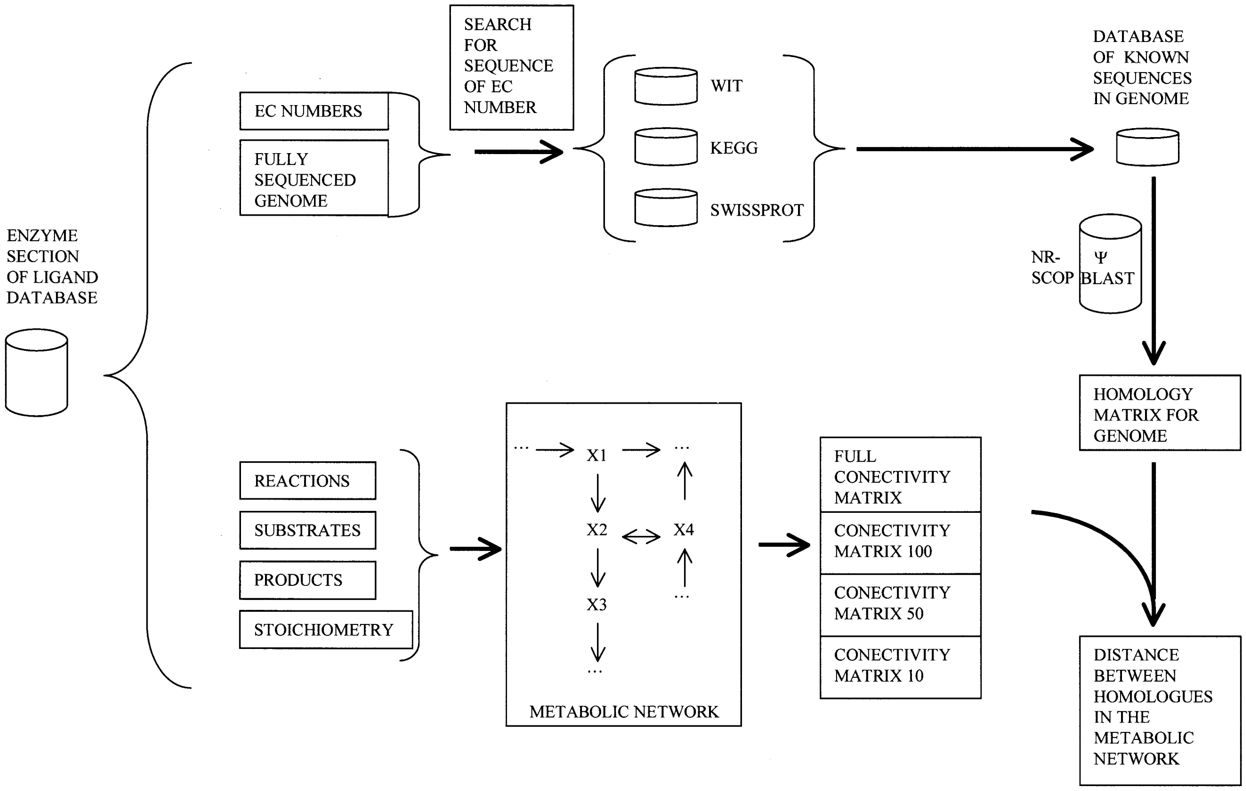

Figure 1. Experimental procedure. See the text for details.

describes the class of the enzyme, namely 1 for oxyreductases, 2 for tranferases, 3 for hydrolases, 4 for lyases, 5 for isomerases and 6 for synthetases. Subsequent digits define further details of function and reactants. If two enzymes belong to the same class in this classification, they are considered to have similar chemical function. Recent studies,[4–10] using both sequence and structure similarity to identify probable homology, have examined the evolution of enzymes in pathways. Copley & Bork[10] studied the evolution of 23 different TIM barrel SCOP superfamilies and found indications that at least 12 of these had a common evolutionary origin. Tsoka & Ouzounis[7] have studied the enzymes of *Escherichia coli* and found that enzymes from the same family are distributed among different pathways. Teichmann *et al.*[5] found that homologues are twice as likely to be found in different pathways than in the same pathway. These studies suggest, as a general model, that new enzymes are more likely to have evolved from enzymes belonging to the same enzyme class than from enzymes in the same pathway, thus following the recruitment model.

One constraint of studying enzyme evolution in a pathway context is that metabolism is a very intricate network, with pathways branching into each other. There are several databases that provide annotated metabolic pathways for different organisms, like EcoCYC,[11] KEGG[12] and WIT[13]. Many enzymes that are thought to have evolved by recruitment may have evolved by retro evolution, being only one or two reactions away from their homologues in a different pathway. Here, we study enzyme evolution in the context of the metabolic network, thus avoiding this problem. This change of context also makes it useful to redefine the retro-evolution and recruitment models as local (homologous enzymes being close to each other in the reaction network, independent of the pathway they belong to) evolution and long-distance (homologous enzymes being far from each other in the reaction network, independent of the pathway they belong to) evolution, respectively.

Another constraint of those studies is the small scale of the study, either in the sense of using a limited number of proteins[9] or that they study only one organism,[4–8,10] usually *E. coli*. There are now several fully sequenced and annotated genomes that lend themselves to the same kinds of study and evolutionary results from several organisms are needed to build a more accurate picture of the evolution of metabolism. Here, a network approach is used to study enzyme evolution in the fully sequenced genomes of 12 organisms, thus removing both constraints.

Building and Analyzing Metabolic Networks

Our approach is summarized in Figure 1. We parsed the information in the LIGAND part of

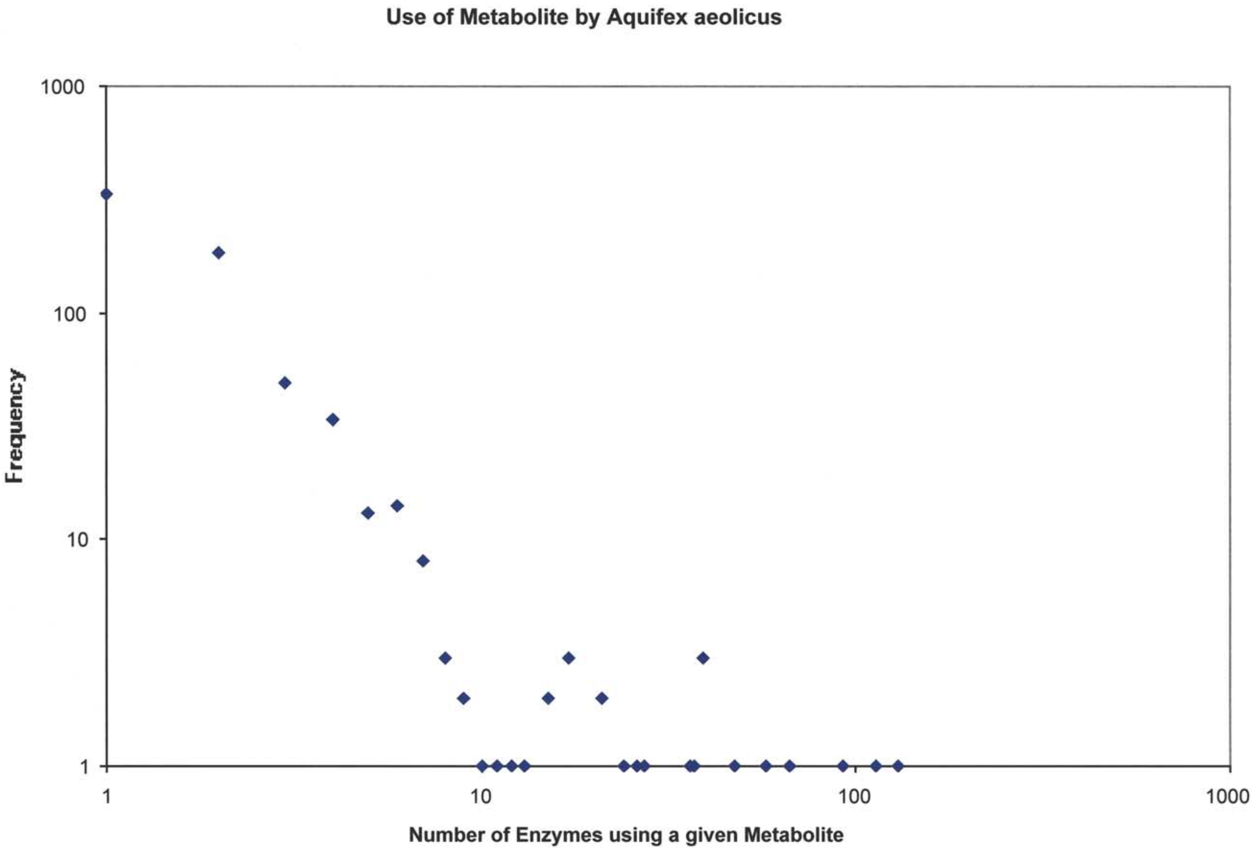

Figure 2. Representative example of a log–log plot of the frequency of the number of enzymes that use any given metabolite. The network is close to scale-free for metabolites being used in less than 100 and more than ten different reactions. Metabolites used in more than 100 reactions are H_2O, H_2O_2, AMP/ADP/ATP, phosphate, NAD(H), NADP(H), H, O_2, CO_2, CoA/Acetyl CoA, acetate, UMP/UDP/UTP, CMP/CDP/CTP, GMP/GDP/GTP, NH_3 and FAD(H).

KEGG [12] and BRENDA† databases to construct a new database, with all of the known enzyme activities defined by the enzyme commission according to their EC number. BRENDA is a database that was used to complete the information in the LIGAND regarding substrates, products and stoichiometry of enzyme reactions. We included information about the substrates, products and stoichiometry for each of the reactions catalyzed by an enzyme, building connectivity matrices between all the enzymes in the network. The connectivity matrices allow us to treat metabolism as a directed-graph and study the topological characteristics of the network. Several connectivity matrices were derived from the database. One matrix considered all metabolites as vertices of the network graph. This matrix includes promiscuous metabolites that are involved in very many reactions, such as water or ATP. Inclusion of these metabolites in the connectivity network will, for example, make it possible for homologous enzymes that have nothing in common but the use of water as a reactant to be only one step apart in the metabolic network. This would lead to spurious suggestion of local evolution. To address this problem, we created other connectivity matrices, considering as vertices of the network those metabolites that are involved in less than 10, 50 or 100 reactions (i.e. with promiscuity indexes lower than 10, 50 or 100, respectively), thus excluding metabolites at different levels of promiscuity. There is a correlation between these connectivity numbers and the number up to which the typical metabolic network is scale-free. Figure 2 shows a typical connectivity plot, which shows that, in logarithmic space, there is an inverse proportion between the number of enzymes using a metabolite and the percentage of metabolites that is used by a given number of enzymes, for metabolites with connectivity index higher than 10 and lower than 100. This proportionality indicates that the metabolic network is a scale-free network; that is, a network that has the same average degree of branching at different levels of detail, for metabolites that are used by more than approximately ten enzymes and less than approximately 100.

The next step in connecting this network perspective and the evolution of enzymes is to obtain the protein sequences for the enzymes in the metabolic networks of individual organisms. We therefore searched SWISSPROT,[14] WIT[13] and KEGG[12]

† http://www.brenda.uni-koeln.de/

Table 1. Homology information for the different organisms that were studied

	Organisms	% Enzymes with at least one homologue	% Enzymes with at least one homologue using SCOP	Number of enzymes with homologues	Number of enzymes with homologues using SCOP	Total number of enzymes	Number of pairs of homologues	Number of pairs of homologues with promiscuity index <100	Number of pairs of homologues with promiscuity index <50	Number of pairs of homologues with promiscuity index <10
Archea	*Aquifex aeolicus*	40.62	43.01	78	83	192	464	405	383	161
	Archaeoglobus fulgidus	38.58	38.58	49	49	127	309	263	263	96
	Methanobacterium thermoautotrophicum	37.14	40.43	52	57	140	258	222	213	81
	Thermotoga maritima	22.22	26.28	30	36	135	237	196	182	79
Eukaryota	*Arabidopsis thaliana*	22.35	27.37	40	49	179	345	222	213	168
	Caenorhabditis elegans	26.49	34.76	49	65	185	438	298	282	179
	Schizosaccharomyces pombe	26.42	27.57	56	59	212	389	269	250	161
Bacteria	*Bacillus subtilis*	46.53	50.35	201	218	432	1909	1246	1140	748
	Escherichia coli	50.62	53.34	326	343	644	4289	2749	2574	2006
	Haemophilus influenzae	32.99	38.18	127	147	385	1306	983	957	566
	Pseudomonas aeruginosa	18.52	21.48	25	29	135	175	88	80	69
	Salmonella typhimurium	18.6	19.82	40	43	215	395	363	313	297

Included is the percentage of enzymes that has homologues as well as the total number of enzymes found in each organism. We include information about the added value of using the structure information in SCOP to determine homologues. Using this information, homologues are found for a further 2% to 8% of the enzymes in each genome, with the exception of *Archaeoglobus fulgidus*, for which there is no information added by using SCOP. The last four columns show the number of homologue pairs for each connectivity matrix, i.e. if we consider as homologues only proteins that are homologous in domains that are not involved in the binding of metabolites with promiscuity index higher than that for the relevant homology matrix.

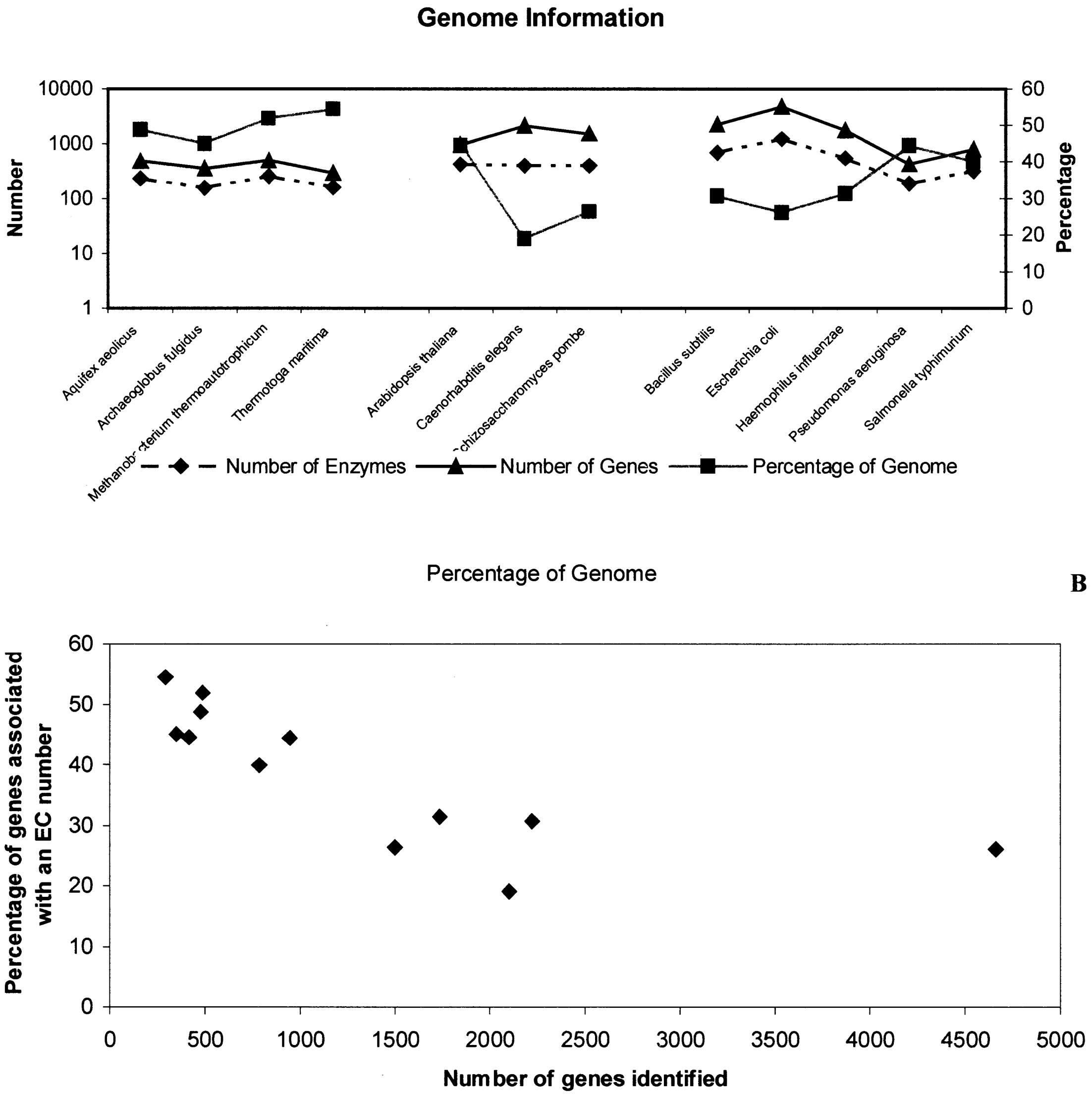

Figure 3. Genome information for the different organisms that were analyzed. (a) Plot including information about the number of genes annotated in each genome and the number and percentage of genes associated with an EC number. (b) Plot of the percentage of the genome that is associated with and EC number as a function of the genome size.

for the sequences of each of the EC numbers in fully sequenced genomes of representative organisms that are publicly available. These organisms are indicated in Table 1. We present results for 12 species with fully sequenced genomes, with at least 100 identified enzyme sequences for each. These species include four Archaea, three Eukaryota and five Bacteria. The choice of not using databases that are committed to one specific organism (e.g. EcoCYC) to obtain protein sequences associated with EC numbers is deliberate and justifiable. These databases are, in principle, more complete and better annotated but there are very few of them available. Thus, we would have had to use data that have different bias for different organisms. By using general databases we aim at decreasing the differences in the bias of genome annotation between different organisms, thus making the data more readily comparable. Because the assignment of protein sequences from these different genomes to metabolic steps is continually being updated, it will be important to repeat this analysis once the available data are augmented substantially.

We ran the sensitive sequence similarity search program PSIBLAST[15] to find the homologues among the different enzymes within each organism. *E*-values smaller than 0.001 were set as sequence homology filters. PSIBLAST identifies a local sequence similarity and thus a significant sequence similarity between a domain in one protein and all or part of the sequence of another protein would be interpreted as a homology between these two proteins. To increase the power of the homology search, we also ran PSIBLAST against the SCOP database.[16] This database collects proteins into hierarchical groups that take into account sequence, 3-D structure and function, thus allowing the determination of homologies that would have been missed if only sequence

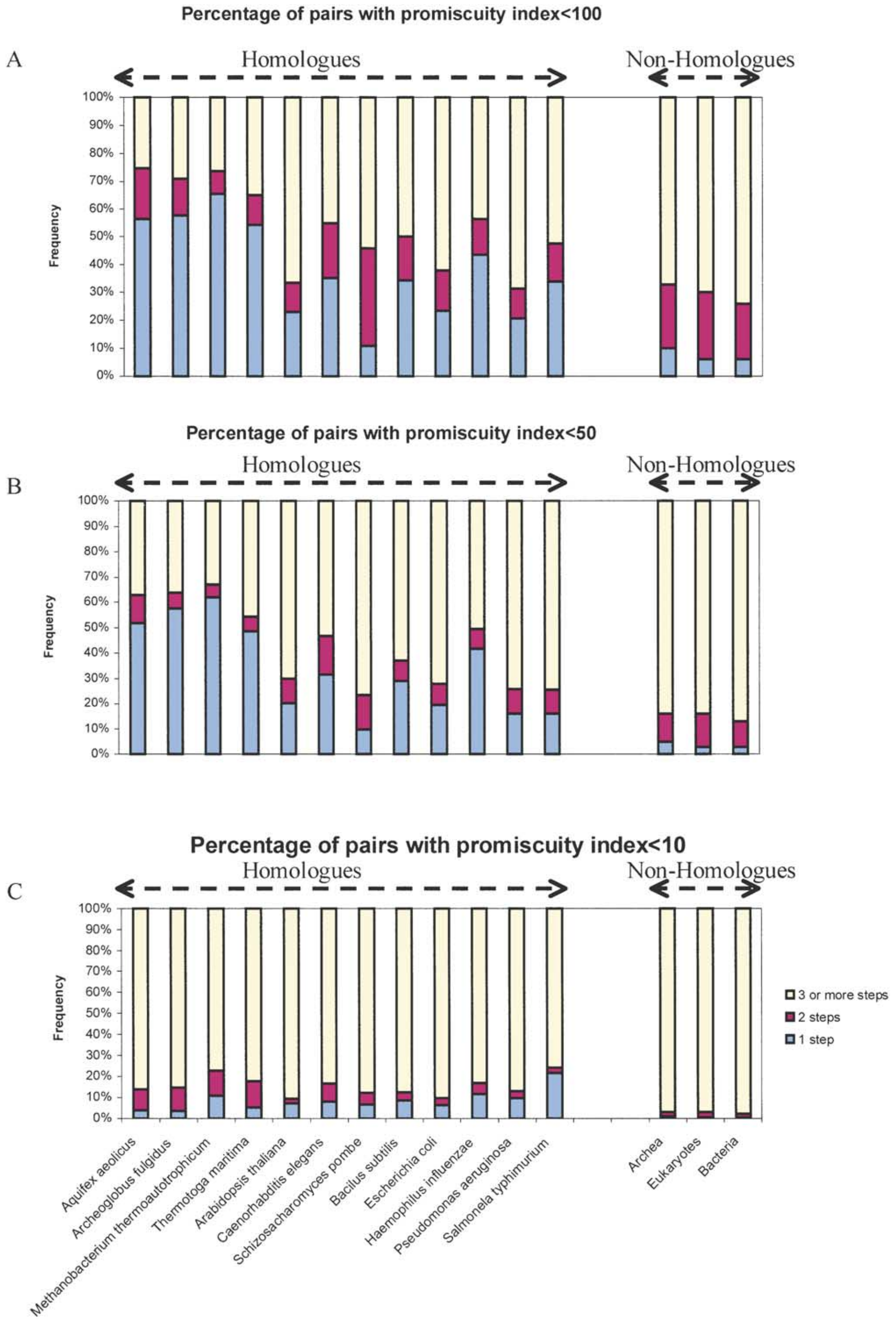

Figure 4. Plot of the frequency of homologue pairs within one, two and three or more steps in the metabolic network of different organisms using different connectivity matrices. In each plot, we present a non-homologous column for Archaea, another for Bacteria and a final one for Eukaryota. These columns give the median distribution for pairs of non-homologues within each of these groups for the different connectivity matrices. The maximum deviation from these values is always smaller than 5%. (a) Connectivity matrix excluding metabolites that are used by more than 100 enzymes (i.e. with promiscuity index >100). (b) Connectivity matrix excluding metabolites with promiscuity index >50. (c) Connectivity matrix excluding metabolites with promiscuity index >10.

information had been considered. Depending on the organism, the percentage of homologues that would have been missed without the use of SCOP is roughly between 1% and 8%, being lower, on average, for Archaea than for Bacteria or Eukaryota. Table 1 shows that, on average, there is a higher percentage of enzymes that have homologues in Archaea than in Bacteria or Eukaryota.

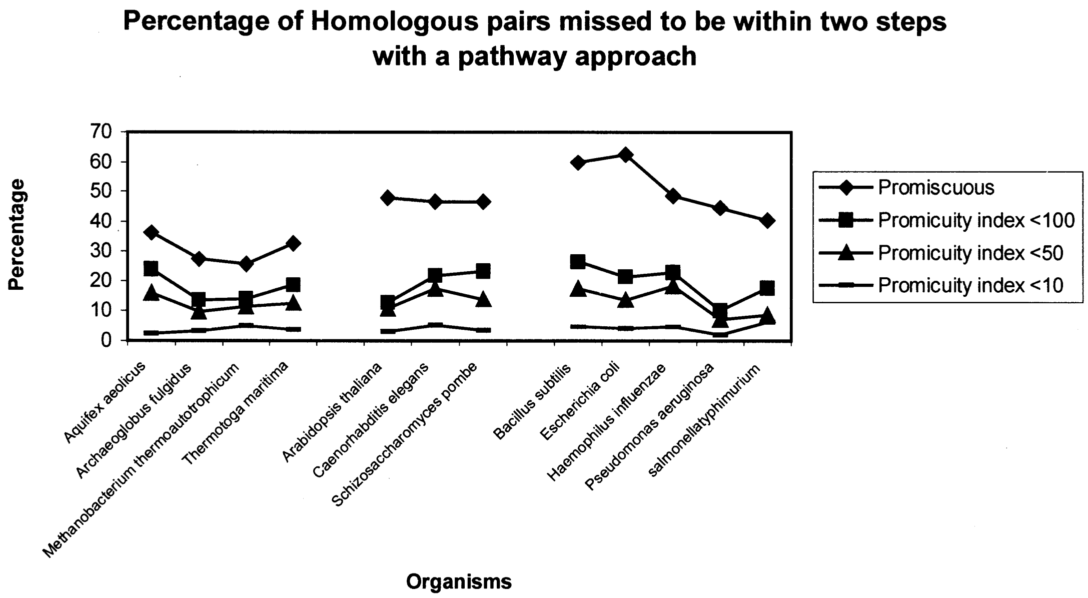

Figure 5. Plot representing the percentage of homologue pairs that would have been missed as being less than three steps away from each other if we had used the KEGG pathway definition for each organism. The different curves represent the results when using different connectivity matrices. The matrix that includes all metabolites is labeled as Promiscuous. The matrix that excludes metabolites with promiscuity index >100 is labeled as Promiscuity index <100. The matrix that excludes metabolites with promiscuity index >50 is labeled as Promiscuity index <50. The matrix that excludes metabolites with promiscuity index >10 is labeled as Promiscuity index <10.

Figure 3(a) shows the total number of genes identified for each studied organism in SWISSPROT, as well as the number and percentage of those genes that have been associated with an EC number. We found that, in general, the percentage of genes associated with an EC number is inversely proportional to the number of genes in the database for genomes smaller than 1500–2000, and appears to remain constant at approximately 20% as the number of genes in the genome of an organism increases above that value. Similar trends are observed for the percentage of genes that is associated with each enzyme class except hydrolases and synthetases. These two classes appear to represent a roughly constant percentage of the total number of genes (data not shown).

Evolution of Enzymes in a Metabolic Network

Figure 4 shows that, on average, long-distance evolution is at least as common as local evolution in the metabolic network. However, local evolution, as measured by the percentage of pairs of homologues that are less than three steps away from each other in the network, is always significantly higher than would be expected by chance alone. This is independent of the promiscuity index that is set as threshold in the connectivity matrix. Random shuffling (see Materials and Methods for an explanation) shows that the patterns of distance in the network have a likelihood of less than 1% of occurring by chance. Figure 4(a) shows the results for a connectivity matrix that includes metabolites with promiscuity indices smaller than 100. Each of the first 12 columns represents an organism and it is shown that a very high percentage of homologues is close to each other in the metabolic network. The last three columns present the average results of the number of steps between members of pairs of non-homologous enzymes for Archaea, Eukaryota and Bacteria, respectively, and show that the distance pattern of pairs of homologues is not just a result of the network connectivity.

If we consider only the network that is scale-free (i.e. Figure 4, with any of the three promiscuity indices), we find that, for each of the three branches in the tree of life, the percentage of enzymes that has evolved locally is higher than would be expected by chance. In general, local evolution of enzymes is more often observed in Archaea than in Bacteria or Eukaryota. When considering pairs of non-homologous enzymes in all species, we find that, on average, these are further away from each other in the network than pairs of homologues.

Figure 4 also provides an example of the shortcomings of using only one organism to study general evolutionary issues. Most of the evolutionary studies of enzymes have been done using the genome of *E. coli*. The percentage of homologues that are less than three steps away from each other in the metabolic network of *E. coli* is below average. Thus, the picture that one gets regarding local evolution of enzymes by studying only that organism is skewed.

To compare how the percentage of pairs of enzyme homologues that are less than three steps

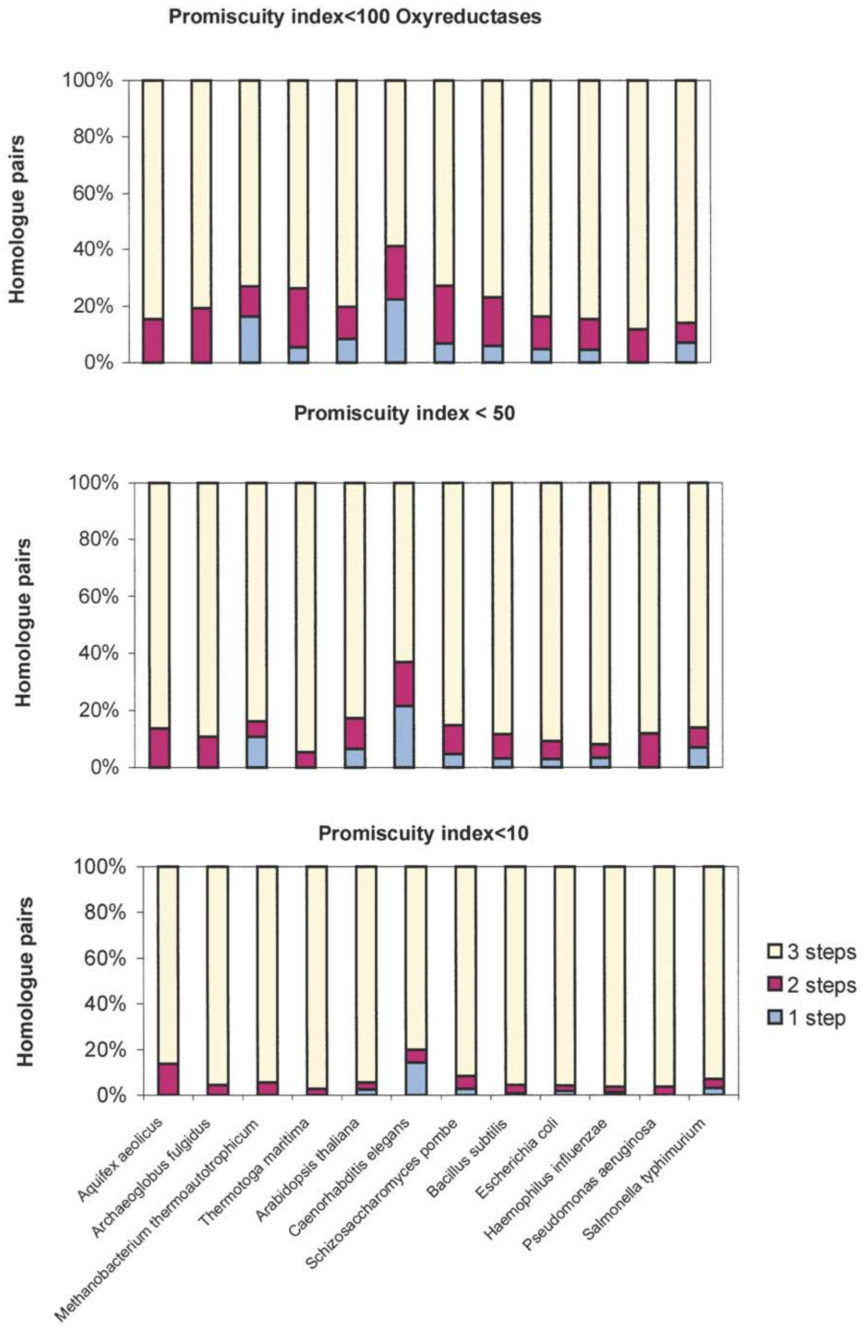

Figure 6 (*legend on page 763*)

away from each other in the network is affected by the use of traditionally defined metabolic pathways, we performed the same studies using the KEGG pathway scheme. The percentage of homologues that are less than three steps away from each other in the network is higher with the network approach than it is when we use the pathway approach. Figure 5 shows the percentage of pairs that would have been missed to be at less than three reaction steps from each other, had we used the traditional pathway definitions of KEGG. For the connectivity matrices with promiscuity index smaller than 100 and smaller than 50, this percentage is between approximately 10% and 30% of all homologue pairs. This is a measure of the added information that we obtain by using the network approach.

Figures 4 and 5 show that the decrease of the number of homologue pairs that are within two steps when comparing a connectivity matrix with promiscuity index >100 with a connectivity matrix with promiscuity index <100 is larger for Bacteria and Eukaryota than for Archaea. This is due to the peculiar metabolism of Archaea. These organisms use a set of enzymes in their anabolism and energy metabolism that is not similar to those used by

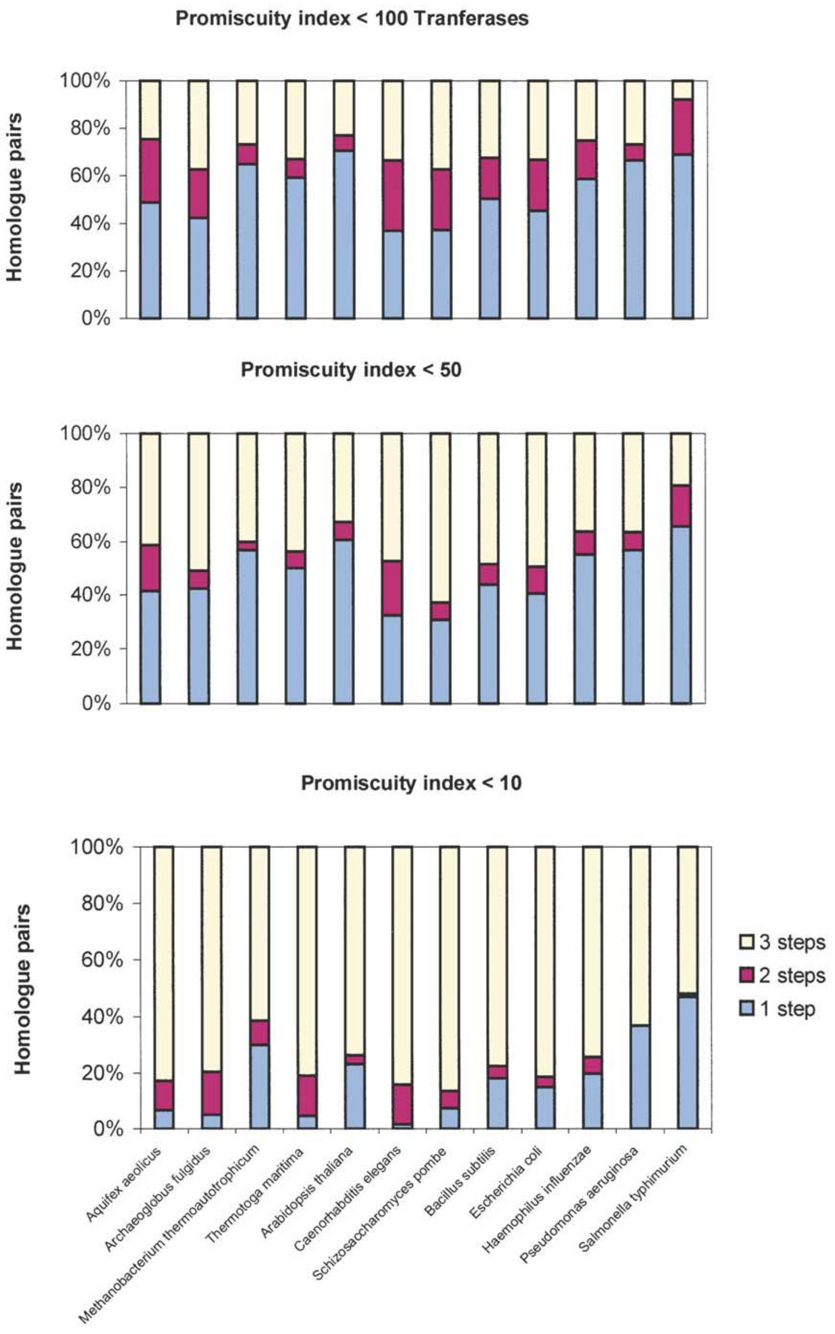

Figure 6 (*legend on page 763*)

Eukaryota and Bacteria, thus making the connectivity of the Archaea metabolic network less sensitive to the removal of metabolite nodes that are promiscuous in the other two realms.

We examined the association between enzyme class and distance in the network between homologues. Figure 6 shows that transferases (class 2) and synthetases (class 6) have a very high percentage of homologue pairs within two reaction steps of each other, independent of the connectivity matrix that is used to measure the distance. For connectivity matrices excluding metabolites with promiscuity indices higher than 10 there is still an average percentage of 60% of homologue pairs with members that are less than three reaction steps away from each other. In general, enzymes in these classes do not use reactants that have high promiscuity indices and thus, the distance between them is not affected when metabolites with high promiscuity indices are excluded from the network connectivity matrix. What this suggests is that the percentage of enzymes from either the synthetases or the transferases classes within three steps of each other is significantly higher than it would be if enzymes had been distributed randomly in the network. This is true

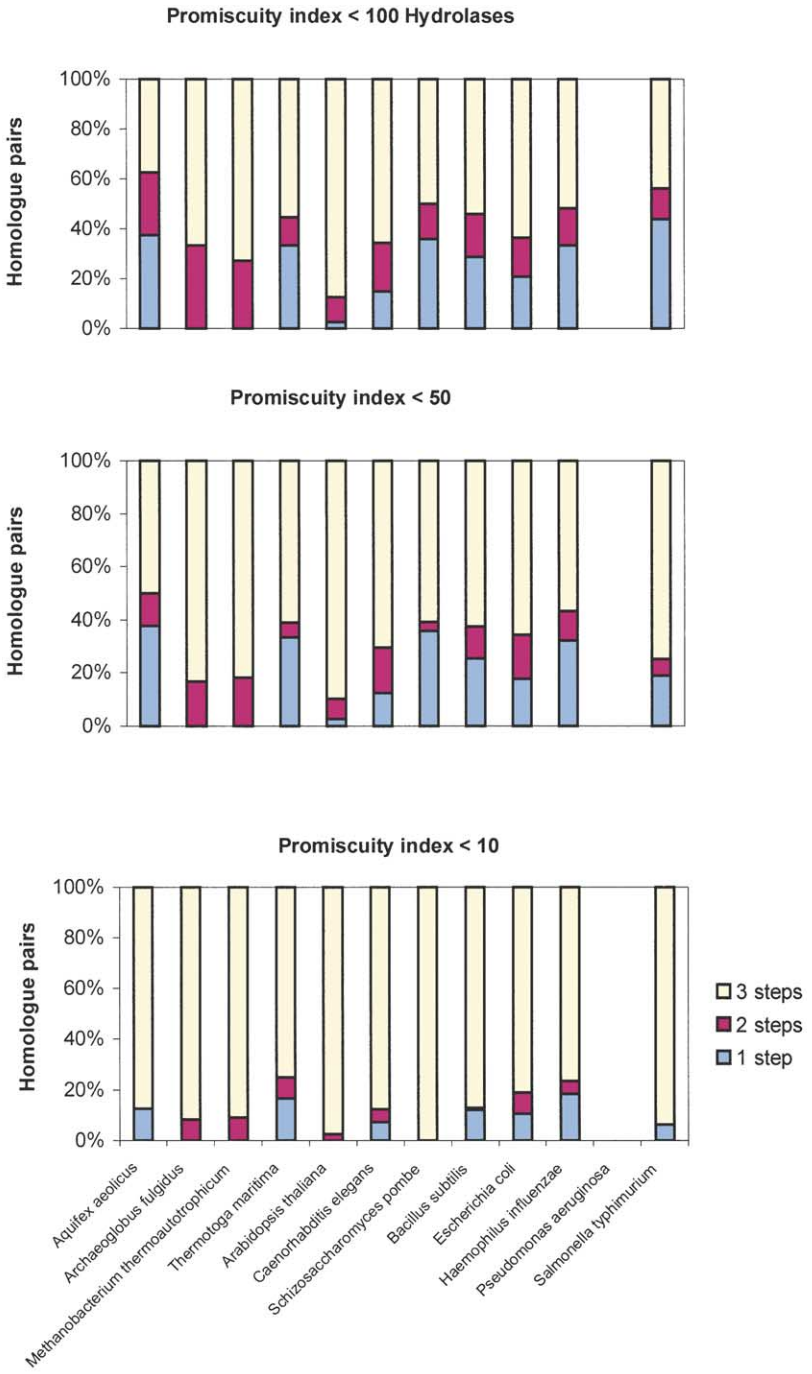

Figure 6 (*legend on page 763*)

also for hydrolases (class 3) and lyases (class 4), although to a lesser extent, with approximately 40% of homologue pairs having members that are less than three steps away from each other in the reaction network.

In contrast, oxyreductases (class 1) and isomerases (class 5) have a very high percentage of homologue pairs that are more than two steps away from each other, with less than 20% of homologue pairs having members that are less than three steps away from each other for connectivity matrices with a promiscuity index equal to or lower than 50. Thus, these enzymes are spread throughout the metabolic network more evenly than the other four classes. Oxyreductases are enzymes that, in general, use metabolites with high promiscuity indices (e.g. NAD(P) or FAD). When these are removed from the connectivity network, the average distance between homologues will increase sharply. However, isomerases still have a lower percentage of homologues less than three steps away from each other even when all

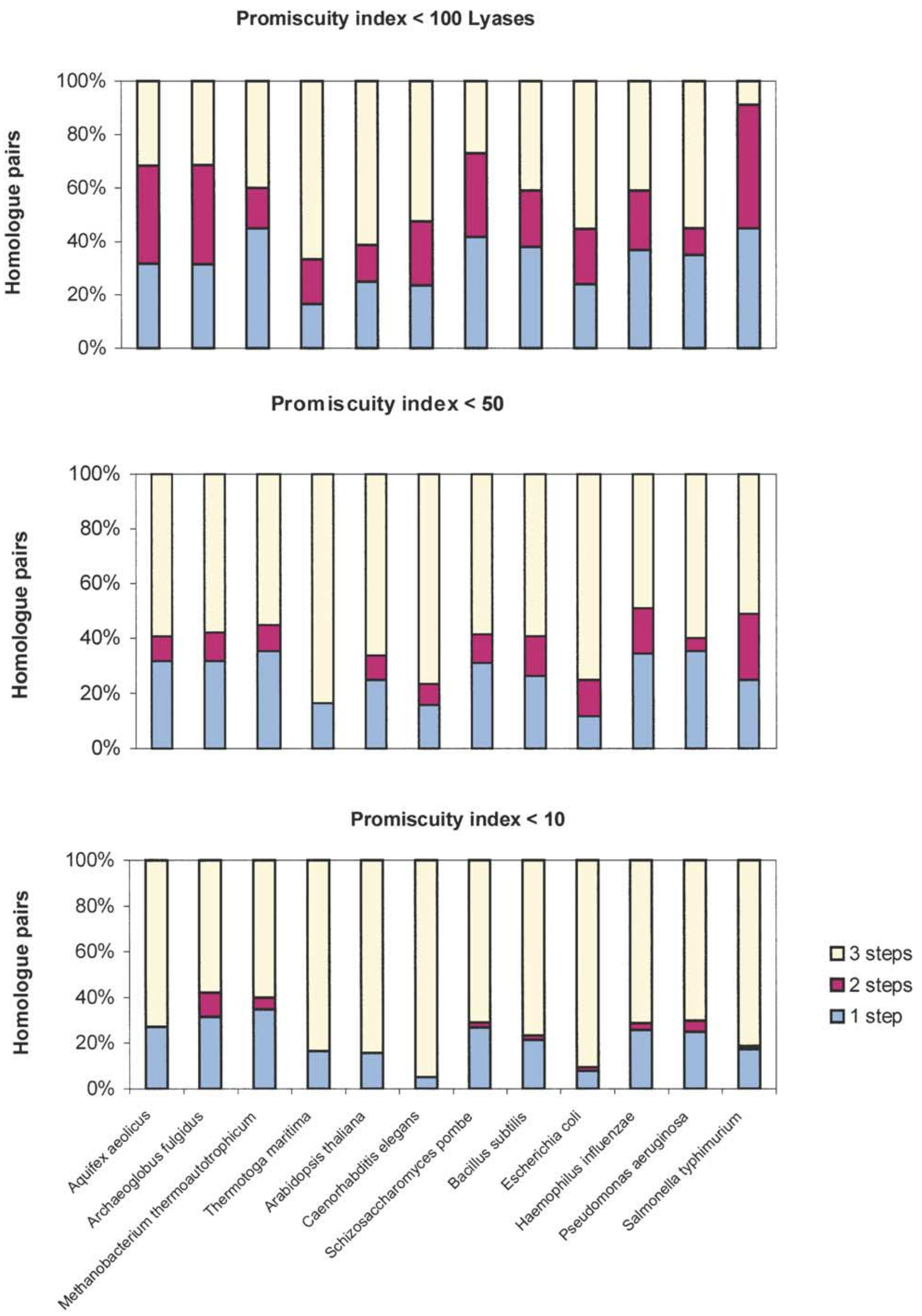

Figure 6 (*legend on page 763*)

metabolites are used in the connectivity matrix. The average distance in the network between homologues is less affected by the decreasing promiscuity indexes in connectivity matrices than for oxyreductases.

Chemistry Conservation in Evolution

Recent general surveys of the structure and function of homologous proteins have highlighted the fact that chemistry tends to be more conserved than substrate specificity.[4,5,17,18] However, as the degree of sequence identity between pairs of enzymes decreases, the extent of commonality of function decreases. Todd *et al*.[6] have shown that some fairly close homologues can differ in chemistry, sometimes having greater similarity of substrate specificity. In keeping with this study, we considered that pairs of enzymes with the same first digit of their four-digit EC number have broadly similar chemistry, since they performed the same general class of reaction. Our data support their result, that homologous enzymes tend

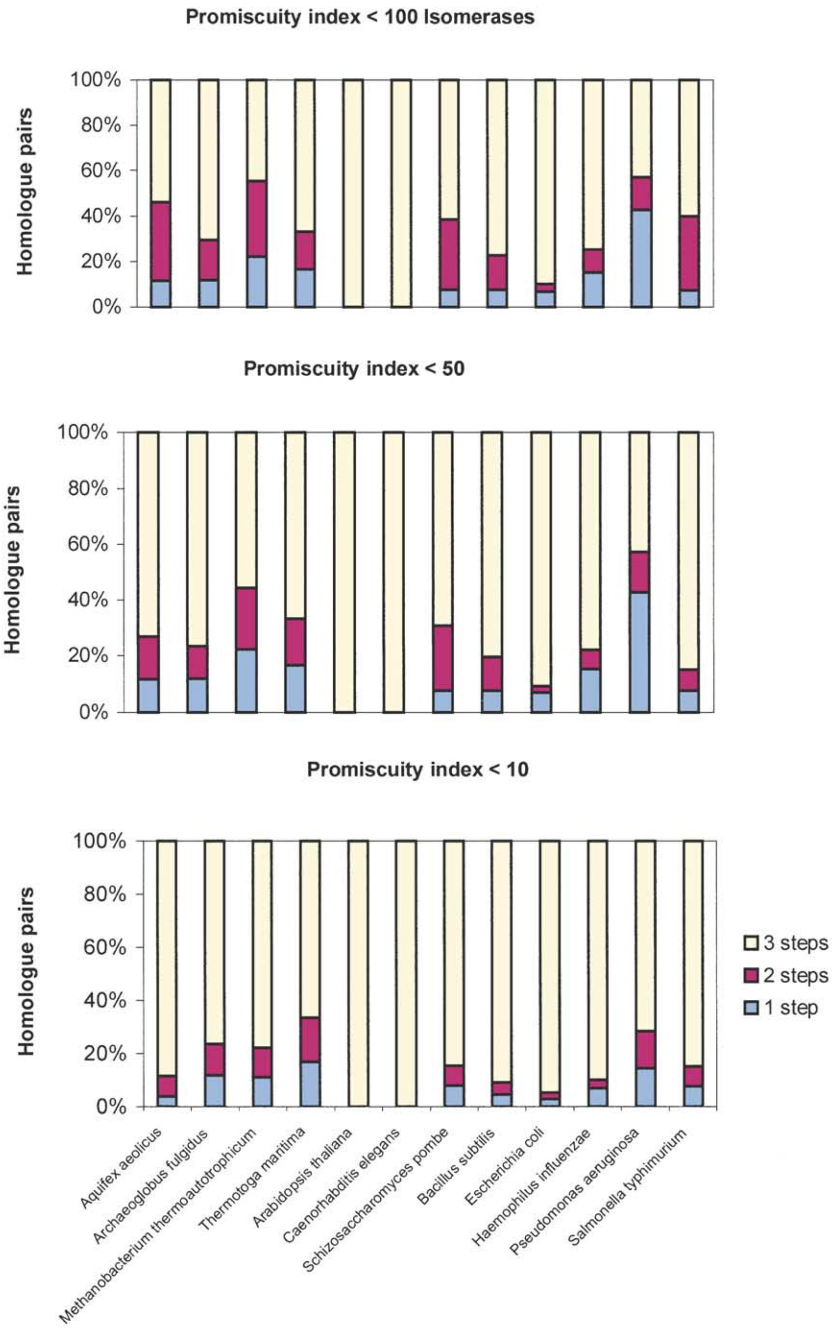

Figure 6 (*legend opposite*)

to have greater conservation of chemistry than non-homologues. Figure 4 shows that there is a high likelihood for enzymes to evolve from other enzymes that are close by in the reaction network of metabolism. Figure 7 shows, for a connectivity matrix with promiscuity index smaller than 50, that the likelihood of two homologues belonging to the same enzyme class is, in general, at least twice as high as that of having homologues with different first digits in their EC numbers. Random shuffling shows that this is the approximate average probability for any random pair of enzymes to share the same first digit in their EC numbers.

To study the association between homology, distance between enzymes in the network and chemistry conservation, we use Table 2, using a promiscuity index equal to or smaller than 50. Table 2 shows measures of how much more or less likely (i.e. the odds) it is for a pair of enzymes to have a given characteristic if they are less than three steps away from each other in the network than if they are three or more steps away from each other in the network. For example, for *Aquifex aeolicus*, the odds that members of a pair of homologues are less than three steps away from each other if they belong to the same enzyme class

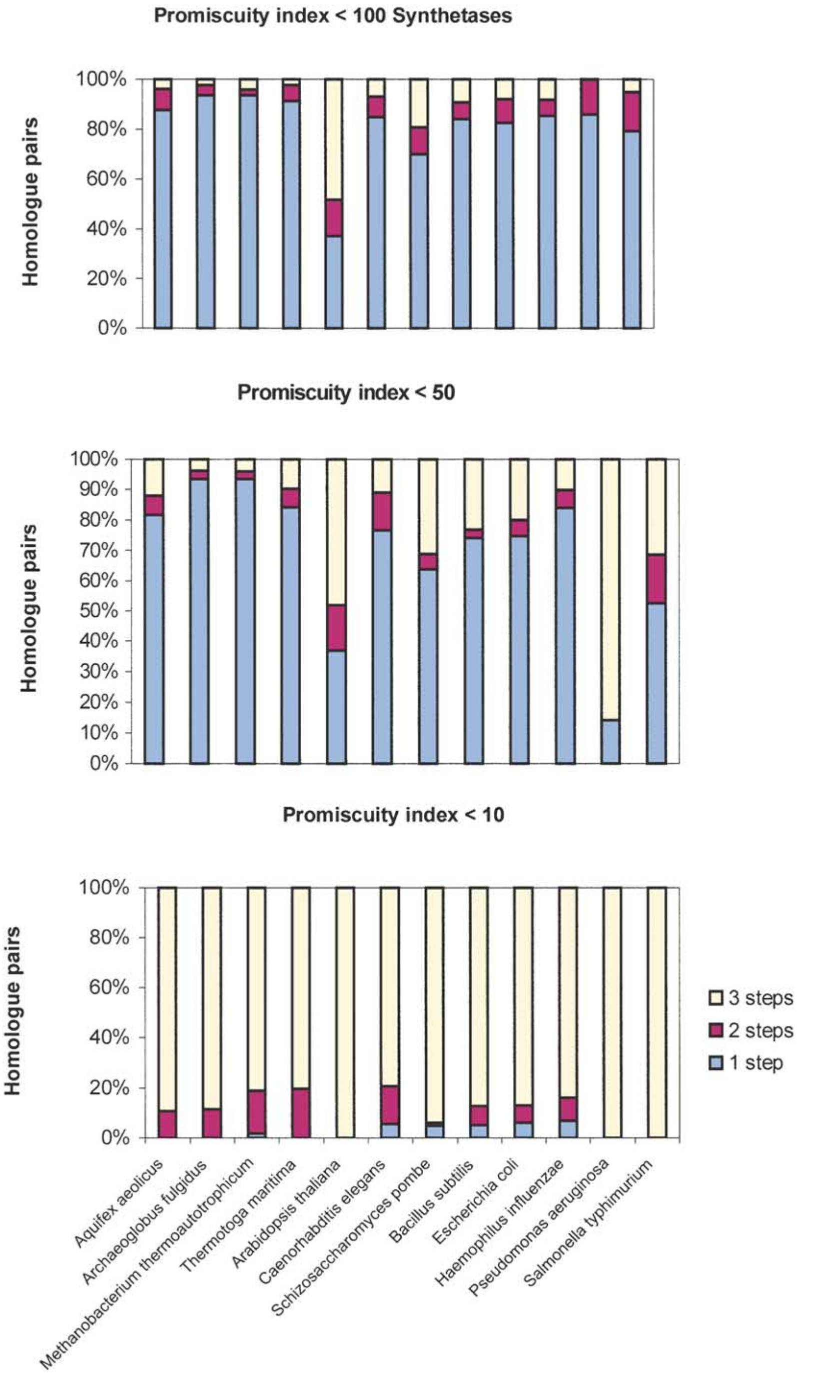

Figure 6. Plots of the frequency of homologue pairs within one, two and three or more steps in the metabolic network of different organisms for the six different enzyme classes using different connectivity matrices. Plots labeled Promiscuity index <100 have been obtained using a connectivity matrix that excludes metabolites with a promiscuity index >100. Plots labeled Promiscuity index <50 have been obtained using a connectivity matrix that excludes metabolites with a promiscuity index >50.

is approximately three times higher than those of being three or more steps away from each other in the network. These odds can be found in Table 2, under same first EC number digit/homologues and they are the ratio between the proportion of homologous pairs that belong to the same class and are less than three steps away from each other to the proportion of pairs of homologues that are three or more steps away from each other and belong to the same enzyme class. Monte Carlo simulations have shown that the average random value for the odds presented in Table 2 is around 0.15. The odds from the homologues columns have a probability of occurring by chance that is smaller than 0.01 in these simulations. On the other hand, the values for the odds of the entries non-homologues/different first EC number digit are always about the average values from the simulations.

To investigate how the distance in the network is associated with homology and chemistry, we use the odds ratios, also given in Table 2. An odds ratio larger than 1 demonstrates an association between proximity of enzymes in the network and homology (odds ratio 2) or chemistry conservation (odds ratio 1). The higher the odds ratio, the stronger is the association.

In general, three types of associations are observed. For *Caenorhabditis elegans*, *Bacillus subtillis*, *Haemophilus influenza*, *Pseudomonas*

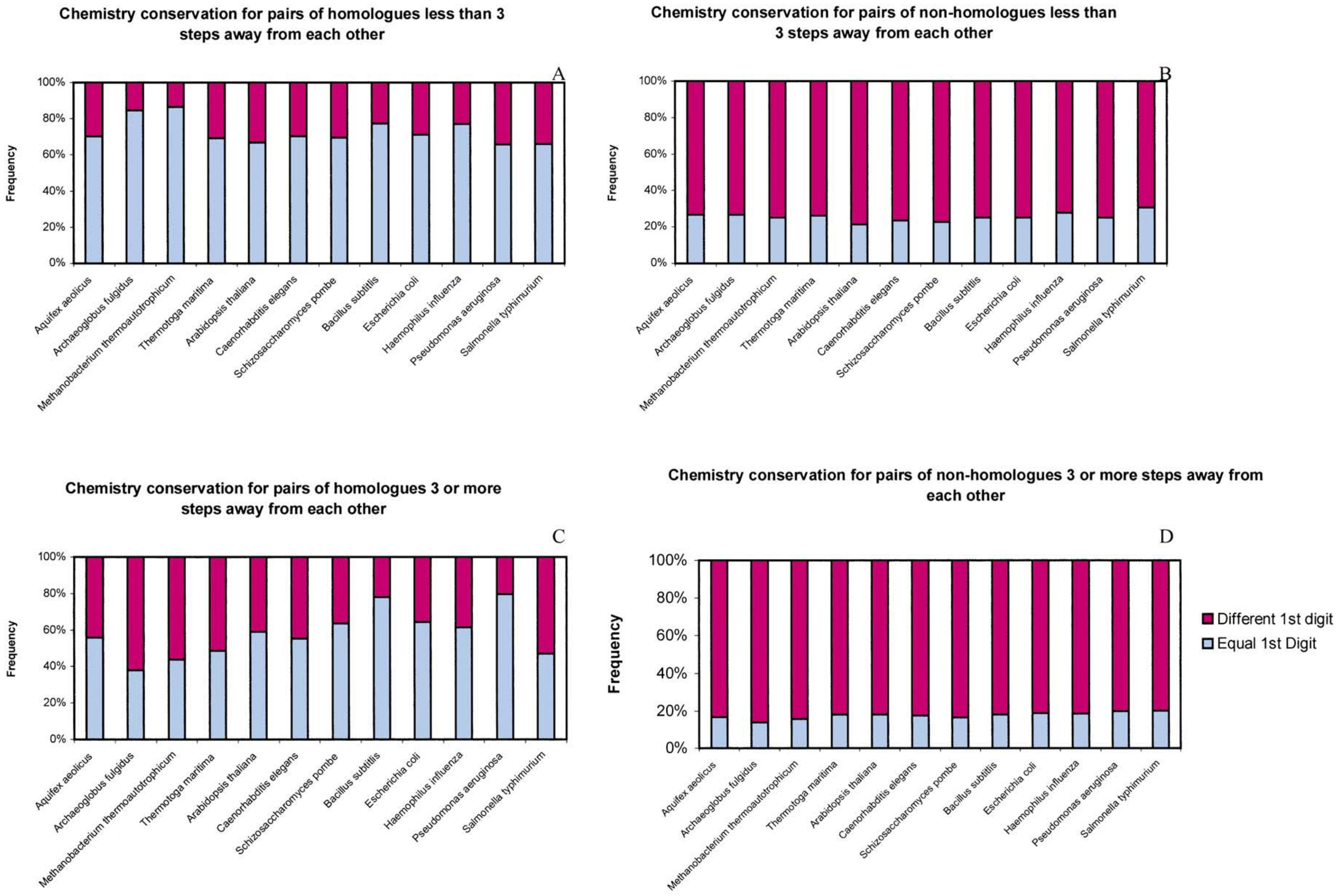

Figure 7 (*legend opposite*)

aeruginosa and *Salmonella typhimurium*, the two values for odds ratio 2 are similar and greater than 1.0 irrespective of the chemistry. For the same organisms, the values for odds ratio 1 are also similar and greater than 1.0 irrespective of homology. The values for odds ratio 2 are greater than for odds ratio 1. This shows that both homology and chemistry are associated with proximity (less than three steps in the network) but that homology has the stronger association.

The second type of association is observed in *A. aeolicus*, *Archaeoglobus fulgidus*, *Methanobacterium thermoautotrophicum* and *Thermotoga maritima*. The odds ratio 1 for pairs of homologues is higher than that for pairs of non-homologues, and so we infer that there is a stronger association between chemistry conservation and proximity of enzyme in the network if the enzymes are homologues than if they are not. This pattern is repeated for the odds ratio 2, showing that the association between homology and proximity of enzymes in the network is strong. When comparing the odds ratio 1 to the odds ratio 2, the latter is higher, showing a stronger association between distance in the network and homology than between distance in the network and chemistry. Nevertheless, we observe a strong association between all the three variables, because the odds ratios for the column homologues and the row same first EC number digit are much higher than the remaining odds ratios. For *C. elegans* and *S. typhimurium*, this association of both homology and chemistry is observed but only to a small extent and for these two species we consider, to a first-order approximation, that homology and chemistry are not associated.

The final grouping is observed in *Arabidopsis thaliana*, *Schizosaccharomyces pombe* and *E. coli*. Similar values greater than 1.0 are observed for all four odds ratios. Thus, there are associations between network proximity and homology, and between network proximity and chemistry. These associations have approximately the same strength and there is no marked reinforcement of the association between proximity in the network and homology (similar chemistry) if enzymes have similar chemistry (are homologous).

It is interesting to note that the type of association between chemistry, distance and homology in *C. elegans* is the same as in Bacteria and distinct from that of the remaining eukaryotes. Similarly, the type of correlation between distance and homology in *E. coli* is the same as that of yeast and cress, and different from that of the other bacteria. The reasons for this association of organisms are not clear and we could not find any biologically relevant reason that could explain these groupings.

In all species we find that there is an association between similar chemistry (as defined by the same first EC digit) and proximity in the network irrespective of homology. This is shown by an odds ratio 1 always being greater than 1. One might expect that this is a trivial consequence of proximate enzymes acting on metabolites with similar chemical structure. However, the definition of the first EC digit is in terms of the general class of chemistry (e.g. oxyreductase) and this is not *a priori* connected with the chemical nature of the metabolite. Furthermore, our results still hold when promiscuous metabolites are not considered in the connectivity of the network. Our observation suggests that it would be useful to undertake a comprehensive analysis of the relationship between conservation of the chemistry of metabolites and the nature of catalyzed steps.

These inter-species differences of association between network proximity, homology and chemistry are complex, and merit more detailed analysis that would be more appropriate for subsequent papers.

Discussion and Conclusions

Our key results are: (1) the percentage of pairs of homologous enzymes that are less than three steps away from each other in the metabolic network is significantly higher than what is expected had the network evolved randomly. (2) There often is a clustering effect of enzymes belonging to the same class in metabolic networks. (3) In several species, these two effects are linked and there is a particularly strong tendency for homologous enzymes with similar chemistry to be found less than three steps away from each other in the network.

Why should local evolution have a higher occurrence than what is expected to occur randomly? Organisms evolve to adapt to their environment and improve their fitness. This is true at every level, including the cellular and metabolic

Figure 7. Plots of the frequency of homologue pairs that belong to the same enzyme class (i.e. same first digit in the EC number; equal first digit in the plot legends) or to different enzyme classes (different first digit in the plot legends) in the metabolic network of different organisms. This provides a measure of chemistry conservation among homologues. (a) Plot of chemistry conservation for pairs of homologues that are less than three steps away from each other in the metabolic network. (b) Plot of chemistry conservation for pairs of non-homologous enzymes that are less than three steps away from each other in the metabolic network. (c) Plot of chemistry conservation for pairs of homologues that are three or more steps away from each other in the metabolic network. (d) Plot of chemistry conservation for pairs of non-homologous enzymes that are three or more steps away from each other in the metabolic network. Comparing (a) and (c) to (b) and (d), respectively, suggests that chemistry conservation is significant in the evolution of enzymes. This is confirmed by Monte Carlo simulations (data not shown).

Table 2. Odds ratios for the correlation between chemistry conservation and distance in the network for the 12 different organisms

Aquifex aeolicus (odds $n(<3$ steps; $M = 283 + 6137)/n(\geq 3$ steps; $M = 100 + 11912))$

	Homologues ($N = 253 + 130$)	Non-homologues ($N = 3610 + 14313$)	Odds ratio 2
Same first EC number digit	3.13	0.33	9.48
Different first EC number digit	0.61	0.18	3.39
Odds ratio 1	5.13	1.83	

Archaeoglobus fulgidus (odds $n(<3$ steps; $M = 187 + 2809)/n(\geq 3$ steps; $M = 76 + 4993))$

	Homologues ($N = 187 + 76$)	Non-homologues ($N = 1482 + 6320$)	Odds ratio 2
Same first EC number digit	4.46	0.38	11.74
Different first EC number digit	0.26	0.14	1.86
Odds ratio 1	17.15	2.71	

Methanobacterium thermoautotrophicum (odds $n(<3$ steps; $M = 155 + 2684)/n(\geq 3$ steps; $M = 58 + 6903))$

	Homologues ($N = 160 + 53$)	Non-homologues ($N = 1917 + 7617$)	Odds ratio 2
Same first EC number digit	3.72	0.36	10.33
Different first EC number digit	0.47	0.16	2.94
Odds ratio 1	7.91	2.25	

Thermotoga maritima (odds $n(<3$ steps; $M = 118 + 2654)/n(\geq 3$ steps; $M = 64 + 7177))$

	Homologues ($N = 113 + 69$)	Non-homologues ($N = 1786 + 7145$)	Odds ratio 2
Same first EC number digit ($N = 1899$)	2.05	0.25	8.20
Different first EC number digit ($N = 7214$)	0.52	0.13	4
Odds ratio 1	3.94	1.92	

Arabidopsis thaliana (odds $n(<3$ steps; $M = 70 + 5375)/n(\geq 3$ steps; $M = 143 + 10433))$

	Homologues ($N = 130 + 83$)	Non-homologues ($N = 2845 + 12963$)	Odds ratio 2
Same first EC number digit	0.51	0.27	1.89
Different first EC number digit	0.30	0.17	1.77
Odds ratio 1	1.70	1.59	

Caenorhabditis elegans (odds $n(<3$ steps; $M = 155 + 4713)/n(\geq 3$ steps; $M = 127 + 12118))$

	Homologues ($N = 178 + 104$)	Non-homologues ($N = 3198 + 13633$)	Odds ratio 2
Same first EC number digit	1.25	0.33	3.79
Different first EC number digit	0.48	0.17	2.82
Odds ratio 1	2.60	1.94	

Schizosaccharomyces pombe (odds $n(<3$ steps; $M = 140 + 5333)/n(\geq 3$ steps; $M = 110 + 16889))$

	Homologues ($N = 163 + 87$)	Non-homologues ($N = 4667 + 17555$)	Odds ratio 2
Same first EC number digit	0.81	0.33	2.45
Different first EC number digit	0.38	0.17	2.24
Odds ratio 1	2.13	1.94	

Bacillus subtilis (odds $n(<3$ steps; $M = 570 + 25040)/n(\geq 3$ steps; $M = 570 + 67702))$

	Homologues ($N = 878 + 262$)	Non-homologues ($N = 18434 + 73738$)	Odds ratio 2
Same first EC number digit	0.67	0.25	2.68
Different first EC number digit	0.41	0.14	2.93
Odds ratio 1	1.63	1.79	

Escherichia coli (odds $n(<3$ steps; $M = 978 + 61438)/n(\geq 3$ steps; $M = 1596 + 143356))$

	Homologues ($N = 1725 + 849$)	Non-homologues ($N = 38911 + 165883$)	Odds ratio 2
Same first EC number digit	0.46	0.25	1.84
Different first EC number digit	0.26	0.13	2.00
Odds ratio 1	1.77	1.92	

Haemophilus influenza (odds $n(<3$ steps; $M = 536 + 23410)/n(\geq 3$ steps; $M = 421 + 49746))$

	Homologues ($N = 670 + 287$)	Non-homologues ($N = 13900 + 59256$)	Odds ratio 2
Same first EC number digit	1.26	0.31	4.06
Different first EC number digit	0.58	0.14	4.14
Odds ratio 1	2.17	2.21	

Pseudomonas aeruginosa (odds $n(<3$ steps; $M = 26 + 2168)/n(\geq 3$ steps; $M = 54 + 6865))$

	Homologues ($N = 61 + 19$)	Non-homologues ($N = 1807 + 7226$)	Odds ratio 2
Same first EC number digit	0.36	0.17	2.12
Different first EC number digit	0.32	0.11	2.91
Odds ratio 1	1.13	1.55	

Salmonella typhimurium (odds $n(<3$ steps; $M = 147 + 6612)/n(\geq 3$ steps; $M = 166 + 16188))$

(*continued*)

Table 2 Continued

Aquifex aeolicus (odds $n(<3$ steps; $M = 283 + 6137)/n(\geq 3$ steps; $M = 100 + 11912)$)

	Homologues ($N = 253 + 130$)	Non-homologues ($N = 3610 + 14313$)	Odds ratio 2
	Homologues ($N = 175 + 138$)	Non-homologues ($N = 4788 + 18012$)	Odds ratio 2
Same first EC number digit	1.26	0.22	5.73
Different first EC number digit	0.57	0.13	4.38
Odds ratio 1	2.21	1.69	

The Table gives odds $n(<3$ steps; $M = a + b)/n(\geq 3$ steps; $M = a + b)$ for specific organisms, where a represents the number of pairs of homologues and b represents the number of pairs of non-homologues. In the homologues/non-homologues boxes, $N = c + d$, where c represents the number of pairs of enzymes that belong to the same class and d represents the number of pairs of enzymes that belong to a different class. The column labeled homologues presents results for pairs of enzymes that are homologues, while the column labeled non-homologues presents results for pairs of enzymes that are not homologues. The rows labeled same first EC number digit present the results for pairs of enzymes that belong to the same class, while the rows labeled different first EC number digit present the results for pairs of enzymes that do not belong to the same class. Each entry in the Table compares the odds that a pair of enzymes has members that are less than three steps away in the metabolic network compared with those of the pair having members that are three or more steps away form each other. For example, the number under same first digit/homologues gives us the odds that the members of the pair are less than three steps away from each other. The column labeled odds ratio 2 presents the odds ratio between pairs of homologues and non-homologues. The row labeled odds ratio 1 shows the odds ratio between pairs of enzymes that belong to the same class and pairs of enzymes that belong to different classes.

levels.[19–24] One should therefore consider the advantage that such a localized evolution might bring to the physiology of the cell. New enzymes appear in a metabolic network by duplication of another enzyme in the genome. It is likely that the new enzyme will retain some partial original activity. This will disrupt the metabolism because there will now be two enzymes producing and consuming the reactants that had been produced and consumed by just one, even though now we also have an enzyme that is necessary for the return of the network to a balanced state. If the original enzyme is close to the new one in the metabolic network, disruption of the physiology of the cell is likely to be restricted to a part of metabolism that is already disrupted. However, if the old enzyme is not close to the new enzyme in the network, then there will be a disruption of the physiological state of another part of metabolism, further decreasing the fitness of the cell. This argument is less compelling if we consider enzymes that use metabolites with a high degree of promiscuity. Imbalanced levels of such a metabolite are likely to disrupt many parts of metabolism, making it less relevant whether an enzyme has evolved locally in the network or not. For example, many oxyreductases use NAD(P) as a reactant. Indeed, once we eliminate this promiscuous metabolite from the connectivity matrix, the average distance between homologues in class 1 increases significantly, as can be seen in Figure 6. These physiological considerations do not explain the clustering effect of chemical function in metabolic networks that our results suggest as a feature in the metabolic networks of many organisms. The clustering can be explained more readily if we consider the constraints that probably exist for enzyme evolution. It seems more likely that an enzyme can be mutated productively into another enzyme that performs a similar function than into one that performs a totally different function. This, combined with the physiological arguments presented above, provides a good background for the evolution of enzymes that have similar chemistry close to each other in the metabolic network, thus leading to clustering of chemical function. We note, however, that our analysis does not take into account complexities such as when enzymes are physically distant from one another by virtue of cellular location or the effects of differential expression during development and/or cell type. Once systematic and reliable database information about these aspects is available, it will be important to take this into account and include it in our analysis.

The present work has highlighted the limits of studying the evolution of enzymes in the context of pathways. Figure 8 shows an example of an enzyme that would have been missed as having evolved locally if we had used the pathway structures in KEGG. Serine tRNA synthetase would not have been determined to be close to prephenate dehydratase when we use the pathway scheme presented in KEGG. Many more examples like this exist. The median percentage of homologues that would not have been found to be close by in the network had we used the KEGG pathway definitions for well-represented genomes is close to 30% or higher if promiscuous metabolites are considered.

The network approach we used in this work provided valuable insight into the evolution of enzymes, at the same time allowing us to look into the topological aspects of metabolism, extending previous results by other groups.[25–29] We determined that approximately 18% of the enzymes represent bridges or bottlenecks in the superset of metabolic networks (i.e. the network that includes all possible enzyme reactions), connecting parts of metabolism that do not exchange material if the bottleneck enzyme is knocked out of the genome. The number of bridges is higher for individual organisms, because organisms do not have all the enzymes in the superset. A more detailed analysis

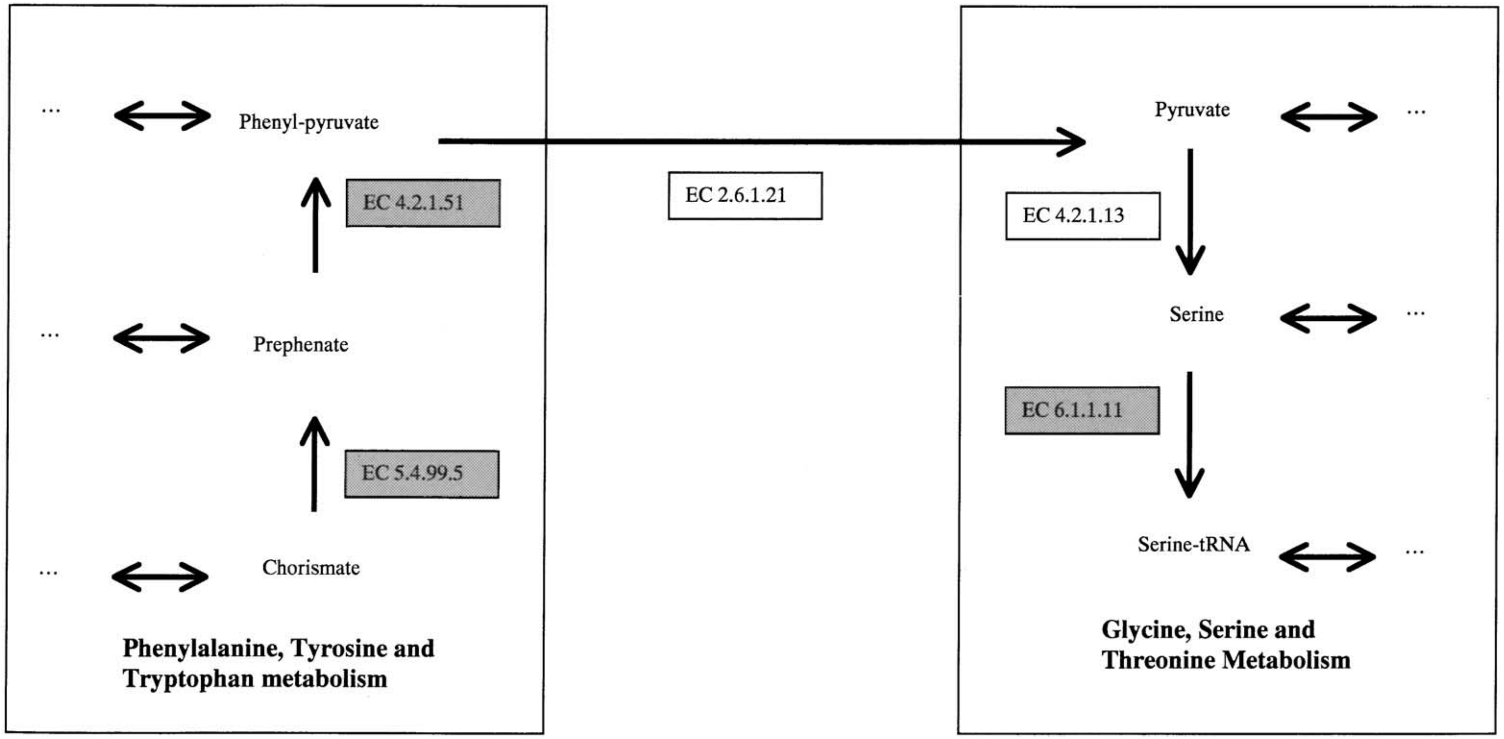

Figure 8. Example of local evolution that would have been missed if the analysis was done using KEGG pathway definitions instead of a network approach: EC 5.4.99.5, chorismate mutase; EC 4.2.1.51, prephenate dehydratase; EC 2.6.1.21, D-alanine transaminase; EC 4.2.1.13, L-serine dehydratase; EC 6.1.1.11, serine-tRNA synthetase. Homologues are presented in gray boxes. EC 6.1.1.11 is only two reactions away from EC 4.2.1.51 but it would not have been detected as being close because it is in a different KEGG pathway.

of each organism identifies a larger number of enzymes that may be organism-specific bottlenecks. However, one must be careful with these organism-specific bottlenecks, because they may be due to errors and limitations in genome annotation. The 18% general bottleneck enzymes are likely to be important targets for mutations causing metabolic diseases and a more detailed study of these enzymes should prove useful.

The use of several fully sequenced genomes in our work circumvents the problem of using only one genome to study the evolution of metabolic networks. The replacement of a pathway approach with a network approach leads to a complex picture of evolution in metabolic networks with a high percentage of non-local evolution but also a significant over-occurrence of local evolution with local islands of homologous enzymes that share similar chemistry.

Materials and Methods

Database construction

We used LIGAND (version 19.0) and BRENDA (Internet version of August 2001†) for Metabolic Network general reconstruction. Version 4.01 of Mathematica[30] was used to analyze the characteristics of metabolic network graphs. We used SWISSPROT (version 39), WIT (version December 1999) and KEGG (current version) for reconstructing the enzyme sequence of each organism and SCOP (version 1.50) for structure-based homology. We ran PSIBLAST version 4.2.3 with the default parameters and an *E*-value cut off of 0.001 to find the homologues among the different enzymes within each organism.

Distance between enzymes and local evolution

Whether two homologues have evolved by retro-evolution or by recruitment is traditionally decided by determining whether they belong to the same pathway(s). Once we remove the traditional definition of pathways, the problem of determining what is local and what is long-distance evolution arises. An extremely conservative approach would consider that local evolution occurs only when homologues are consecutive enzymes in the network. However, this criterion is much stricter than what is used in traditional studies, where enzymes are considered to have retro-evolved if they belong to the same pathway, independent of their distance in the pathway. Connectivity studies (Jeong *et al.*,[26] and this study, data not shown) show that the average number of reactions for a given metabolite to be transformed into itself again is between 3 and 5. Thus, we defined the threshold distance for local evolution in the network as being 2, which is the only integer larger than 1 and smaller than 3.

Random shuffling

Random shuffling was done using Mathematica.[30] To determine how significant the results for distance between enzymes, chemistry conservation analysis, odds and odds ratios are for each organism, we built a square matrix where each row (column) represents an enzyme (indexed in our database). The entries in the row (column) are 1 if the enzyme in the column (row) is homologous to the enzyme in the row (column) and 0 otherwise. The index that identifies the columns is then shuffled randomly (and the row index is made consistent with it). We then analyze distance and chemistry conservation in the random network. This random procedure was repeated 1000 times for each organism and each time the appropriate numbers were stored. Histograms were built with those numbers and proportions to compare with the actual values of the odds and odds ratios. Note that tests such as χ^2 or contingency tables would be inappropriate due to non-independence of the numbers. If enzymes A, B and C are homologues, and A is one step from B, and B is one step from C, then A and C must be less than three steps apart and thus the result from A–C in not independent of the results for A–B and B–C.

† http://www.brenda.uni-koeln.de/

Acknowledgments

R.A. was supported by the Imperial Cancer Research Fund (ICRF) and the BBSRC. R.A.G.C. was supported by the ICRF. M.J.E.S. was supported by the ICRF and Imperial College. We thank Dr Peter Sasieni (ICRF) for statistical advice.

References

1. Horowitz, N. H. (1945). On the evolution of biochemical syntheses. *Proc. Natl Acad. Sci. USA*, **31**, 152–157.
2. Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425.
3. Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov, E., Kyrpides, N. *et al.* (2000). EWIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucl. Acids Res.* **28**, 123–125.
4. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (1992). *Enzyme Nomenclature*, Academic Press, New York.
5. Teichmann, S. A., Rison, S. C. G., Thornton, J. M., Riley, M., Gough, J. & Chothia, C. (2001). The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *J. Mol. Biol.* **311**, 693–708.
6. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
7. Tsoka, S. & Ouzounis, C. A. (2001). Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*. *Genome Res.* **11**, 1503–1510.
8. Petsko, G. A., Kenyon, G. L., Gerlt, J. A., Ringe, D. & Kozarich, J. W. (1992). On the origin of enzymatic species. *Trends Biochem. Sci.* **18**, 372–376.
9. Saqi, M. A. S. & Sternberg, M. J. E. (2001). A structural census of metabolic networks for *E. coli*. *J. Mol. Biol.* **313**, 1195–1206.

10. Copley, R. R. & Bork, P. (2000). Homology among $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* **303**, 627–640.
11. Gerlt, J. A. & Babbit, P. C. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.* **70**, 209–246.
12. Karp, P. D. (2000). An ontology for biological function based on molecular interactions. *Bioinformatics*, **16**, 269–285.
13. Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **28**, 27–30.
14. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
15. Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I. *et al.* (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucl. Acids Res.* **29**, 2994–3005.
16. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
17. Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer from genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233–249.
18. Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Struct. Funct. Genet.* **42.1**, 98–107.
19. Savageau, M. A. (1976). *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*, Addison-Wesley, Reading, MA.
20. Hlavacek, W. S. & Savageau, M. A. (1997). Completely uncoupled and perfectly coupled gene expression in repressible systems. *J. Mol. Biol.* **266**, 538–558.
21. Alves, R. & Savageau, M. A. (2000). Effect of overall feedback inhibition in unbranched biosynthetic pathways. *Biophys. J.* **79**, 2290–2304.
22. Heinrich, R. & Schuster, S. (1998). The modeling of metabolic systems, structure, control and optimality. *Biosystems*, **47**, 61–77.
23. Melendez-Hevia, E. & Isidoro, A. (1985). The game of the pentose phosphate cycle. *J. Theor. Biol.* **117**, 251–263.
24. Miettenthal, J. E., Yuan, A., Clarke, B. & Scheeline, A. (1993). Design metabolism: alternative connectivities for the pentose phosphate pathway. *Bull. Math. Biol.* **60**, 815–856.
25. Bhalla, U. S. & Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways. *Science*, **283**, 381–387.
26. Jeong, H., Tombor, B., Albert, R., Oltval, Z. N. & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.
27. Wagner, A. & Fell, D. A. (2001). The small world inside large metabolic networks. *Proc. Roy. Soc. Lond. ser. B*, **268**, 1803–1810.
28. Banavar, J. R., Maritan, A. & Rinaldo, A. (1999). Size and form in efficient transportation networks. *Nature*, **399**, 130–132.
29. Qian, J., Luscombe, N. M. & Gerstein, M. (2001). Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.* **313**, 673–681.
30. Wolfram, S. (2001). *Mathematica: A System for Doing Mathematics by Computer*, Addison-Wesley, Reading, MA.

Edited by F. E. Cohen

(Received 3 January 2002; received in revised form 20 March 2002; accepted 16 May 2002)