THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 286, NO. 51, pp. 43994–44004, December 23, 2011

Origins of Specificity and Promiscuity in Metabolic Networks[S]

Received for publication, June 29, 2011, and in revised form, November 1, 2011 Published, JBC Papers in Press, November 3, 2011, DOI 10.1074/jbc.M111.274050

Pablo Carbonell[‡1], **Guillaume Lecointre**[§], **and Jean-Loup Faulon**[‡2]

From the [‡]*Institute of Systems and Synthetic Biology, University of Evry, 91030 Evry, France and* [§]*UMR 7138 Systématique Adaptation Evolution, Département Systématique et Evolution, Muséum National d'Histoire Naturelle, 75005 Paris, France*

Background: How enzymes evolved to their present form is linked to how extant metabolic pathways emerged.
Results: Chemical diversity of reactions parallels enzyme phylogenetic diversity across the tree of life.
Conclusion: Enzyme promiscuity plays a prominent role in the evolution of metabolic networks.
Significance: Learning about the mechanisms of enzyme evolution might assist us with the identification of primeval catalytic functions and minimal metabolism.

How enzymes have evolved to their present form is linked to the question of how pathways emerged and evolved into extant metabolic networks. To investigate this mechanism, we have explored the chemical diversity present in a largely unbiased data set of catalytic reactions processed by modern enzymes across the tree of life. In order to get a quantitative estimate of enzyme chemical diversity, we measure enzyme multispecificity or promiscuity using the reaction molecular signatures. Our main finding is that reactions that are catalyzed by a highly specific enzyme are shared by poorly divergent species, suggesting a later emergence of this function during evolution. In contrast, reactions that are catalyzed by highly promiscuous enzymes are more likely to appear uniformly distributed across species in the tree of life. From a functional point of view, promiscuous enzymes are mainly involved in amino acid and lipid metabolisms, which might be associated with the earliest form of biochemical reactions. In this way, results presented in this paper might assist us with the identification of primeval promiscuous catalytic functions contributing to life's minimal metabolism.

Recent phylogenetic studies of cellular metabolism suggest a core set of highly conserved enzymes involved in amino acid, energy, carbohydrate, and lipid metabolism, which is likely to be associated with the ancestral form of the extant metabolic network (1, 2). However, the essentiality of this core for life and the existence of a universally essential minimal gene set are currently under debate, because essential genes show a diverse overall organization across multiple organisms in metabolic maps (3). This diversity might be explained, in part, by the role that enzyme promiscuity plays, providing a common background level of metabolism to the organisms (4). Thus, how enzymes have evolved to their present form in organisms appears to be linked to the question of how metabolic pathways emerged and configured extant metabolic networks (5). In fact, although many enzyme families and superfamilies sharing structural and functional features have been identified, it has been recognized that distant evolutionary relationships are expected to be difficult to identify solely from global sequence comparison (6). Here, we propose to extend previous studies of metabolic network evolution (7–10) by studying the evolutionary aspect of enzyme promiscuity (*i.e.* of the latent capabilities of enzymes to broaden their specificity to substrates or to process diverse biochemical reactions) (11, 12).

An early theory on how metabolic pathways have evolved was the retrograde evolution model, proposed by Horowitz (13), which states that the earliest biosynthetic pathways evolved in a backward direction in response to depletion of substrates from the environment. In the retrograde model, the recruitment of an enzyme capable of synthesizing the depleted substrate from some other available precursor brings a selective advantage to the organism. In contrast, the patchwork evolution model originally proposed by Jensen (14) states that primitive enzymes possessed broad substrate specificity and that gene duplication and divergence led to the specialized and increased metabolic efficiency observed in extant enzymes. Both models might be interpreted in the context of a similar evolutionary mechanism, driven either by selective pressure of the substrate in the retrograde model or by the chemical reaction in the patchwork model (7). A current view assumes that although the retrograde model could have played a role on the evolution of the ancestral form of the metabolic network, the patchwork evolution model is the predominant player in modern metabolic enzyme evolution (9, 15).

As has been observed by Tawfik's group (16), an intriguing aspect of evolved enzymes is that their promiscuous activities or latent functions can usually be enhanced and diversified without impairing the traditional function. There are, however, some practical limits in the efficiency that modern enzymes have achieved, as if they had evolved and become specialized up to the required level in the cell to perform their task (17), suggesting that keeping some level of latent promiscuous capabilities might be a selective advantage allowing adaptation to environmental pressure (18). In many cases, such enhancement of latent catalytic activities does not seem to trade off with parallel decreases in the original function, in contradiction with the commonly accepted assumption that broad substrate acceptance generally comes at the price of low reaction turnover

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Tables S1 and S2 and Figs. S1–S10.

[1] Supported by Genopole through an Actions Thématiques Incitatives de Genopole grant.
[2] Supported by Agence Nationale de la Recherche Chaire d'Excellence. To whom correspondence should be addressed: Genopole Campus 1, Genavenir 6, 5 Rue Henri Desbruères, 91030 EVRY Cedex, France. Fax: 33-1-6947-44-37; E-mail: jean-loup.faulon@issb.genopole.fr.

TABLE 1
Data set of enzyme entries, organisms, and their catalytic activities
For each database, values are given for enzymes belonging to organisms in the tree of life and for total enzymes (in parentheses).

	Subset (total data set) from KEGG with annotated reactions and organisms in the tree of life	Subset (total data set) from MetaCyc with annotated reactions and organisms in the tree of life
Total entries	152,041 (404,205)	1080 (7287)
Non-promiscuous enzymes	148,594 (394,732)	927 (6465)
Promiscuous enzymes	3447 (9293)	153 (822)
Total EC numbers	770 (2039)	766 (2018)
Total organisms	356 (1096)	120 (1130)

numbers (19). Mutations in enzymes that are neutral with respect to the protein's primary biological function can provide in many cases a way to induce substantial changes in other promiscuous functions present in the enzyme at lower efficiency levels (20). Subsequent gene duplications might allow natural selection to transform one of the genes into a protein with high efficiency for the new functional role (21).

Furthermore, enzyme evolution might be influenced by the structure and function of the metabolic network. For instance, enzymes in central parts of the metabolism, such as the tricarboxylic acid cycle, evolve more slowly than less connected enzymes (22). Also, enzymes carrying high metabolic fluxes under natural biological conditions experience higher evolutionary constraints. Genes encoding enzymes with high connectivity and high metabolic flux have higher chances to retain duplicates in evolution. Similarly, highly expressed enzymes have been observed to evolve slowly (23). Finally, another mechanism that we want to examine in this study is how the environment influences enzyme evolution. The presence of substrates plays a pivotal role in determining whether latent catalytic abilities become manifest in a novel enzyme (24), as has been observed in recent years with the emergence of enzymes that degrade novel synthetic chemicals and with the alarming evolution of drug resistance.

In our study of origins of enzyme multispecificity and promiscuity, we plan to overcome the limitations of sequence comparison studies by performing a comparison between enzyme promiscuity and phylogenetic diversity of species. Our hypothesis is that if promiscuous enzymes are more evolvable (*i.e.* they have a greater capacity to evolve, they should be present in distant species, and their corresponding common ancestors should be deeper in the tree of life). Conversely, non-promiscuous enzymes, which exhibit specific catalytic activities, might be found within smaller groups of more closely related species. Nevertheless, our approach based on enzyme promiscuity might also be prone to investigation bias, since a well known observed fact is that the number of identified cross-reactants increases exponentially with the number of tested ligands (18). Therefore, a rigorous definition of enzyme promiscuity needs to be introduced and applied to an unbiased data set. The term "enzyme promiscuity," however, has been used with different interpretations. Several authors reserve the term "promiscuity" in order to describe enzyme activities other than their native function (18), a sort of adventitious secondary activity (25), which might have appeared accidentally or have been induced (17), whereas "multispecificity" is used to designate enzymes with the ability to transform a whole range of substrates (18). A further classification of enzyme promiscuity distinguishes between three different forms (17): enzyme condition promiscuity (other reaction conditions than the natural one), enzyme substrate promiscuity (broad substrate specificity), and enzyme catalytic promiscuity (different chemical transformations). To consider two chemical transformations as different, the functional groups involved and/or the transition states of the two reactions must differ (26).

In order to clarify these concepts, it is necessary to define a quantitative measure of the levels of enzyme promiscuity. Nath and Atkins (21) introduced a quantitative index of promiscuity, which quantifies the degree of similarity between different substrates of the same enzyme. This index, however, does not take into account the chemical transformation, and it is, therefore, more suitable for measuring cross-reactivity between multispecific enzymes. Khersonsky and Tawfik (18) propose the use of EC number comparison in order to assess the degree of promiscuity. Multispecific enzymes should differ at most in the fourth digit, whereas differences in the third, second, or even first digits would refer to catalytic promiscuity. The authors, however, pointed out some cases where EC numbers of even the first digit are misleading, because they actually refer to transformations with considerable similarity in the chemistry of catalysis. In order to circumvent the limitations of these definitions, we propose here to use a quantitative measure of similarity between reactions based on molecular signatures, a representation of the molecular graph (27), which provides a consistent way to characterize enzyme multispecificity and promiscuity (28).

EXPERIMENTAL PROCEDURES

Data—We downloaded from the KEGG (release version 50) (29) and MetaCyc (release version 15.1) (30) databases the list of proteins with annotated catalytic activity. Enzymes annotated with more than one EC number, a basic definition of enzyme promiscuity, correspond to ~2% of the total (Table 1). Annotations corresponding to partial EC numbers were not considered in the study. Redundancy in the KEGG database was removed within each set at a threshold of 90% of sequence similarity. In our study, we were interested in those sequences from the data set belonging to organisms in the reference tree of life (see definition below). Additionally, in order to apply our analysis of chemical similarity, the information in KEGG about the biochemical reactions that catalyze the enzymes was collected. Finally, we obtained 152,041 enzyme sequences from KEGG verifying simultaneously these specifications (Table 1). In the case of MetaCyc, we were able to identify 1080 entries verifying these specifications (Table 1).

TABLE 2

Number of species by taxonomic group

Shown are the numbers of species contained in different taxonomic groups in the reference tree of life.

Taxonomic group	Species in taxonomic group
Archaea	Crenarchaeota: 3
	Euryarchaeota: 9
	Nanoarchaeota: 1
Eubacteria	Actinobacteria: 5
	Bacteroidetes/Chlorobi group: 4
	Chlamydiae/Verrucomicrobia group: 2
	Chlroflexi: 1
	Cyanobacteria: 4
	Deinococcus-Thermus: 2
	Firmicutes: 10
	Fusobacteria: 1
	Planctomycetes: 2
	Proteobacteria: 35
	Spirochaetes: 3
	Tenericutes: 3
	Thermotogae: 1
Eukaryota	Alveolata: 4
	Amoebozoa: 1
	Animals: 10
	Excavobionta: 2
	Fungi: 3
	Plants (*i.e.* Chlorobionta): 3

We used as the reference tree of life the phylogenetic tree reconstructed by Ciccarelli *et al.* (31) from a multiple alignment of 31 DNA orthologs occurring in 191 organisms with sequenced genome. In this set, data were curated by these authors in order to avoid horizontal gene transfer events, and phylogenetic inferences were made by maximum likelihood and the JTT substitution matrix by using PHYML (32). In our study, we grouped together branches of the same genus in this reference tree, resulting in 108 species (see supplemental Table S1). The apparent speed of sequence evolution can be measured as the cumulative branch length distance from a species to the root, being the root placed through automatic midpoint rooting. The choice of such a rooting is not a problem for the purpose of our study because midpoint rooting falls somewhere between the three domains, and our results are actually based on what happens within each of them. In the present paper, we refer only to relative divergence times (*i.e.* the ordering of divergence times of taxonomic groups), not to absolute ones. Divergence times of broad taxonomic groups, when considered, are those given in Ref. 33.

To link enzyme and phylogenetic information, organisms in the databases were mapped into the tree of life. Different strains from the same species were assigned to the same organism. When several organisms were matched to the same taxon in the tree of life, values were averaged. We were able to assign 108 species in the tree of life to 356 organisms in KEGG (33%) and 120 in MetaCyc (11%) (See Table 2 and supplemental Tables S1 and S2 for details).

Evaluation of Enzyme Promiscuity—The term "enzyme promiscuity" is used throughout to refer to the general concept of the ability of an enzyme to process multiple substrates or reactions. Furthermore, we introduce more specific definitions of enzyme promiscuity in Table 3 at the enzyme level by looking at the chemical diversity of the reactions that the enzyme can process, the enzyme chemical diversity, and at the catalytic activity level by looking at the list of reactions that can be simultaneously processed by enzymes annotated for this activity, the enzyme latent promiscuity.

TABLE 3

Summary of enzyme promiscuity and phylogenetic diversity measures used in the present study

Term	Level	Definition	Formula
Promiscuous annotations for an EC number	EC number	EC numbers that have been simultaneously annotated with a given EC_i	$p(EC_i) = EC_{i1} \dots EC_{im}$
EC number latent promiscuity	EC number	Total number of promiscuous annotations for a given EC_i	$lp(EC_i) = \lvert p(EC_i) \rvert = m$
Organisms annotated with an EC number	EC number	Organisms containing genes that have been annotated for a given EC_i	$o(EC_i) = o_{i1} \dots o_{ir}$
EC number phylogenetic diversity	EC number	Average intertaxa branch-length distance $\delta(o_{ij}, o_{jk})$ between organisms o_{ij} and o_{ik} annotated with a given EC_i	$pd(EC_i) = \frac{1}{r(r-1)} \sum_{j=1}^{\lvert o(EC_i) \rvert} \sum_{k=j+1}^{\lvert o(EC_i) \rvert} \delta(o_{ij}, o_{ik})$
Promiscuous annotations for a reaction	Reaction	Reactions that have been simultaneously annotated with a given reaction R_i	$p(R_i) = R_{i1} \dots R_{in}$
Reaction chemical diversity	Reaction	Average chemical dissimilarity $T_c(R_{ij}, R_{ik})$ between promiscuous annotations R_{ij} and R_{ik} for a given reaction R_i	$d(R_i) = 1 - \frac{1}{n(n-1)} \sum_{j=1}^{\lvert p(R_i) \rvert} \sum_{k=j+1}^{\lvert p(R_i) \rvert} T_c(R_{ij}, R_{ik})$
Organisms annotated with a reaction	Reaction	Organisms containing genes that have been annotated for a given reaction R_i	$o(R_i) = o_{i1} \dots o_{iq}$
Reaction phylogenetic diversity	Reaction	Average intertaxa branch-length distance $\delta(o_{ij}, o_{jk})$ between organisms o_{ij} and o_{ik} annotated with a given reaction R_i	$pd(R_i) = \frac{1}{q(q-1)} \sum_{j=1}^{\lvert o(R_i) \rvert} \sum_{k=j+1}^{\lvert o(R_i) \rvert} \delta(o_{ij}, o_{ik})$

First, we evaluated enzyme chemical diversity by computing the average chemical dissimilarity between the reactions in the list of biochemical reactions annotated for the enzymes in the databases. Chemical similarity was evaluated for pairs of reactions R_A and R_B using the well known Tanimoto coefficient $T_c(R_A,R_B)$, a number ranging between 0 (dissimilar reactions) and 1 (identical reactions). In this study, T_c was evaluated by using the molecular signature of the reactions (28). Molecular signatures are vectors whose components correspond to canonical representations of the graph associated with molecular structures (27). If $G = (V,E)$ is a molecular graph, where vertices V correspond to atoms and edges E correspond to bonds, then the molecular signature of G is given by Equation 1,

$$\sigma(G) = \sum_{x \in V} \sigma(x) \qquad \text{(Eq. 1)}$$

where $\sigma(x)$ represents the atomic signature of G rooted at atom x. The molecular signature of reaction R,

$$s_1S_1 + \dots + s_nS_n \rightarrow p_1P_1 + \dots + p_mP_m$$

REACTION 1

where s_i and p_i are the stoichiometric coefficients of substrates S_i and products P_j, is defined by the difference between the signatures of products and substrates,

TABLE 4
Enzyme efficiency data set
Shown are the numbers of catalytic reactions, activities, and organisms with experimental efficiency from BRENDA, which were considered in our study.

Definition	Total
Catalytic reactions	4313
EC numbers	861
Reactions with (k_{cat}/K_m) full information	485
EC with (k_{cat}/K_m) full information	210
Organisms	620
Organisms (k_{cat}/K_m) full information	151

$$\sigma(R) = \sum_{j=1}^{m} p_j \sigma(P_j) - \sum_{i=1}^{n} p_i \sigma(S_i) \quad \text{(Eq. 2)}$$

Based on the previous definition, reaction chemical similarity between the pair of reactions R_A and R_B is given by Equation 3,

$$T_c(R_A,R_B) = \frac{\|\sigma(R_A) \cdot \sigma(R_B)\|}{\|\sigma(R_A)\|^2 + \|\sigma(R_B)\|^2 - \|\sigma(R_A) \cdot \sigma(R_B)\|} \quad \text{(Eq. 3)}$$

where the chemical points represent the dot product between the signatures.

Second, our definition of latent enzyme promiscuity for one EC number was given by the number of different activities where enzymes annotated with this activity have been observed to participate. It was computed as follows. We listed all enzymes that were annotated with a given EC number, and we counted the total number of different EC numbers where these enzymes have been annotated.

Evaluation of Phylogenetic Diversity—The phylogenetic distance between two species o_i and o_j was computed as the inter-taxa branch length distance $\delta(o_i,o_j)$ of the two taxa within the tree of life (31); this is generally called a patristic distance. This value corresponds to their average sequence divergence in the multiple alignment of 31 orthologs (31). Phylogenetic diversity of a set of organisms was computed as in (34), by averaging the patristic or taxonomic distance between individual organisms. Phylogenetic diversity of a reaction or EC number was estimated by computing the phylogenetic diversity of organisms annotated with this reaction or catalytic function.

Catalytic Function Classification and Efficiency Estimation—We used two types of functional classification for enzymes, one based on pathway modules and another on ortholog groups. KEGG pathway modules for each catalytic function (EC number) were downloaded from the KEGG site. Cluster of ortholog groups in *Escherichia coli* were downloaded from the NCBI, National Institutes of Health, Web site. Catalytic efficiency was computed from the BRENDA database (35) (see Table 4). For each EC number, we stored the values of K_m and K_{cat} of those reactions involving native enzymes (*i.e.* excluding mutants) where we were able to match substrate information with KEGG by means of the following procedure. For each entry in Brenda corresponding to an enzyme, we went through the list of substrates processed by this enzyme, flagging the entry as consistent if the same substrate was annotated in KEGG for the EC number that corresponds to the enzyme. Finally, we collected all available catalytic efficiency values in the consistent entries.

RESULTS

Reaction and Annotation Level Enzyme Promiscuity—To characterize enzyme promiscuity, our basic definition is given by the chemical diversity found among the reactions that an enzyme can process. Typically, the number of biochemical reactions annotated for an enzyme is higher than the number of catalytic functions according to their EC classification because enzymes can use the same catalytic activity to process more than one substrate (supplemental Fig. S1). Therefore, in order to get a more accurate estimate of the enzyme catalytic capabilities, enzyme promiscuity at the reaction level was measured by computing the average chemical dissimilarity (see Table 3) between the reactions in the list of biochemical reactions annotated for each enzyme in the KEGG (29) or MetaCyc database (30). Contrarily to the two previous parameters (number of annotated reactions and EC numbers), our definition based on reaction chemical diversity provides an overall estimate of enzymatic capabilities and a quantitative characterization of enzyme promiscuity, which is less likely to be affected by the bias arising from redundant annotations or arbitrary/inconsistent classifications (36). In fact, it is not always the case that an enzyme annotated with a large number of reactions is able to process chemically more diverse substrates (see supplemental Fig. S2). The number of substrates, however, might be underestimated when they are labeled uniquely in generic terms, such as "alcohol" or "dialkyl ketone," an issue that was found in 33 EC numbers (4.3% of total ECs in the data set) in our data set. Of these, only 15 EC numbers (1.9% of total ECs in the data set) were annotated uniquely with one generic substrate. The effect of the presence of this small set of enzymes mislabeled as non-promiscuous, thus, is not going to alter significantly any general trend observed in the results.

To further characterize enzyme promiscuity, our second definition is at the EC number annotation level, which might be more properly defined as *latent* enzyme promiscuity because its definition is the number of different activities where enzymes have been annotated simultaneously with a given EC number. Therefore, this value is a measure of the total number of potential catalytic activities that an enzyme can process, based on the different functions that have been observed within its orthologs across the species.

Both definitions of promiscuity at the reaction and annotation activity level are interrelated because enzymes with the ability of catalyzing a greater chemical diversity are potentially more likely to be involved in a greater number of catalytic functions across organisms. On average, we observed this trend ($r = 0.77$, $p = 2.6 \times 10^{-2}$) (see Fig. 1*A*). In some cases, however, enzymes annotated with one EC number, and therefore with no latent promiscuity, are actually able to process a significant number of chemically diverse substrates, a fact that could be related to the EC number definition. For example, aspartate aminotransferases (EC 2.6.1.1), which are enzymes with no latent promiscuity according to KEGG annotation, can potentially process tyrosine, phenylanine, and trytophan as well (37), making its reaction chemical diversity 0.51. This value is one of the highest in the group of EC numbers with no latent promiscuity. Therefore, reaction level chemical diversity is in general a

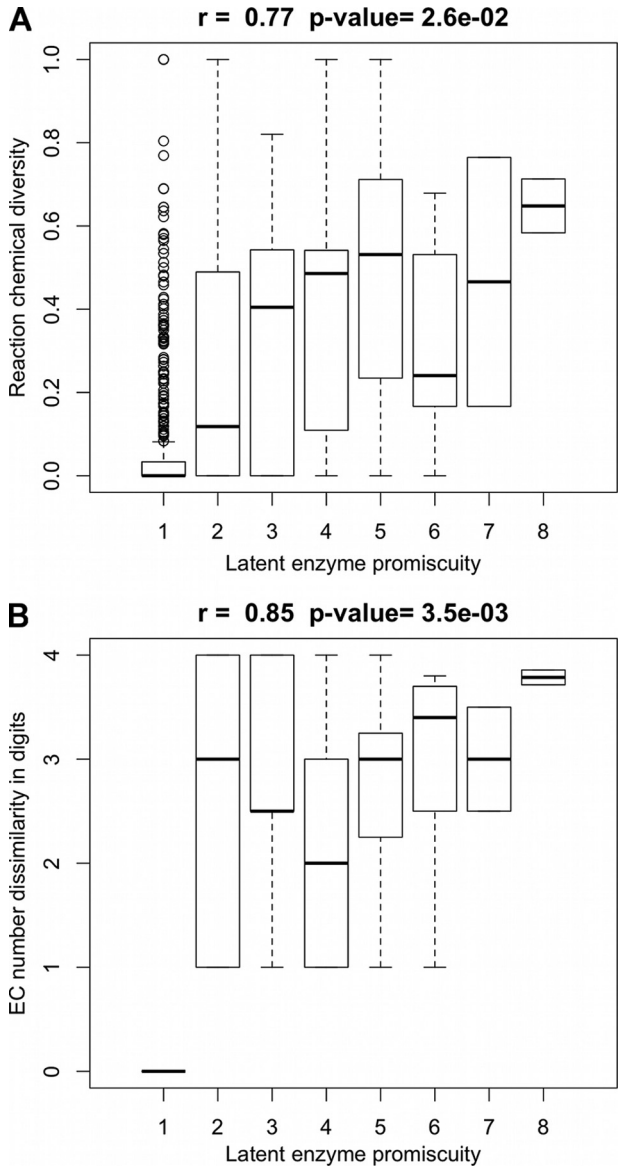


FIGURE 1. **Relationship between latent promiscuity and reaction diversity.** *A*, relationship between latent promiscuity and substrate chemical diversity; *B*, comparison between the measure of promiscuity at the annotation level (EC number) and based on EC dissimilarity in digits.

more accurate promiscuity measure than EC number annotation level.

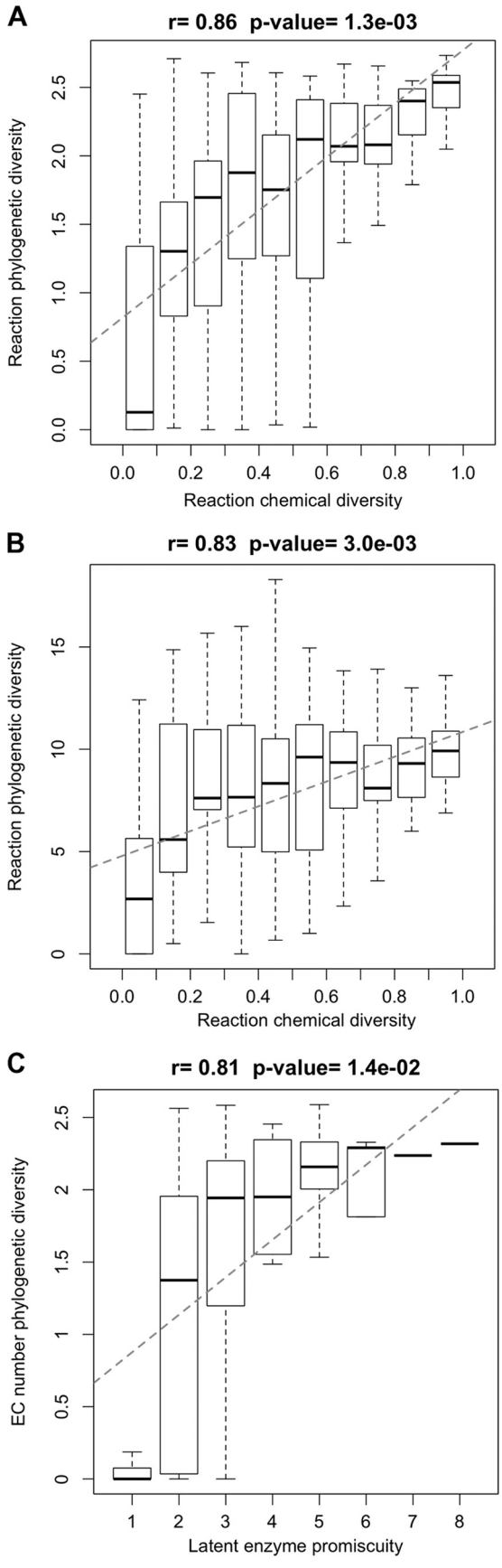
To examine how both of our definitions of promiscuity relate to the definition in Ref. 18, which is based on EC difference in digits, we plot in Fig. 1*B* for each enzyme the relationship between EC number dissimilarity and latent promiscuity. As it is shown in the plot, higher values of latent enzyme promiscuity usually correspond to a greater dissimilarity in digits, whereas lower values appear more scattered according to their EC number dissimilarity. Nevertheless, both measures are highly correlated ($r = 0.85$, $p = 3.5 \times 10^{-3}$).

Phylogenetic Distance in Trees of Life and Enzyme Promiscuity—To perform our study of origins of enzyme promiscuity, we need an unbiased and comprehensive data set of phylogenetic and metabolic data. Phylogenetic diversity of any subset in a list of species might be estimated once a reference tree of life is established. As described under "Experimental Procedures," we took here as a reference the tree in Ref. 31, where genes in 191 species (see Table 1) with completely annotated genomes were carefully selected in order to minimize horizontal gene transfer effects. The number of species was further reduced to 108 when branches of the same genus were grouped together (see supplemental Tables S1 and S2). In order to estimate phylogenetic diversity, we took two approaches: tree distances calculated from sequence data following the procedure in the reference tree (31) and a tree based solely on evolutionary relationships through the use of phylogenetic systematics (38) (see dendrograms in supplementary Fig. S3).

The reason for using two reference trees in this study is motivated by the fact that patristic distances (sum of the length of the branches) extracted from sequence alignment data should be used with caution. We hope that such pairwise distances will reflect phylogenetic diversity; however, some species with a much higher rate of DNA change than others will bias the estimation by having a higher terminal branch length. To limit the risk of facing such bias, results are presented in this section comparatively for the estimation of phylogenetic diversity using pairwise patristic distances on the reference tree based on sequence data and for the estimation of phylogenetic diversity using a purely taxonomic (or topological) approach by counting the number of nodes separating each pair of species in the reference tree based on evolutionary relationships (*i.e.* by looking at the tree without its branch lengths).

To investigate how effects associated with the sequence data selection, such as the unequal rate of change among selected orthologs, the effects from horizontal gene transfer events (39, 40), and species sampling (41), might be introducing some bias in the tree of life, we evaluated the distribution of phylogenetic distances, finding that some taxa such as "prokaryotes" were overrepresented in our reference tree due to their greater availability of sequenced genomes. In fact, we observed that the distribution of phylogenetic distances (patristic distances from a maximum likelihood tree calculated from aligned DNA sequences) between organisms in the reference tree (see "Experimental Procedures" for definitions) follows a bimodal distribution (see supplemental Fig. S4), illustrating the fact that some phyla are overrepresented in the tree, as is the case for proteobacteria and firmicutes (see Table 1). From this form of distribution, we might expect that phylogenetic distances between pairs in a random set of species would not monotonically increase with sample size (see supplemental Fig. S5). Therefore, we can assume for this study that obtaining higher average phylogenetic distances can be properly interpreted as greater phylogenetic diversity or spread of species across the tree of life rather than a bias coming from a tree artifact.

To further evaluate the effect of investigation bias, we relied on the BRENDA database (35), which is one of the most comprehensive resources on experimental enzymatic data. One can hypothesize that enzymes present in many organisms might have been investigated in more detail, which could have led consequently to its annotation with multiple reactions. We assumed that the number of entries in Brenda for a given EC number provides a good estimate of the extent of the overall number of experimental studies performed on the enzyme. On



the other hand, the presence of the enzyme in multiple organisms can be quantified, as previously, by computing the average phylogenetic distance between all organisms annotated with this enzyme. In our tests, the comparison between these two measures revealed no significant correlation, suggesting that the number of experimental studies performed on enzymes is not directly related to their phylogenetic extent in the tree of life, as can be seen in supplemental Fig. S6, *A* and *B*. This result rules out a potential artifact in our observations due to investigation bias.

To test the hypothesis that those catalytic functions in enzymes with greater chemical diversity are to be found and shared across more distant species in the tree of life and therefore come from deeper ancestral enzymes in the tree of life, we measured the chemical diversity of an enzyme by using the promiscuity definitions previously introduced. That hypothesis is preferring an ACCTRAN optimization of homoplasies, favoring reversions over convergences; indeed, part of the tested hypothesis is that losing chemical diversity by progressive specialization is easier than having greater (twice or more) chemical diversity formed by ancestral enzymatic reactions. Conversely, enzymatic reactions that process specific substrates might be expected to be found usually confined within groups of more closely related species. Because catalytic specialization is an emergent property driven by evolution, based on our hypothesis, it should be observed at both enzyme and catalytic activity levels of promiscuity.

Similarly to the chemical distance between the substrates for a reaction pair, we measured the phylogenetic diversity of the reaction pair, which we define as the average phylogenetic distance between all of the organisms that can process that pair of reactions. In Fig. 2*A*, we plot the relationship between the reaction chemical diversity and reaction phylogenetic diversity of enzymes in our set. Interestingly, we found a significant high correlation. Therefore, this result suggests that the capability of an enzyme to process more than one substrate is intimately related to its evolvability. Reactions that are catalyzed by a highly specific enzyme are shared by poorly divergent species (*i.e.* specific taxa in the tree of life), suggesting a later emergence of this function during evolution. In contrast, promiscuous enzymes are more likely to appear uniformly distributed across species in the tree of life.

The conclusion does not change significantly according to the type of distance used to calculate species pairwise distances in the reaction phylogenetic diversity (Fig. 2, *A* and *B*); the correlation coefficient is $r = 0.86$, $p = 1.3 \times 10^{-3}$ when using patristic distances of a tree based on sequence data and $r = 0.83$, $p = 3.0 \times 10^{-3}$ when using taxonomic distances in the tree, showing that the first estimation of the correlation was not affected by too strong inequalities in DNA rate of change among lineages.

FIGURE 2. **Relationship between reaction chemical diversity and reaction phylogenetic diversity.** *A*, phylogenetic distances from pairwise genetic distances calculated from multiple alignments; *B*, phylogenetic distances calculated from pairwise taxonomic node count in the tree of life. *C*, relationship between latent enzyme promiscuity and phylogenetic diversity of catalytic functions (EC numbers).

Similar results were obtained in Fig. 2*C* at annotation level ($r = 0.81$, $p = 1.4 \times 10^{-2}$). In this case, we studied the relationship between all potential activities and phylogenetic distance of EC numbers, which was computed as the average distance between the species annotated with this particular EC number. This common trend reveals that not only are biochemical reactions found across more diverse species more likely to be processed by promiscuous enzymes, but also catalytic functions present in a greater diversity of species are potentially more promiscuous.

To evaluate the consistency of our previous results, which were based on KEGG, we recomputed them on MetaCyc, which has a higher curation level in terms of enzyme function annotation. For instance, although latent enzyme promiscuity values were found similar for both databases (correlation between promiscuity values $r = 0.88$, $p = 7.1 \times 10^{-4}$, shown in supplemental Fig. S7), there were nevertheless specific cases where EC numbers that appeared in KEGG as non-promiscuous were found in MetaCyc annotated as promiscuous. This result is due to the fact that KEGG annotations are performed, in their majority, by automated methods. Therefore, KEGG annotations might have overlooked some of the species-specific enzymatic functions of the multifunctional proteins present in the tree of life, a problem that could lead our study to underestimate the actual distribution of enzymatic functions across species in the tree. Thus, to evaluate any bias arising from the way KEGG is annotated, our previous tests were repeated on the MetaCyc database. The data set that we were able to process from MetaCyc was, in turn, considerably smaller than the one from KEGG. This difference in size is due to the fact that MetaCyc does not contain complete genomes as KEGG does but rather a small number of highly curated reference enzymes for each biochemical activity. The results nonetheless followed the same trend as before, despite the way enzyme promiscuity was measured either from the chemical diversity of their reactions ($r = 0.74$, $p = 1.5 \times 10^{-2}$; supplemental Fig. S8) or as latent promiscuity ($r = 0.94$, $p = 4.7 \times 10^{-4}$; supplemental Fig. S9). The fact that the results from the highly curated MetaCyc database closely mirror the results from KEGG provides some confirmation that the larger but much less accurate KEGG data set is providing meaningful results.

Chemical Diversity and Latent Enzyme Promiscuity across Species—Our previous findings show a close relationship between reaction chemical diversity and evolutionary time as measured through phylogenetic divergence. Based on this result, we might expect to see some distinctive shift in species with highly specialized catalytic activities. In fact, using our measure of reaction chemical diversity, it is possible to parallel successive divergence times of groups in the tree of life of species with metabolic annotations. In Fig. 3, *A* and *B*, reaction chemical diversity of enzymes was averaged and plotted for each organism and taxonomic group. Fig. 3, *C* and *D*, shows the results obtained by using latent enzyme promiscuity. Interestingly, taxonomic groups going from less to more specialized catalytic function appear in the following order archaea, bacteria, fungi, plants, and animals, which is the order of divergence times of each group (33), illustrating the fact that groups of later divergence in evolution are in general more specialized in their catalytic functions. We found that the separation between phylogenetic groups Archaea, Eubacteria, and Eukaryota, represented in Fig. 4, *A* and *B*, is significant in both latent enzyme promiscuity and reaction chemical diversity ($p < 2.2 \times 10^{-16}$ for a multivariate analysis of variance test).

Reaction chemical diversity and latent enzyme promiscuity together provide an account of how catalytic capabilities of organisms have evolved and adapted to their environments. For instance, euryarchaeota, which can both process many substrates and contain high latent promiscuity, are an example of extremophiles, which can survive under extreme conditions and therefore need a highly adaptable metabolic system. The same goes for deinococcus-thermus, which are a small group of bacteria highly resistant to environmental hazards. To generalize an evolutionary interpretation of the plot (Fig. 4, *A* and *B*), we must keep in mind that all species here are today's products of genetic lineages that do have the same evolutionary time since the origins of life. However, those species do not face the same chemical changes in their environments, and their position in the plot could also reflect past adaptive pressures. We infer that organisms showing high latent enzyme promiscuity and high reaction chemical diversity (Fig. 4, *A* and *B*, *top right*) must be pioneers that have to face strong chemical variations in their recently colonized environments. Organisms showing low latent enzyme promiscuity and high reaction chemical diversity (Fig. 4, *A* and *B*, *top left*) must be "old warriors" that have been accustomed to facing strong environmental changes since long ago. If the data set can be considered as "transfer-free," the spirochaetes and tenericutes are "old warriors" in the sense that their ancestors might have faced very strong environmental changes in a recent past and therefore they still kept a high reaction chemical diversity but do not have to face strong environmental pressure anymore. In other words, they are not in a pioneering situation anymore or are less so than in their recent past. Similarly, organisms exhibiting high latent enzyme promiscuity and low reaction chemical diversity (Fig. 4, *A* and *B*, *bottom right*) must be organisms that recently arrived in stable chemical conditions, like, for instance, recent parasites, commensals, or symbionts, such as those observed in actinobacteria, a bacteria with mutualistic tendencies (42). Last, organisms with low latent enzyme promiscuity and low reaction chemical diversity (Fig. 4, *A* and *B*, *bottom left*) must be highly integrated organisms. Pluricellularity, for instance, is buffering the genetic/enzymatic effects of environmental chemical changes. Moreover, multicellular mobile organisms have the choice to rapidly escape from chemical hazards and actively reach chemical optima. It is not surprising to find animals in that area of the plot. Plants are multicellular organisms; however, they cannot move and escape chemical aggression. This may be why we find them at the middle of the plot, with moderate latent enzyme promiscuity and moderate reaction chemical diversity.

Metabolic Functions of Promiscuous Enzymes—We did not observe a link between enzyme promiscuity and gene essentiality (3, 43) (*i.e.* with genes that are absolutely required for cell survival). However, analysis of metabolic functions found in promiscuous enzymes, given in Table 5, revealed that catalytic activities with a higher degree of latent promiscuity (≥5) are

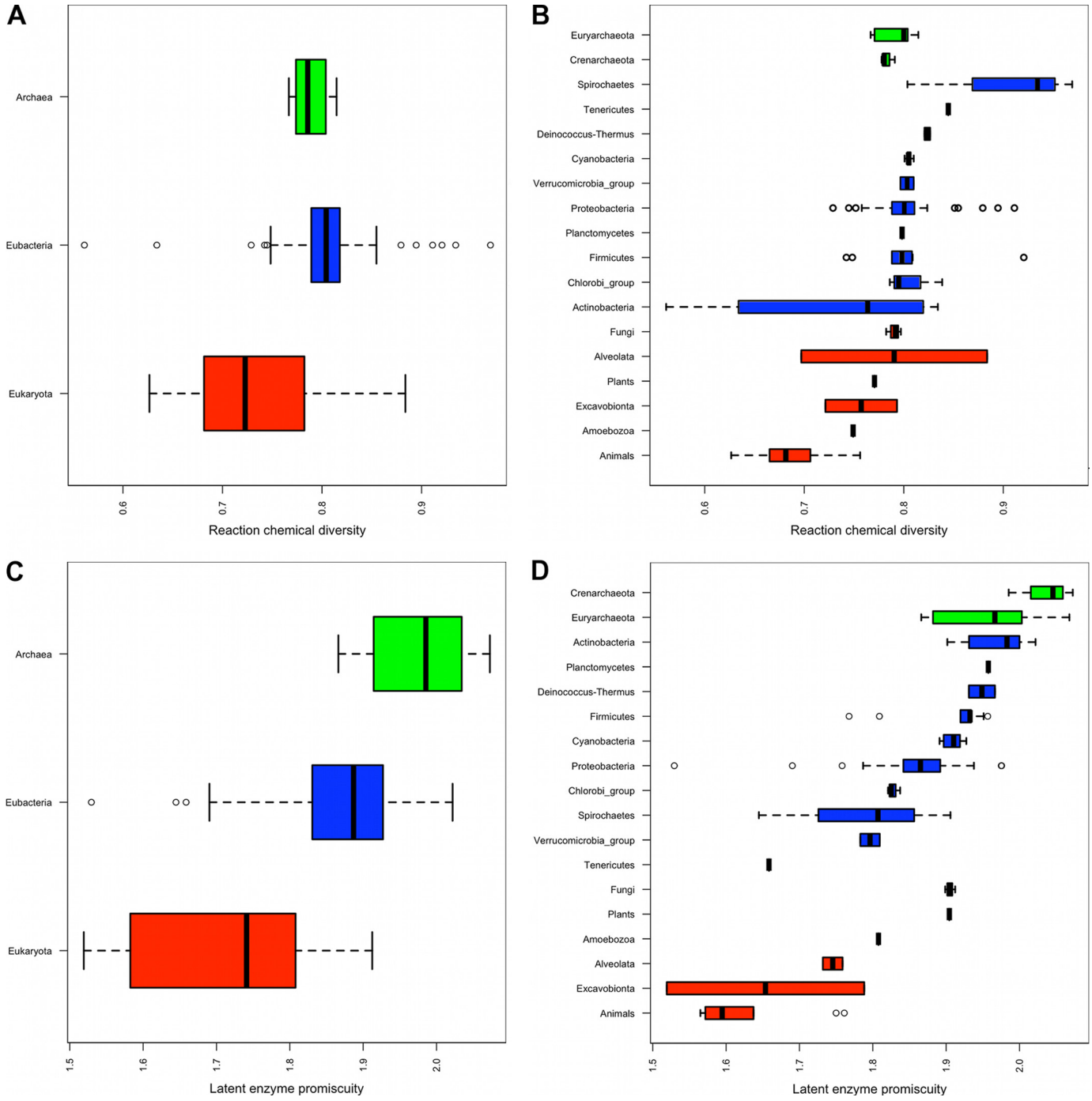


FIGURE 3. **Enzyme promiscuity at different levels of taxonomic grouping.** *A* and *B*, distribution of reaction chemical diversity; *C* and *D*, latent enzyme promiscuity at two different levels of taxonomic grouping.

mainly involved in amino acid ($p = 1.7 \times 10^{-2}$) and lipid metabolism ($p = 1.9 \times 10^{-2}$). This fact was seen both in the whole data set using the functional classification of pathways where they are involved (KEGG modules) (44) and using the NCBI clusters of ortholog groups for enzyme sequences in *E. coli*, shown in Fig. 5, *A* and *B*), respectively. Based on our previous finding of a greater phylogenetic extent of promiscuous enzymes, amino acid and lipid metabolisms might be associated here with the earliest form of biochemical reactions, which still are being processed by multifunctional enzymes (45, 46). It is not surprising; free amino acids are found in abiotic environments like interstellar meteorites and ice (47, 48). They must have been one of the very first complex organic compounds (*i.e.* with all main atomic species of life: carbon, hydrogen, oxygen, nitrogen, and sulfur) already available to the earliest protobionts and enzymatic functions). Enzymes performing amino acid catabolism and anabolism must have been the very first ones in life. Conversely, metabolisms of glycans ($p = 7.4 \times 10^{-3}$) and secondary metabolites ($p = 4.4 \times 10^{-1}$) are catalyzed by enzymes with a low degree of latent promiscuity, probably therefore of more recent origin. Their greater specificity might be linked to the fact that secondary

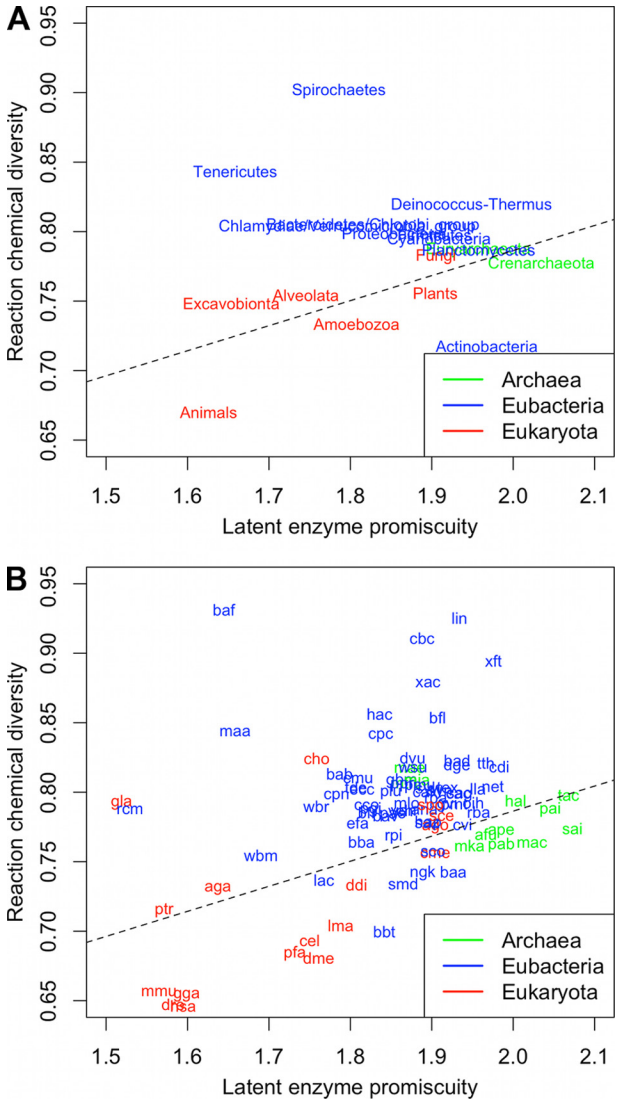


FIGURE 4. **Scatter plot of latent enzyme promiscuity and reaction chemical diversity in organisms.** *A*, grouping by taxonomic groups; *B*, grouping at the organism level.

TABLE 5

Enrichment (+) or depletion (−) in the metabolic function according to KEGG module classification for the set of enzymes with high latent promiscuity (≥5) compared with the total set in Fig. 5*A*

Significant p values ($p < 5.0 \times 10^{-2}$) are shown in boldface type.

Metabolism	p value	Enrichment/Depletion
Amino acids	$\mathbf{1.7 \times 10^{-2}}$	+
Lipids	$\mathbf{1.9 \times 10^{-2}}$	+
Central + energy	1.1×10^{-1}	−
Glycans	$\mathbf{7.4 \times 10^{-3}}$	−
Secondary	4.4×10^{-1}	−
Nucleotides	9.7×10^{-1}	−

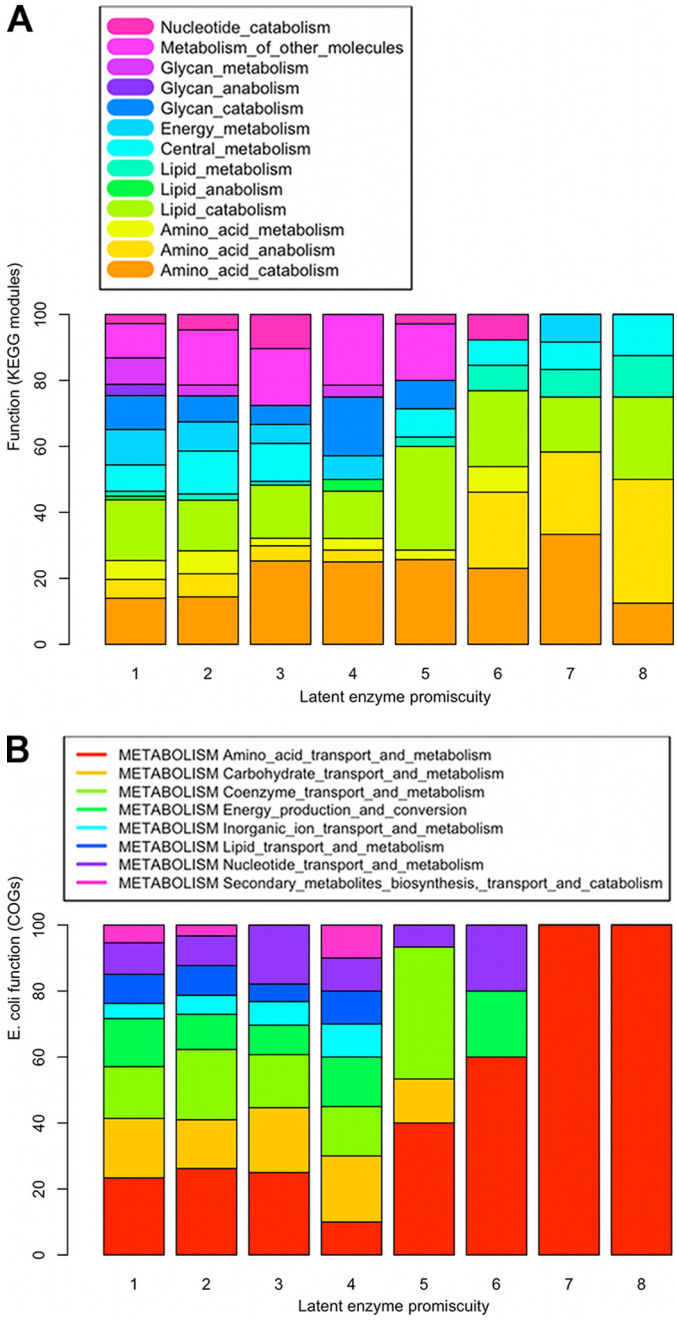


FIGURE 5. **Distribution of metabolic functions on enzymes depending on their latent promiscuity.** *A*, KEGG orthologies; *B*, *E. coli* based on clusters of ortholog groups.

metabolism is generally associated with defense mechanisms specific to organisms, and glycan pathways are usually associated with multicellular developmental processes and restricted to the metazoa (43).

Catalytic Efficiency of Promiscuous Enzymes—We used our definition of latent promiscuity in order to investigate the previously reported experimental observation that enhancement of latent catalytic activities does not seem often to trade off with parallel decreases in the original function (16, 24), in contradiction with the commonly accepted assumption that broad substrate acceptance generally comes at the price of low reaction turnover numbers (19). In our results, catalytic efficiency, estimated from experimental data as described under "Experimental Procedures," does not seem to be related in general to the latent promiscuity of enzymes because the trend that we observed remains flat for most of the values of latent enzyme promiscuity ($r = 0.02$, $p = 0.27$; see supplemental Fig. S10). This result, thus, corroborates the observed fact (16) that enzyme promiscuity can be in many cases achieved without compromising efficiency, suggesting that although there is a trend in evolution toward

specialization, catalytic efficiency has remained on average at the same level during the history of life, perhaps due to energy constraints.

DISCUSSION

In this work, we have shown that enzyme promiscuity plays a prominent role in the study of the evolution of metabolic networks. A distinctive property of our given definition of promiscuity, which is based on the chemical diversity of catalytic reactions that enzymes can process, is that it does not depend on protein sequence comparisons. As a matter of fact, the same evolutionary trend in enzyme promiscuity was observed in the tree of life independently of the way pairwise distances between species were measured, either based on sequence data or on pure taxonomic hierarchy. We find that the fact that a particular enzyme catalyzes a chemically diverse set of substrates is related to the wider spread of the underlying processing reactions across the tree of life (Fig. 2, *A* and *B*), suggesting an older origin for these nonspecific enzymes. In turn, enzymes that process more specific substrates are usually shared by organisms of lower divergence times (*i.e.* constrained to some smaller taxa in the tree of life, which might be interpreted as a later emergence of their function during evolution). These observations are confirmed when we look at catalytic functions that potentially can be performed simultaneously, a property that we term latent enzyme promiscuity. Again, the fact that a catalytic function is potentially more promiscuous is related to the fact that the activities are more widespread across the tree of life (Fig. 2*C*).

The evolutionary aspect of enzyme promiscuity appears to be a feature with broad applicability. In particular, it was shown in this study that promiscuity at both the enzyme and catalytic level paralleled divergence times in the tree of life. These results can provide new clues to explain the mechanism of how enzymatic activities evolve and specialize under environmental pressure. Namely, because promiscuous catalysis is mainly found in metabolic functions associated with amino acid metabolism, our hypothesis suggests an early origin of these functions, whereas glycan and secondary metabolisms appear as highly specialized functions of later origin.

We observe, however, that catalytic efficiency of enzymes is a property essentially independent of their latent promiscuity. This intriguing result might suggest that despite the general trend in evolution toward specialization, energy constraints have kept average catalytic efficiency at the same level through the history of life. Open questions remain in the study of how metabolic fluxes are related to enzyme promiscuity and whether these results can be extended to other biological networks, such as transcription regulatory and signaling networks. A better understanding of the underlying model of enzyme evolution will have important implications for protein design through directed and *in silico* evolution and might assist us with the identification of primeval promiscuous catalytic activities in metabolic pathways contributing to life's minimal metabolism.

REFERENCES

1. Peregrin-Alvarez, J. M., Tsoka, S., and Ouzounis, C. A. (2003) *Genome Res.* **13,** 422–427
2. Peregrín-Alvarez, J. M., Sanford, C., and Parkinson, J. (2009) *Genome Biol.* **10,** R63
3. Gerdes, S., Edwards, R., Kubal, M., Fonstein, M., Stevens, R., and Osterman, A. (2006) *Curr. Opin. Biotechnol.* **17,** 448–456
4. Danchin, A. (2009) *Curr. Opin. Biotechnol.* **20,** 504–508
5. Alves, R., Chaleil, R. A., and Sternberg, M. J. (2002) *J. Mol. Biol.* **320,** 751–770
6. O'Brien, P. J., and Herschlag, D. (1999) *Chem. Biol.* **6,** R91–R105
7. Rison, S. C., and Thornton, J. M. (2002) *Curr. Opin. Struct. Biol.* **12,** 374–382
8. Rison, S. C., Teichmann, S. A., and Thornton, J. M. (2002) *J. Mol. Biol.* **318,** 911–932
9. Yamada, T., and Bork, P. (2009) *Nat. Rev. Mol. Cell Biol.* **10,** 791–803
10. Freilich, S., Goldovsky, L., Ouzounis, C. A., and Thornton, J. M. (2008) *BMC Evol. Biol.* **8,** 247
11. de Groot, M. J., van Berlo, R. J., van Winden, W. A., Verheijen, P. J., Reinders, M. J., and de Ridder, D. (2009) *Bioinformatics* **25,** 2975–2982
12. Marcet-Houben, M., Puigbò, P., Romeu, A., and Garcia-Vallve, S. (2007) *Bioinformation* **2,** 135–144
13. Horowitz, N. H. (1945) *Proc. Natl. Acad. Sci. U.S.A.* **31,** 153–157
14. Jensen, R. A. (1976) *Annu. Rev. Microbiol.* **30,** 409–425
15. Light, S., and Kraulis, P. (2004) *BMC Bioinformatics* **5,** 15+
16. Khersonsky, O., Roodveldt, C., and Tawfik, D. S. (2006) *Curr. Opin Chem. Biol.* **10,** 498–508
17. Hult, K., and Berglund, P. (2007) *Trends Biotechnol.* **25,** 231–238
18. Khersonsky, O., and Tawfik, D. S. (2010) in *Comprehensive Natural Products II: Chemistry and Biology* (Mander, L., and Lui, H. W., eds) pp. 48–90, Elsevier, Oxford
19. Jakoby, W. B., and Ziegler, D. M. (1990) *J. Biol. Chem.* **265,** 20715–20718
20. Bloom, J. D., Romero, P. A., Lu, Z., and Arnold, F. H. (2007) *Biol. Direct* **2,** 17
21. Nath, A., and Atkins, W. M. (2008) *Biochemistry* **47,** 157–166
22. Vitkup, D., Kharchenko, P., and Wagner, A. (2006) *Genome Biol.* **7,** R39
23. Pál, C., Papp, B., and Hurst, L. D. (2001) *Genetics* **158,** 927–931
24. Kurtovic, S., Shokeer, A., and Mannervik, B. (2008) *J. Mol. Biol.* **382,** 136–153
25. Copley, S. D. (2003) *Curr. Opin Chem. Biol.* **7,** 265–272
26. Kazlauskas, R. J. (2005) *Curr. Opin. Chem. Biol.* **9,** 195–201
27. Faulon, J. L., Collins, M. J., and Carr, R. D. (2004) *J. Chem. Inf. Comput. Sci.* **44,** 427–436
28. Carbonell, P., and Faulon, J. L. (2010) *Bioinformatics* **26,** 2012–2019
29. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008) *Nucleic Acids Res.* **36,** D480–D484
30. Caspi, R., Altman, T., Dale, J. M., Dreher, K., Fulcher, C. A., Gilham, F., Kaipa, P., Karthikeyan, A. S., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Paley, S., Popescu, L., Pujar, A., Shearer, A. G., Zhang, P., and Karp, P. D. (2010) *Nucleic Acids Res.* **38,** D473–D479
31. Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006) *Science* **311,** 1283–1287
32. Guindon, S., and Gascuel, O. (2003) *Syst. Biol.* **52,** 696–704
33. Hedges, S. B., and Kumar, S. (2009) *The Timetree of Life*, pp. 3–18, Oxford University Press, New York
34. Warwick, R. M., and Clarke, K. R. (1995) *Mar. Ecol. Prog. Ser.* **129,** 301–305
35. Chang, A., Scheer, M., Grote, A., Schomburg, I., and Schomburg, D. (2009) *Nucleic Acids Res.* **37,** D588–D592
36. Markov, G., Lecointre, G., Demeneix, B., and Laudet, V. (2008) *BioEssays* **30,** 349–357
37. Rothman, S. C., and Kirsch, J. F. (2003) *J. Mol. Biol.* **327,** 593–608
38. Lecointre, G., and Le Guyader, H. (2007) *The Tree of Life: A Phylogenetic Classification*, pp. 18–24, Belknap Press of Harvard University Press, Cambridge, MA
39. Doolittle, W. F., and Bapteste, E. (2007) *Proc. Natl. Acad. Sci. U.S.A.* **104,**

2043–2049
40. Lopez, P., and Bapteste, E. (2009) *C. R. Biol.* **332,** 171–182
41. Lecointre, G., Philippe, H., Vân Lê, H. L., and Le Guyader, H. (1993) *Mol. Phylogenet. Evol.* **2,** 205–224
42. Kaltenpoth, M. (2009) *Trends Microbiol.* **17,** 529–535
43. Baba, T., Ara, T., Hasegawa, M., Takai, Y., and Okumura, Y. (2006) *Mol. Syst. Biol.* **2,** 2006.0008
44. Yamada, T., Kanehisa, M., and Goto, S. (2006) *BMC Bioinformatics* **7,** 130
45. Cunchillos, C., and Lecointre, G. (2003) *J. Biol. Chem.* **278,** 47960–47970
46. Meierhenrich, U. J., Muñoz Caro, G. M., Bredehöft, J. H., Jessberger, E. K., and Thiemann, W. H. (2004) *Proc. Natl. Acad. Sci. U.S.A.* **101,** 9182–9186
47. Muñoz Caro, G. M., Meierhenrich, U. J., Schutte, W. A., Barbier, B., Arcones Segovia, A., Rosenbauer, H., Thiemann, W. H., Brack, A., and Greenberg, J. M. (2002) *Nature* **416,** 403–406
48. Bernstein, M. P., Dworkin, J. P., Sandford, S. A., Cooper, G. W., and Allamandola, L. J. (2002) *Nature* **416,** 401–403