Evolution of Domain Architectures and Catalytic Functions of Enzymes in Metabolic Systems

Summit Suen[1], Henry Horng-Shing Lu[2], and Chen-Hsiang Yeang[1,*]

[1]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

[2]Institute of Statistics, National Chiao-Tung University, Hsinchu, Taiwan

*Corresponding author: E-mail: chyeang@stat.sinica.edu.tw.

Accepted: August 20, 2012

Abstract

Domain architectures and catalytic functions of enzymes constitute the centerpieces of a metabolic network. These types of information are formulated as a two-layered network consisting of domains, proteins, and reactions—a domain–protein–reaction (DPR) network. We propose an algorithm to reconstruct the evolutionary history of DPR networks across multiple species and categorize the mechanisms of metabolic systems evolution in terms of network changes. The reconstructed history reveals distinct patterns of evolutionary mechanisms between prokaryotic and eukaryotic networks. Although the evolutionary mechanisms in early ancestors of prokaryotes and eukaryotes are quite similar, more novel and duplicated domain compositions with identical catalytic functions arise along the eukaryotic lineage. In contrast, prokaryotic enzymes become more versatile by catalyzing multiple reactions with similar chemical operations. Moreover, different metabolic pathways are enriched with distinct network evolution mechanisms. For instance, although the pathways of steroid biosynthesis, protein kinases, and glycosaminoglycan biosynthesis all constitute prominent features of animal-specific physiology, their evolution of domain architectures and catalytic functions follows distinct patterns. Steroid biosynthesis is enriched with reaction creations but retains a relatively conserved repertoire of domain compositions and proteins. Protein kinases retain conserved reactions but possess many novel domains and proteins. In contrast, glycosaminoglycan biosynthesis has high rates of reaction/protein creations and domain recruitments. Finally, we elicit and validate two general principles underlying the evolution of DPR networks: 1) duplicated enzyme proteins possess similar catalytic functions and 2) the majority of novel domains arise to catalyze novel reactions. These results shed new lights on the evolution of metabolic systems.

Key words: protein domain architecture, metabolic reaction, parsimony, max-product, protein duplication.

Introduction

Metabolic systems are one of the most ancient and essential systems of living organisms. Their importance to life, universality, and complexity are unequivocal. On the one hand, the metabolism of many essential nutrients is highly conserved across all life forms on earth. On the other hand, diverse systems have been tailored to meet the differential metabolic demands in organisms living in distinct environments. These phenomena have amazed generations of scientists and inspired a great number of research endeavors in modern biology.

The term "metabolic network" is overloaded as it encompasses at least three types of relations. First, one can focus on the metabolites (substrates) of reactions and construct a "metabolite-centric" network. Second, alternatively one can focus on the catalytic functions of enzymes and construct an "enzyme-centric" network. Third, most enzyme proteins are composed of polypeptide subunits called domains. Each domain possesses a distinct structural and functional characteristic, and novel proteins can be formed by recombinations of limited domains. Thus, one can focus on the domain architectures of enzymes and construct a "domain-centric" network. In cellular organisms, the evolution of metabolic systems is driven by the evolution of enzyme proteins. Furthermore, the phylogenetic relations of enzymes are revealed by their protein sequences. To study the evolution of metabolic systems, we construct a network of domains, proteins, and reactions to simultaneously characterize the domain architectures and catalytic functions of enzyme proteins. We term this network a "domain-protein-reaction network."

The evolution of domain architectures and catalytic functions of proteins in metabolic systems has been investigated in many previous studies. By comparing the domain architectures from a wide range of species, generic patterns of domain evolution have been uncovered. For instance, Chothia et al. (2003) showed the majority of protein domains appeared before the prokaryote–eukaryote split. Wuchty (2001) demonstrated a power-law distribution on the co-occurrence of protein domains. Apic et al. (2001) observed prevalent domain recombinations across all three kingdoms of life. Teichmann et al. (2001) indicated frequent turnover of substrate binding domains and stability of catalytic domains in enzymes. Vogel et al. (2005) demonstrated the conservation of domain architectures in specific orders. Schmidt et al. (2007) reviewed the molecular mechanisms originating novel domains, and Kaessmann et al. (2002) demonstrated signatures of domain shuffling in human genomes. Behzadi and Vingron (2006) proposed an optimization algorithm to reconstruct the domain architectures of ancestral proteins from those of the extant species. Wiedenhoeft et al. (2011) proposed an algorithm to construct a novel network-like structure representing domain architectures. Bjorklund et al. (2005) calculated the differences of domain architectures between homologous proteins and observed that the evolution of most multidomain proteins could be explained by stepwise insertions of single domains and tandem duplications of domains. Recurrence of domain architectures in evolution was investigated by multiple authors (e.g., Gough 2005; Przytycka et al. 2006; Forslund et al. 2007), whereas disparate conclusions were reached from those studies. Primary mechanisms of altering the repertoire of domain architectures—including protein duplications (Ohno 1970), domain recruitments (Teichmann et al. 2001), gene loss (Kunin and Ouzounis 2003), and horizontal gene transfers (HGTs; Pal et al. 2005)—have also been investigated and reviewed in (Bornberg-Bauer et al. 2005).

Tracing the history of enzymes belonging to specific functional classes or pathways also leads to important discoveries. For instance, Morowitz (1999) and Peregrin-Alves et al. (2003) identified a conserved central core of metabolic reactions shared by all cellular organisms. This core set consists of pathways involved in metabolism of carbohydrates, nucleotides, and amino acids. Freilich et al. (2005) found that eukaryotes expanded metabolic networks by increasing functional redundancy, whereas prokaryotes' evolution was dominated by broadening the reaction repertoire. In particular, mammals revealed a massive expansion of enzymes involved in signaling and degradation. Pfeiffer et al. (2005) observed from simulation studies that specialized metabolic systems could be derived from a small multifunctional subsystem. Borenstein et al. (2008) identified a metabolic network's "seed set," the set of compounds that are exogenously acquired, and inferred the environmental conditions of organisms from their seed sets. Freilich et al. (2008) classified metabolic pathways by their emerging times along the lineage from the common ancestor of cellular organisms to human. Mithani et al. proposed a stochastic model for metabolic network evolution by treating metabolites as nodes and reactions as hyperedges in hypergraphs. The rates of adding or deleting the reaction hyperedges could be independent, neighbor-dependent (Mithani et al. 2009), or the hybrid of the two models (Mithani et al. 2010).

Simultaneous characterization of domain architectures and catalytic functions of enzymes can shed more light on the evolution of metabolic systems. Two remarkable examples are validation of the patchwork model for metabolic system evolution and reconstruction of the phylogenetic relations of metabolic pathways. By examining the pathways containing homologs of newly recruited domains, it is discovered that a novel metabolic pathway tends to recruit domains from diverse existing pathways rather than inheriting from a single source (Schmidt et al. 2003). Furthermore, by comparing the functional annotations of enzymes containing each domain family, Caetano-Anolles et al. attempted to reconstruct the phylogenetic relations of metabolic pathways (Caetano-Anolles and Caetano-Anolles 2003; Caetano-Anolles et al. 2009). In addition, previously we examined process-specific evolutionary patterns of domain compositions and metabolic reactions between human and *Escherichia coli* (Yeang and Baas 2009).

Despite the values of these studies, a principled approach to reconstruct the history of metabolic network evolution and a comprehensive categorization of its underlying mechanisms are yet to be established. Given the domain architectures and catalytic functions of enzymes—domain–protein–reaction (DPR) networks—in a number of extant species, our goals are to 1) reconstruct the evolutionary history of these networks, 2) categorize the mechanisms of metabolic system evolution in terms of network operations, and 3) detect the enriched types of evolutionary mechanisms for each pathway and relate the enriched evolutionary patterns with the functions of pathways. To fulfill these goals, we propose an algorithm to reconstruct the DPR networks of ancestral species from observed data that minimize the total number of network alterations along all lineages in the phylogeny. We confirm the accuracy of the reconstruction algorithm by simulation studies and cross validations on the real data sets. By applying the inference algorithm to the DPR networks of 13 selected species, we find that prokaryotes and eukaryotes share dominant evolutionary mechanisms in the early stage but diverge substantially along each clade. Refined analysis indicates heterogeneous patterns of evolutionary mechanisms for distinct metabolic pathways. For instance, although the pathways of steroid biosynthesis, protein kinases, and glycosaminoglycan biosynthesis all contribute critically to opisthokonta-specific physiology, their evolution of domain architectures and catalytic functions follows distinct patterns. Steroid biosynthesis is enriched with reaction creations but retains a relatively conserved repertoire of proteins and domain architectures. Protein kinases retain conserved reactions but possess many novel domains and proteins.

In contrast, glycosaminoglycan biosynthesis has high rates of reaction/protein creations and domain recruitments. Finally, we elicit and validate two general principles underlying the evolution of DPR networks: 1) duplicated enzyme proteins possess similar catalytic functions and 2) the majority of novel domains arise to catalyze novel reactions. The results shed new lights on the evolution of metabolic networks.

Materials and Methods

Evolutionary History of DPR Networks

We define a DPR network as a two-layered graph $G = (V_D \cup V_P \cup V_R, E_{DP} \cup E_{PR})$ consisting of three types of nodes—domain families V_D, proteins V_P, and reactions V_R—and two types of edges—domain-protein edges E_{DP} and protein-reaction edges E_{PR}. For conciseness, we will use the term "domains" to denote both domain families and members of the family that appear in specific proteins. A domain-protein pair $(d, p) \in E_{DP}$ is adjacent in G if domain d appears in protein p. A protein-reaction pair $(p, r) \in E_{PR}$ is adjacent in G if protein p catalyzes reaction r. Therefore, G simultaneously characterizes domain architectures and catalytic functions of enzymes in a metabolic system. Notice that a DPR network may include nonmetabolic enzyme proteins such as protein kinases and protein glycosylation enzymes.

Given a collection of extant species S and their phylogenetic tree T_S, each species (extant or ancestral) in T_S possesses a DPR network. The DPR network of an extant species can be extracted from the data of domain architectures and enzymatic functions. However, the DPR networks of ancestral species and their inheritance relations to the extant DPR networks cannot be observed. An objective of this work is to reconstruct the evolutionary history of the DPR networks in multiple species.

We define an evolutionary history over a phylogeny T_S as a tuple $H = \{G, \mathrm{Pa}\}$. $G \equiv \{s \in T_S \mid G_s\}$ denotes the collection of the DPR networks for all species in T_S, where G_s is the DPR network of species s. $\mathrm{Pa} \equiv (\bigcup_{s \in T_S} V_P(G_s)) \times (\bigcup_{s \in T_S} V_P(G_s) \cup \phi)$ maps each protein p to its parent $\mathrm{Pa}(p)$. $\mathrm{Pa}(p) = \phi$ if p is newly created and has no parent. A toy example of the evolutionary history of DPR networks is illustrated in the left part of figure 1. A reconstruction algorithm takes species phylogeny T_S and extant DPR networks $\{s \in S \mid G_s\}$ as inputs and returns an evolutionary history H. The following missing information has to be imputed: 1) domains, proteins, and reactions of each ancestral species, 2) domain-protein and protein-reaction edges in the ancestral species, and 3) the parents of all proteins in each species.

Similar to other reconstruction algorithms (e.g., Hein 1990; Fong et al. 2007; Pinney et al. 2007; Ma et al. 2008), we employ a parsimonious assumption to infer the evolutionary history of DPR networks. Parsimony demands the number of changes over an evolutionary history to be minimized.

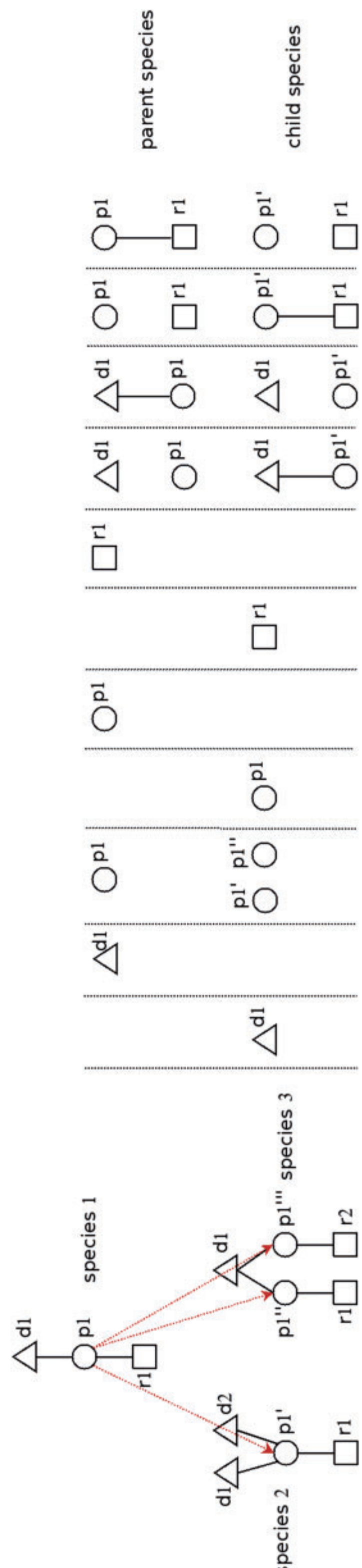

FIG. 1.—Left: A toy example of the evolutionary history of DPR networks. Species 1 has two child species: species 2 and 3. Triangles denote domains, circles denote proteins, and squares denote reactions. Solid, undirected edges denote domain-protein and protein-reaction edges. Dotted, directed edges denote phylogenetic relations of proteins. In species 1, protein p_1 consists of one domain d_1 and catalyzes reaction r_1. From species 1 to species 2, a new domain d_2 is recruited into protein p_1', which is inherited from p_1. From species 1 to species 3, p_1 is duplicated into p_1'' and p_1''', and p_1''' catalyzes a new reaction r_2. Right: Eleven types of changes in the DPR networks from a parent species to a child species. From left to right: (1) domain creation, (2) domain deletion, (3) protein duplication, (4) protein creation, (5) protein deletion, (6) reaction creation, (7) reaction deletion, (8) domain-protein edge addition, (9) domain-protein edge deletion, (10) protein-reaction edge addition, and (11) protein-reaction edge deletion.

There are 11 types of changes between two DPR networks belonging to a parent–child species pair. They are listed as follows and illustrated in the right part of figure 1.

1. Domain creation: A domain node appears in the child species but not in the parent species.
2. Domain deletion: A domain node appears in the parent species but not in the child species.
3. Protein duplication: A protein in the parent species has multiple children in the child species.
4. Protein creation: A protein in the child species has no parent.
5. Protein deletion: A protein in the parent species has no child in the child species.
6. Reaction creation: A reaction node appears in the child species but not in the parent species.
7. Reaction deletion: A reaction node appears in the parent species but not in the child species.
8. Domain-protein edge addition: A domain-protein edge is absent in the parent species and present in the child species.
9. Domain-protein edge deletion: A domain-protein edge is present in the parent species and absent in the child species.
10. Protein-reaction edge addition: Similar to domain-protein edge addition.
11. Protein-reaction edge deletion: Similar to domain-protein edge deletion.

Casting Parsimonious Reconstruction of the Evolutionary History of DPR Networks as Statistical Inference

A parsimonious solution minimizes the total number of network changes between each parent–child species pair in the phylogeny. For each species, we define discrete variables pertaining to the following features of the evolutionary history H: the parent label of each protein, binary variables indicating whether each protein is valid, whether each protein is newly created, whether each domain-protein or protein-reaction edge is present or absent. Formally, for each species $s \in T_S$ we define the following variables:

— $l^s_{p_j}$: distinct label of protein p_j. In this work, we fixed $l^s_{p_j} = j$.
— $a^s_{p_j}$: parent label of protein p_j. $a^s_{p_j} = 0$ if p_j has no parent.
— $\pi^s_{d_i,p_j}$: a binary variable indicating the presence of a domain-protein edge (d_i, p_j).
— $\pi^s_{p_j,r_k}$: a binary variable indicating the presence of a protein-reaction edge (p_j, r_k).
— $v^s_{p_j}$: a binary variable indicating that p_j is newly created in s.
— $\mu^s_{p_j}$: a binary variable indicating that p_j is a valid protein in s.

An instantiation of values of all these variables specifies an evolutionary history of DPR networks over a species tree T. The cost function of an evolutionay history is the total number of network changes (node creations/duplications/deletions and edge additions/deletions) summed over all consecutive species pairs in phylogeny T_S:

$$
\begin{aligned}
C = \sum_{(s_1,s_2)\in T} \Bigg[& \sum_{j_1 \in s_1} \left| \sum_{j_2 \in s_2} \bar{\delta}\left(l^{s_1}_{p_{j_1}}, a^{s_2}_{p_{j_2}}\right)\mu^{s_2}_{p_{j_2}} - 1 \right| \mu^{s_1}_{p_{j_1}} \\
& + \sum_{j_2 \in s_2} v^{s_2}_{p_{j_2}} \mu^{s_2}_{p_{j_2}} + \sum_{j_2 \in s_2} v^{s_2}_{p_{j_2}} \mu^{s_2}_{p_{j_2}} \left(\sum_i \pi^{s_2}_{d_i,p_{j_2}} + \sum_k \pi^{s_2}_{p_{j_2},r_k} \right) \\
& + \sum_{i,j_1,j_2} \bar{\delta}\left(l^{s_1}_{p_{j_1}}, a^{s_2}_{p_{j_2}}\right) \delta\left(\pi^{s_1}_{d_i,p_{j_1}}, \pi^{s_2}_{d_i,p_{j_2}}\right) \mu^{s_1}_{p_{j_1}} \mu^{s_2}_{p_{j_2}} \\
& + \sum_{j_1,j_2,k} \bar{\delta}\left(l^{s_1}_{p_{j_1}}, a^{s_2}_{p_{j_2}}\right) \delta\left(\pi^{s_1}_{p_{j_1},r_k}, \pi^{s_2}_{p_{j_2},r_k}\right) \mu^{s_1}_{p_{j_1}} \mu^{s_2}_{p_{j_2}} \Bigg].
\end{aligned}
\tag{1}
$$

where $\bar{\delta}(a,b) = 1$ if $a = b$ and 0 otherwise, and $\delta(a,b) = 1 - \bar{\delta}(a,b)$. The six terms in equation (1) correspond to the costs of protein duplications/deletions (term 1), protein creations (term 2), domain-protein and protein-reaction edge additions to new proteins (terms 3 and 4), and domain-protein and protein-reaction edge additions or deletions on inherited proteins (terms 5 and 6).

Three additional constraints are introduced to specify the relations of variables:

$$
\begin{aligned}
& v^{s_2}_{p_{j_2}} = \prod_{j_1 \in s_1} \left[\delta\left(l^{s_1}_{p_{j_1}}, a^{s_2}_{p_{j_2}}\right)\mu^{s_1}_{p_{j_1}} + 1 - \mu^{s_1}_{p_{j_1}} \right]. \\
& \text{if } \mu^{s_1}_{p_j} = 0 \text{ then none of the parent labels is } l_j. \\
& \text{if } v^{s_2}_{p_j} = 1 \text{ then } a^{s_2}_{p_j} = 0.
\end{aligned}
\tag{2}
$$

The first constraint stipulates that an inherited protein in s_2 has a valid parent protein in s_1. The second constraint stipulates that only valid proteins can be parents. The third constraint states a tautology that a newly created protein has no parent. A parsimonious evolutionary history minimizes the cost function in equation (1) subjected to the constraints in equation (2).

Constrained optimization of a complex cost function of integer variables is generally NP hard. Therefore, we translate the cost function and constraints into a joint likelihood function of a probabilistic graphical model and apply a belief propagation algorithm to find a (approximate) maximum-likelihood solution of the model. This approach has the merits of simplicity—belief propagation essentially performs summations/maximizations and multiplications—and flexibility—arbitrary objective functions and constraints can be incorporated in the model. In practice, it has been successfully applied to large-scale problems of deciphering error-correction codes (Kschischang et al. 2001), inferring the functions of protein–DNA and protein–protein interactions (Yeang et al. 2004), and determining the causal orders of genes in the regulatory network (Vaske et al. 2009).

We convert the combinatoiral optimization problem into a statistical inference problem by building a multiplicative likelihood function from the exponentiated cost and constraints.

$$
\begin{aligned}
L &= \prod_{(s_1,s_2)\in T}\prod_{i,j_1,j_2,k}\prod_{m=1}^{10} f_m^{(s_1,s_2)}.\\
f_1^{(s_1,s_2)}\left(l_{p_{j_1}}^{s_1}, a_{p_{j_2}}^{s_2}, \mu_{p_{j_2}}^{s_2}, \gamma_{j_1,j_2}\right) &= \begin{cases} 1 & if\, \gamma_{j_1,j_2} = \bar{\delta}\left(l_{p_{j_1}}^{s_1}, a_{p_{j_2}}^{s_2}\right)\mu_{p_{j_2}}^{s_2}. \\ 0 & \text{otherwise.}\end{cases}\\
f_2^{(s_1,s_2)}\left(\gamma_{j_1,1},\cdots,\gamma_{j_1,M},\mu_{p_{j_1}}^{s_1}\right) &= e^{-\left|\left(\sum_{j_2}\gamma_{j_1,j_2}\right)-1\right|\mu_{p_{j_1}}^{s_1}}.\\
f_3^{(s_1,s_2)}\left(v_{p_{j_2}}^{s_2}, \mu_{p_{j_2}}^{s_2}\right) &= e^{-v_{p_{j_2}}^{s_2}\mu_{p_{j_2}}^{s_2}}.\\
f_4^{(s_1,s_2)}\left(v_{p_{j_2}}^{s_2}, \pi_{d_i,p_{j_2}}^{s_2}, \mu_{p_{j_2}}^{s_2}\right) &= e^{-v_{p_{j_2}}^{s_2}\pi_{d_i,p_{j_2}}^{s_2}\mu_{p_{j_2}}^{s_2}}.\\
f_5^{(s_1,s_2)}\left(v_{p_{j_2}}^{s_2}, \pi_{p_{j_2},r_k}^{s_2}, \mu_{p_{j_2}}^{s_2}\right) &= e^{-v_{p_{j_2}}^{s_2}\pi_{p_{j_2},r_k}^{s_2}\mu_{p_{j_2}}^{s_2}}.\\
f_6^{(s_1,s_2)}\left(l_{p_{j_1}}^{s_1}, a_{p_{j_2}}^{s_2}, \pi_{d_i,p_{j_1}}^{s_1}, \pi_{d_i,p_{j_2}}^{s_2}, \mu_{p_{j_1}}^{s_1}, \mu_{p_{j_2}}^{s_2}\right) &= e^{-\bar{\delta}\left(l_{p_{j_1}}^{s_1}, l_{p_{j_2}}^{s_2}\right)\delta\left(\pi_{d_i,p_{j_1}}^{s_1}, \pi_{d_i,p_{j_2}}^{s_2}\right)\mu_{p_{j_1}}^{s_1}\mu_{p_{j_2}}^{s_2}}.\\
f_7^{(s_1,s_2)}\left(l_{p_{j_1}}^{s_1}, a_{p_{j_2}}^{s_2}, \pi_{p_{j_1},r_k}^{s_1}, \pi_{p_{j_2},r_k}^{s_2}, \mu_{p_{j_1}}^{s_1}, \mu_{p_{j_2}}^{s_2}\right) &= e^{-\bar{\delta}\left(l_{p_{j_1}}^{s_1}, l_{p_{j_2}}^{s_2}\right)\delta\left(\pi_{p_{j_1},r_k}^{s_1}, \pi_{p_{j_2},r_k}^{s_2}\right)\mu_{p_{j_1}}^{s_1}\mu_{p_{j_2}}^{s_2}}.\\
f_8^{(s_1,s_2)}\left(v_{p_{j_2}}^{s_2}, a_{p_{j_2}}^{s_2}, l_{p_1}^{s_1},\cdots,l_{p_M}^{s_1},\mu_{p_1}^{s_1},\cdots,\mu_{p_M}^{s_1}\right) &= \begin{cases} 1 & \text{if } v_{p_{j_2}}^{s_2} = \prod_{j_1\in s_1}\left[\delta\left(l_{p_{j_1}}^{s_1}, a_{p_{j_2}}^{s_2}\right)\mu_{p_{j_1}}^{s_1} + 1 - \mu_{p_{j_1}}^{s_1}\right]. \\ 0 & \text{other wise.}\end{cases}\\
f_9^{(s_1,s_2)}\left(l_{p_j}^{s_1}, \mu_{p_j}^{s_1}, a_{p_1}^{s_2},\cdots,a_{p_M}^{s_2}\right) &= \begin{cases} 1 & \text{if } \left(\mu_{p_j}^{s_1}=1\right)\vee\left[\left(\mu_{p_j}^{s_1}=0\right)\wedge\left(l_{p_j}^{s_1}\neq a_{p_1}^{s_2}\right)\wedge\cdots\wedge\left(l_{p_j}^{s_1}\neq a_{p_M}^{s_2}\right)\right]. \\ 0 & \text{other wise.}\end{cases}\\
f_{10}^{(s_1,s_2)}\left(v_{p_j}^{s_2}, a_{p_j}^{s_2}\right) &= \begin{cases} 1 & \text{if } \left(a_{p_j}^{s_2}=0\right)\wedge\left(v_{p_j}^{s_2}=1\right). \\ 1 & \text{if } \left(a_{p_j}^{s_2}>0\right)\wedge\left(v_{p_j}^{s_2}=0\right). \\ 0 & \text{other wise.}\end{cases}
\end{aligned}
\tag{3}
$$

Multiplication is taken over the indices of adjacent species pairs (s_1, s_2) in the phylogeny, domains (i), parent–child protein pairs (j_1 and j_2), reactions (k), and factor types (m). Factors $f_1^{(s_1,s_2)} - f_7^{(s_1,s_2)}$ correspond to the terms in the cost function in equation (1), and factors $f_8^{(s_1,s_2)} - f_{10}^{(s_1,s_2)}$ correspond to the constraints in equation (2).

Equation (3) is the joint likelihood function of a factor graph model (Kschischang et al. 2001). The model can be visualized as a bipartite graph. Nodes of the graph correspond to variables and factors in the likelihood function. An edge (x, f) between a variable x and a factor f denotes that x is an argument of f.

The max-product algorithm is employed to find an approximate solution of maximum likelihood variable configurations of equation (3) (Kschischang et al. 2001). In brief, message functions are defined over edges in the factor graph. In each iteration, messages are updated according to the messages incident from neighbors. A variable $\rightarrow$ factor message is the product of other messages incident to the variable. A factor $\rightarrow$ variable message is the max marginalization of the product of the factor and other messages incident to the factor. Message updates continue until all variables converge. The belief function of a variable is the product of all messages incident to the variable. An optimal configuration is obtained by iteratively fixing variables according to belief functions. Detailed operations are explained in Kschischang et al. (2001) and the supplementary text, Supplementary Material online.

Simultaneous optimization of all variables with max product may give a poor solution due to the highly underconstrained nature of the problem. There exist a large number of optimal/suboptimal configurations, and most of them are likely unrealistic. To alleviate this problem, we partition the variables into two sets—protein parent labels and edge presence—and iteratively fix one set of variables and infer the other until both converge. The initial values of protein parent labels are obtained from the template phylogenetic relations of observed proteins (gene trees) according to their sequences and domain compositions.

Constructing the Initial Protein Trees

The initial protein trees are inferred from the domain compositions and sequences of all enzyme proteins in the extant species. First, we extract unique domain compositions from all proteins and identify the (ancestral or extant) species where each domain composition first appears. Second, we employ a heuristic to build phylogenetic trees of domain compositions. It incrementally merges similar domain compositions with the constraint that younger domain compositions (domain compositions that first appear in more recent nodes in the species tree) should be placed under older domain compositions in the phylogenetic trees. In brief, traversing from the root of the species tree, it undertakes three types of merging operations sequentially in each species: incorporating the domain compositions containing inherited domains (the domains arising before the current species) with the existing phylogenetic trees, incorporating the domain compositions containing both inherited and novel domains (the domains arising in the current species), and incorporating the domain compositions containing novel domains alone. Each merging operation places a younger subtree of domain compositions under a descendant node of an older subtree. The outputs of this step are a collection of phylogenetic trees for domain compositions.

Fourth, we collect protein sequences with each unique domain composition and construct their phylogenetic tree using the MUSCLE program (Edgar 2004). MUSCLE automatically aligns the protein sequences and returns a nearest-neighbor tree derived from the aligned sequences. The protein tree associated with a domain composition is attached under its corresponding node in the domain composition trees. The results are merged phylogenetic trees of domain compositions and proteins.

To verify the robustness of the reconstructed networks against the choice of phylogenetic tree inference algorithms, we also construct the initial phylogenetic trees of protein sequences with each unique domain composition using the PhyloBayes 3.3 program (Lartillo et al. 2009). PhyloBayes applies the Monte Carlo Markov Chain (MCMC) sampling on phylogenetic tree structures and reports the consensus phylogenetic tree. The protein tree associated with a domain composition is attached under its corresponding node in the domain composition trees.

Fifth, the merged phylogenetic trees of domain compositions and protein sequences are reconciled with the species tree. A reconciliation algorithm (Zmasek and Eddy 2001) is employed to label the gene tree nodes with duplication and speciation events as well as the species nodes where they are located. The reconciled proteins trees may contain deep subtrees within each species. The factor graph model of the protein trees with intraspecies subtrees is very complicated and prone to overfitting, as the number of variable configurations grows exponentially with the intraspecies subtree depth. Because alterations of domain architectures and catalytic functions of proteins are relatively rare, with a sufficient coverage of sampled species, the intraspecies alteration events can be discarded. Therefore, we convert subtrees within a species into flat structures. In flattened structures, the species of a parent–child pair in the protein trees are also a parent–child pair in the species tree. The reconciled protein trees use the information of the species phylogeny and the most recent common ancestors of each domain composition. Thus, they are rooted trees.

A Reconstruction Algorithm of the Evolutionary History of DPR Networks

Figure 2 illustrates the algorithm of reconstructing the DPR network evolutionary history. The inputs of the reconstruction algorithm are the sequences, domain compositions, and catalytic functions of enzyme proteins in a number of extant species, and the phylogenetic tree of these species. The outputs are the inferred evolutionary history of DPR networks over the species tree. The initial protein trees are inferred from protein domain compositions and sequences using the aforementioned heuristic. The factor graph model described in equation (3) is then constructed. The initial values of protein parent labels are set according to the initial protein trees, whereas the binary variables of domain-protein and protein-reaction edge presence of extant species are fixed according to the observed DPR networks. Hidden variables are divided into two groups: 1) parent labels of proteins in all species and 2) domain-protein and protein-reaction edge presence of ancestral species. The max-product algorithm is invoked iteratively to impute the hidden variable values. In each iteration, we first fix the values of parent labels and infer the values of edge presence, then fix the values of edge presence and infer the values of parent labels. Iterations continue until all variables converge to fixed values. The converged variable configurations are transformed into an evolutionary history of DPR networks.

Data Sources

We extracted the DPR networks of 13 species from the databases of domain compositions and metabolic reactions (Pfam, Bateman et al. 2002; Uniprot, The UniProt Consortium 2011; BioCyc, Krummenacker et al. 2005). Supplementary table S1, Supplementary Material online, reports the summary information of the DPR networks of the 13 extant species. The taxonomic hierarchies of the 13 selected species were extracted from the National Center for Biotechnology Information Taxonomy Database (http://www.ncbi.nlm.nih.gov/Taxonomy/). Intermediate nodes with only one child in the hierarchies were collapsed. The collapsed species tree consists of 13 terminal (extant) nodes and 10 ancestral nodes and is shown in figure 3.

These species were chosen for their relatively rich information about domain architectures and metabolic reactions. Seven species are prokaryotes, and six species are eukaryotes. All prokaryotes are well-known pathogens: *E. coli*, *Bacillus anthracis*, *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Shigella flexneri*, and *Vibrio cholerae*. Eukaryotes include

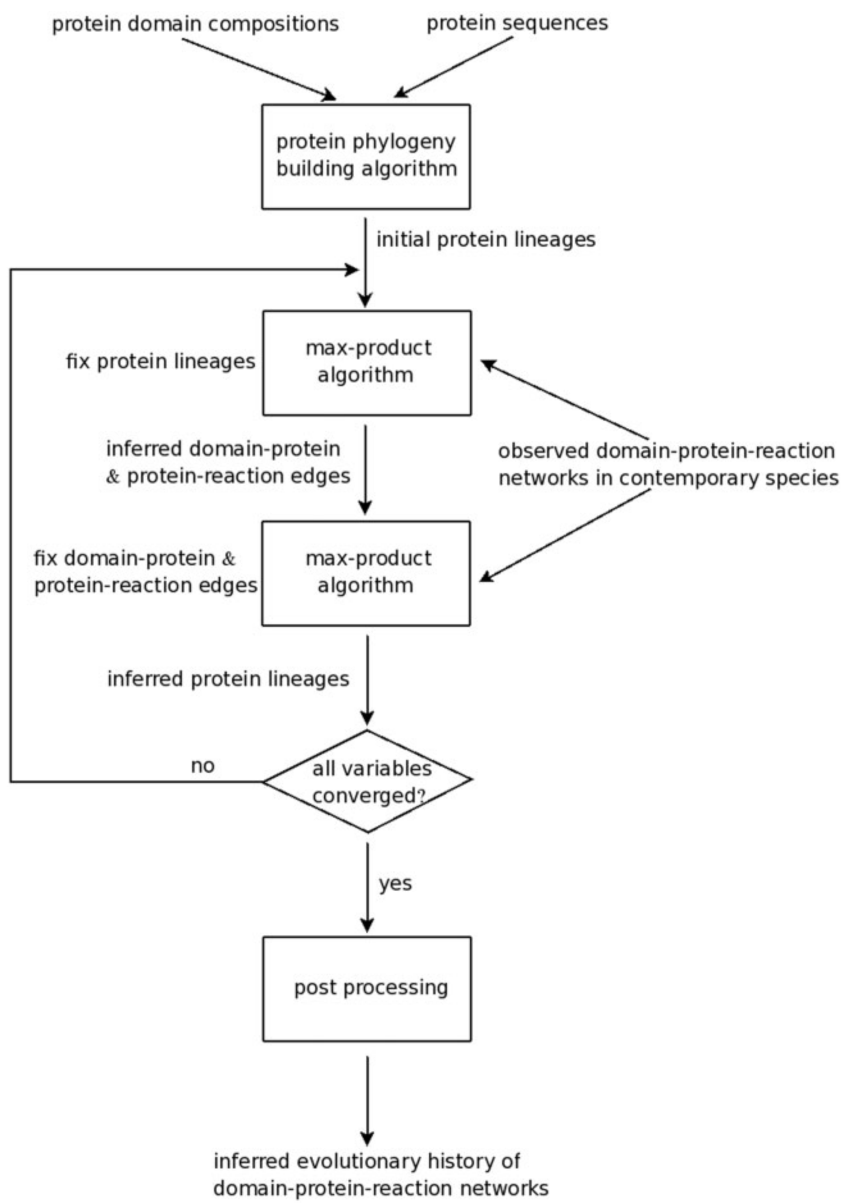

Fig. 2.—Schematic of the DPR network reconstruction algorithm. A collection of protein trees is inferred from their domain compositions and sequences alone. These protein trees set the initial values of protein lineage variables. With observed DPR networks in the contemporary species as inputs, the max-product algorithm is iteratively applied to infer the values of one set of variables (e.g., domain-protein and protein-reaction edges) by fixing the values of the other set of variables (e.g., protein lineages). Iteration continues until all variable values converge. The converged variable configuration is postprocessed to generate the inferred evolutionary history.

three mammals—human (*Homo sapiens*), mouse (*Mus musculus*), and cow (*Bos taurus*), fruit fly (*Drosophila melanogaster*), budding yeast (*Saccharomyces cerevisiae*), and *Plasmodium falciparum* (malaria parasites). Along the prokaryotic lineage, *B. anthracis* is a firmicute, *M. tuberculosis* is an actinobacteria, whereas the remaining five species are all proteobacteria. Along the eukaryotic lineage, mammals and fruit flies belong to coelomata; budding yeasts and animals belong to opisthokonta; and *P. falciparum* is a protist. The network sizes (total numbers of nodes and edges) of these species are positively correlated with their total numbers of genes ($R^2 = 0.7732$, supplementary fig. S1, Supplementary Material online). Human and mouse have considerably larger networks and proteomes than all the other species. Furthermore, *E. coli* has a disproportionally large network but comparable proteome size with other microbes. For instance, the network size and gene number of *E. coli* are 8,105 and 4,200, whereas those of *V. cholerae* are 5,632 and 3,828, respectively. Because gene function annotations in most species are far from complete, the disparity of network

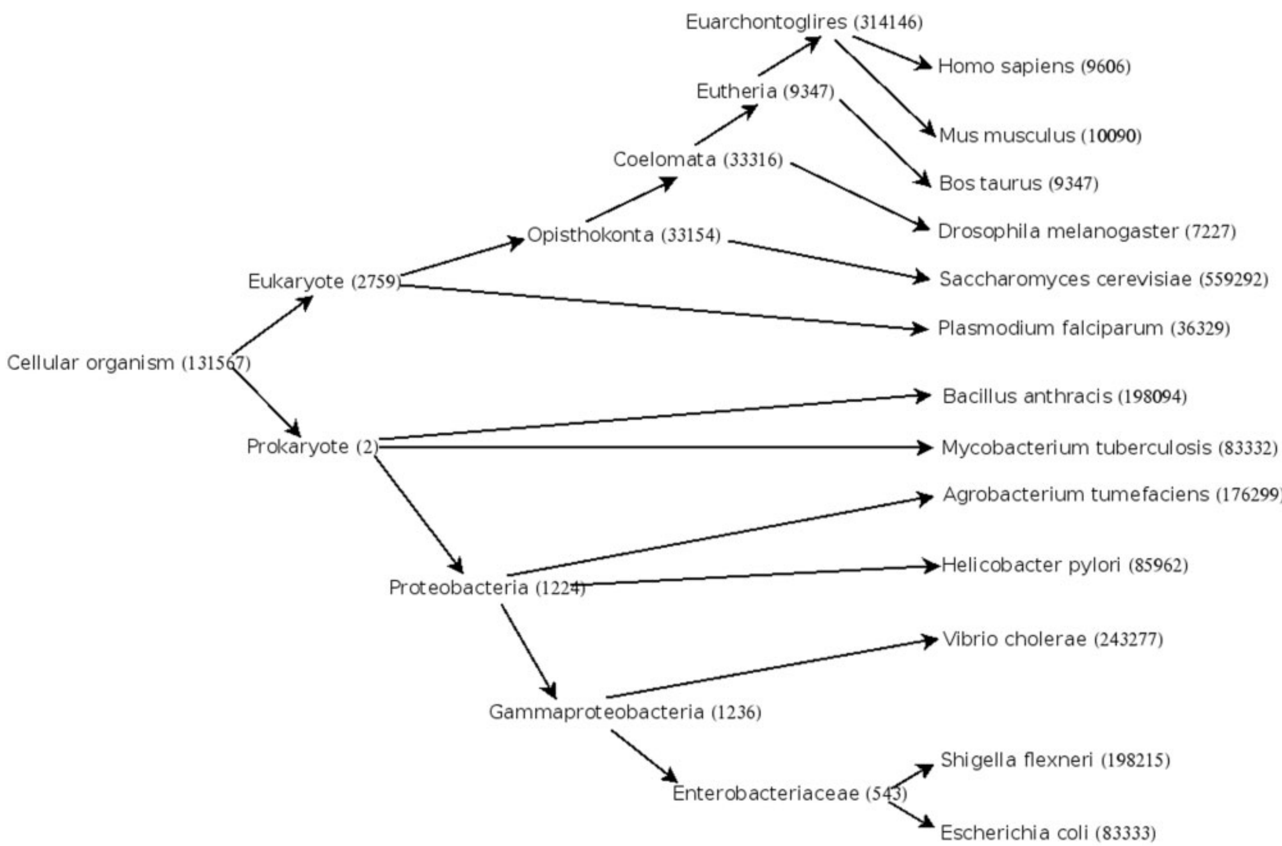

Fig. 3.—Phylogenetic tree of 13 selected species according to the National Center for Biotechnology Information (NCBI) taxonomy database. Branch lengths are not scaled to the evolutionary distances between genomes.

sizes is likely due to paucity of information on other microbes rather than expansion of *E. coli* networks during evolution.

A total of 386 metabolic pathways were downloaded from the KEGG database (Kanehisa and Goto 2000), and 750 metabolic pathways were extracted from BioCyc. In addition, reactions belonging to EC subclasses 2.7.10, 2.7.11, 2.7.12, 2.7.14, and 2.7.99 were labeled as the protein kinase pathway.

Results

Validation of the Reconstruction Algorithm

We validated the accuracy of the reconstruction algorithm using both simulation studies and cross validations on the data sets of 13 selected species. The purpose of simulation studies is to demonstrate that simultaneous reconstruction of all types of DPR network features (protein phylogenies, domain-protein, and protein-reaction edges) outperforms separate reconstructions of protein phylogenies and edge presences. A benchmark algorithm is to first reconstruct the protein phylogenies from sequences alone, then fixes the protein parent labels with reconstructed phylogenies and inferred the edge presences with dynamic programming. Sequence-based reconstruction algorithms (such as neighbor-joining and maximum likelihood methods) are error prone, as sequence evolution may not follow the models underlying the reconstruction algorithms. We expect to recover the DPR network evolutionary history in presence of these errors by incorporating the information of domain compositions and enzymatic functions of proteins in the model. Because our goal is not to evaluate particular sequence-based reconstruction algorithms but to demonstrate the merits of joint optimization, we skipped the steps of simulating sequence evolution and reconstructing phylogenies from simulated sequences. Instead, we explicitly perturbed the simulated protein phylogenies and treated the perturbed phylogenies as the outputs generated by a sequence-based reconstruction algorithm.

In simulation studies, a species tree of 5–20 nodes and a DPR network in the root species were randomly generated. We then sampled network change operations—additions and deletions of nodes and edges as well as duplications of protein nodes—sequentially to simulate the evolution of the DPR networks. The sampled protein trees were perturbed by reassigning parents to randomly selected nodes. The perturbed gene trees were treated as the phylogenies obtained from a sequence-based reconstruction algorithm and determined the initial values of protein parent labels for the max-product algorithm. Three parameters were varied: the total DPR network size, the rate of network evolution events, and the rate of perturbations to the initial gene trees. A total of 100 trials were performed for each parameter setting. The DPR networks of extant species and the perturbed gene trees were revealed to the inference algorithm. Error rates were measured by the difference between simulated and inferred networks normalized by the simulated network size. Error rates of our reconstruction algorithm and the benchmark method were reported. The experimental procedures of simulations are elaborated in section Materials and Methods and supplementary text, Supplementary Material online.

The top part of figure 4 shows the distributions of error rates with varying perturbation rates on gene trees. Error rates remain low (the mean value ≤ 0.1) for both methods when perturbation rates ≤ 0.01. Higher perturbation rates shift the error rate distributions to the right. However, the mean error rate was < 0.2 even when half of the proteins are assigned to wrong parents in the initial gene trees. The max-product algorithm significantly outperforms dynamic programming at high perturbation rates. This is expected because the benchmark method takes the perturbed gene trees as given and thus is sensitive to perturbation rates. In contrast, the max-product algorithm incorporates extant DPR networks to correct the phylogenetic relations of proteins and thus is more robust against gene tree errors. Error rates are robust against network sizes and increase with the rate of network change events (supplementary figs. S2 and S3, Supplementary Material online). The two methods exhibit similar error rates with varying network sizes and network evolution rates.

To further validate the reconstruction algorithm on real data sets, we hid features on selected domain-protein and protein-reaction pairs of extant species in the test set and predicted these features from the remaining training set. We extracted the DPR networks of 13 extant species. In each trial, we randomly included four types of pairs to the test set: 1) domain-protein edges, 2) domain-protein pairs that were not edges, 3) protein-reaction edges, and 4) protein-reaction pairs that were not edges. The presence/absence of these pairs was treated as hidden variables and were imputed by the reconstruction algorithm. The imputed values of the leave-out pairs were then compared with their true values. Error rates were measured by the ratios of the numbers of discrepant pairs and the test set sizes. The ratio of the test set size relative to the training set size + the test set size varied from 0.1 to 0.6, and 100 random trials were undertaken for each test set size. Separate inferences of protein phylogenies and edge presences were again used as the benchmark. The experimental procedures of cross validations are elaborated in section Materials and Methods and supplementary text, Supplementary Material online.

The bottom part of figure 4 shows the mean error rates for four types of pairs and the aggregate error rates with varying sizes of test sets. Intriguingly, max product and benchmark exhibit opposite patterns of mean error rates. Max product achieves 92–96% sensitivity (accuracy on edges) and 75–85% specificity (accuracy on nonedges) on domain-protein pairs, and 66–77% sensitivity and 83–87% specificity on protein-reaction pairs. In contrast, the benchmark algorithm achieves 80–92% sensitivity and 87–92% specificity on domain-protein pairs, and 25–55% sensitivity and 94–98% specificity in protein-reaction pairs. Overall, max product has higher sensitivity but lower specificity than the benchmark. However, max product has less mistakes on negative instances than the benchmark on positive instances, resulting in superior overall accuracy rates (82–85% vs. 75–82%). Furthermore, protein-reaction pairs yield substantially worse error rates than domain-protein pairs. The error rates of both methods are relatively robust against the test set sizes.

Evolutionary History of DPR Networks

We applied the reconstruction algorithm to infer the evolutionary history of the DPR networks of selected species. Figure 5 and supplementary table S2, Supplementary Material online, summarize the information of the evolution of the entire networks. The total numbers of network changes along each branch of the species tree and their contributions among distinct types of changes provide the following insights regarding the evolution of metabolic networks.

First, the common ancestor of cellular organisms (taxid 131,567) possesses a smaller network (network size 4,167) than human (network size 15,671), *E. coli* (network size 8,105), and most other extant species. This is sensible since the network of the eukaryote–prokaryote common ancestor constitutes the conserved core of metabolic systems (Morowitz 1999; Peregrin-Alves et al. 2003) and thus should be smaller than its derived descendants. Moreover, in spite of the disparate gaps of total network sizes, the majority of domains and reactions in eukaryotes and prokaryotes already appear in their common ancestor. The cellular organism common ancestor contains 888 domains and 1,307 reactions, whereas human has 1,488 domains and 1,784 reactions, and *E. coli* has 1,054 domains and 1,512 reactions. The network size differences are primarily attributed to proteins, domain-protein, and protein-reaction edges. Concordant with the findings in Chothia et al. (2003), the results suggest that most elementary components of metabolic systems (reactions and domains fulfilling certain catalytic functions) arise before the eukaryote/prokaryote split. Complexity of the systems accrues by protein duplications, recombinations of domain compositions, and reassignments of catalytic functions to orthologous and paralogous proteins.

Second, network sizes drop along many branches of the species tree. For instance, there is a significant decrease of network sizes from the common ancestor of opisthokonta (fungus/animal group, taxid 33,154, network size 7,920) to *S. cerevisiae* (network size 4,025) and from the common

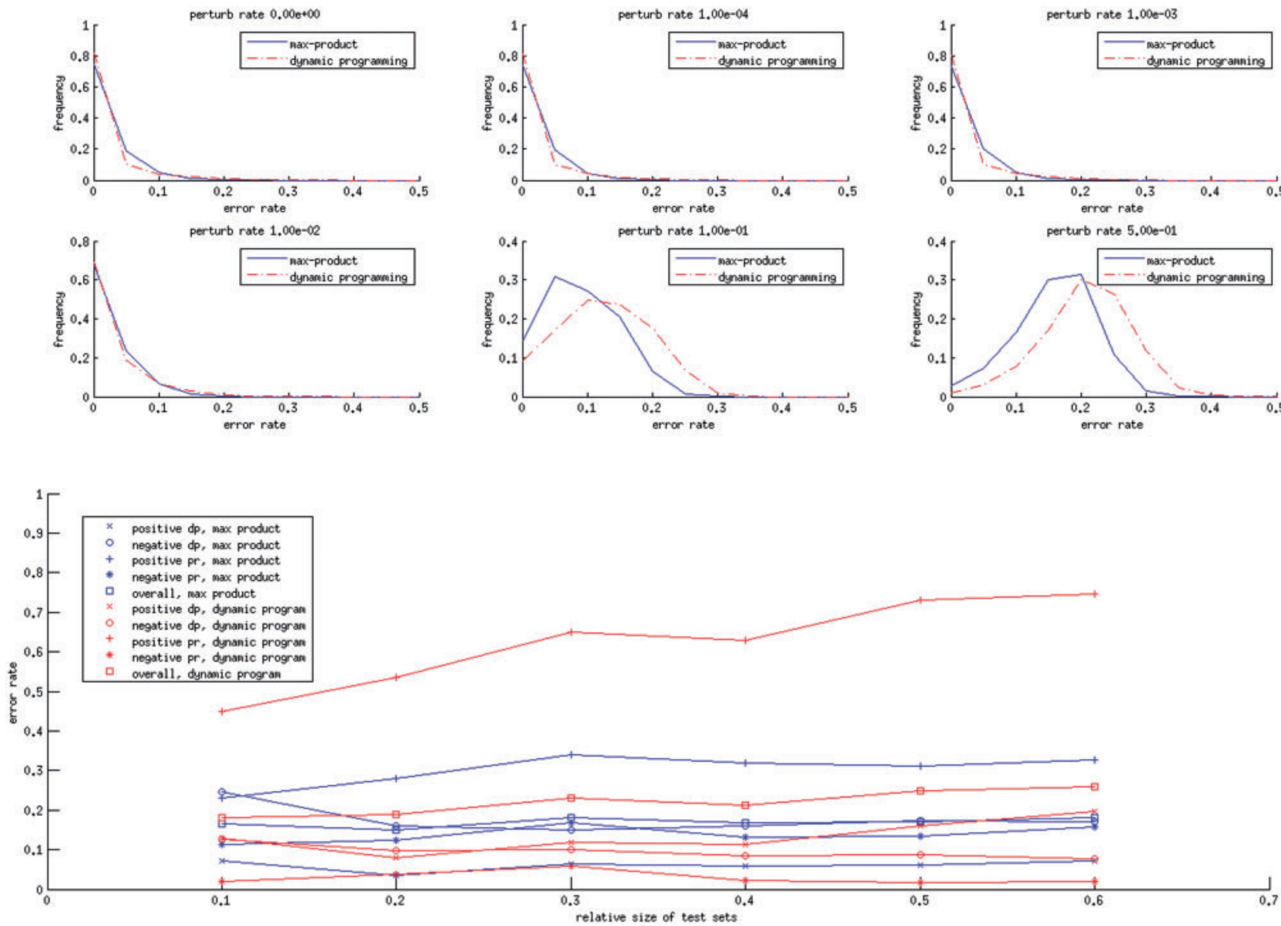

Fig. 4.—Validations of the reconstruction algorithm. Top: Distribution of error rates on simulated data, with varying perturbation rates on the initial protein trees. From top-left to bottom-right, the perturbation rate varies from 0 to 0.5. The error rate distributions of the max-product algorithm (solid blue) and dynamic programming (dashed red) are plotted. Bottom: Sensitivities, specificities, and overal error rates of cross-validation predictions on real data sets, with varying ratios of test set sizes and training set sizes. Crosses: sensitivities of domain-protein edges. Circles: specificities of domain-protein edges. Plus signs: sensitivities of protein-reaction edges. Asterisks: specificities of protein-reaction edges. Squares: overall rates. Blue symbols: max-product prediction outcomes. Red symbols: dynamic programming prediction outcomes.

ancestor of eukaryotes (taxid 2,759, network size 7,058) to *P. falciparum* (network size 3,381). Since the networks of human and *E. coli* are thoroughly annotated, their common ancestor (the common ancestor of cellular organisms) should retain a relatively complete conserved network. Consequently, we suspect that the majority of node and edge deletions in the DPR networks reflect missing information rather than real evolutionary processes. In contrast, additions and duplications of nodes and edges are more likely to reflect lineage-specific evolutionary processes.

Third, protein duplications are the dominant events of network evolution besides node and edge deletions. They comprise approximately 20% of the total network change events (10,808 of 52,669, supplementary table S2, Supplementary Material online). Frequent protein duplications indicate the prevalence of paralogous proteins with similar or identical domain architectures. These duplicated proteins may catalyze identical (isozymes) or distinct reactions (neofunctionizations or subfunctionizations). Functional analysis of some duplicated proteins are discussed later.

Fourth, the branches from the root to eukaryote and prokaryote ancestors possess strikingly similar contributions of network change mechanisms: protein duplications (51% along the eukaryotic branch and 44% along the prokaryotic branch), domain-protein edge additions (14% and 10%), domain creations (13% and 10%), protein-reaction edge additions (9% and 12%), reaction creations (7% and 5%), and protein creations (4% and 6%). Similarity of network sizes and relative contributions of network change mechanisms between eukaryote and prokaryote ancestors indicates similar evolutionary history of metabolic networks in the early stage of life. Furthermore, these early branches have higher fractions of domain, protein, and reaction creations than most other branches, suggesting that "innovation events,"

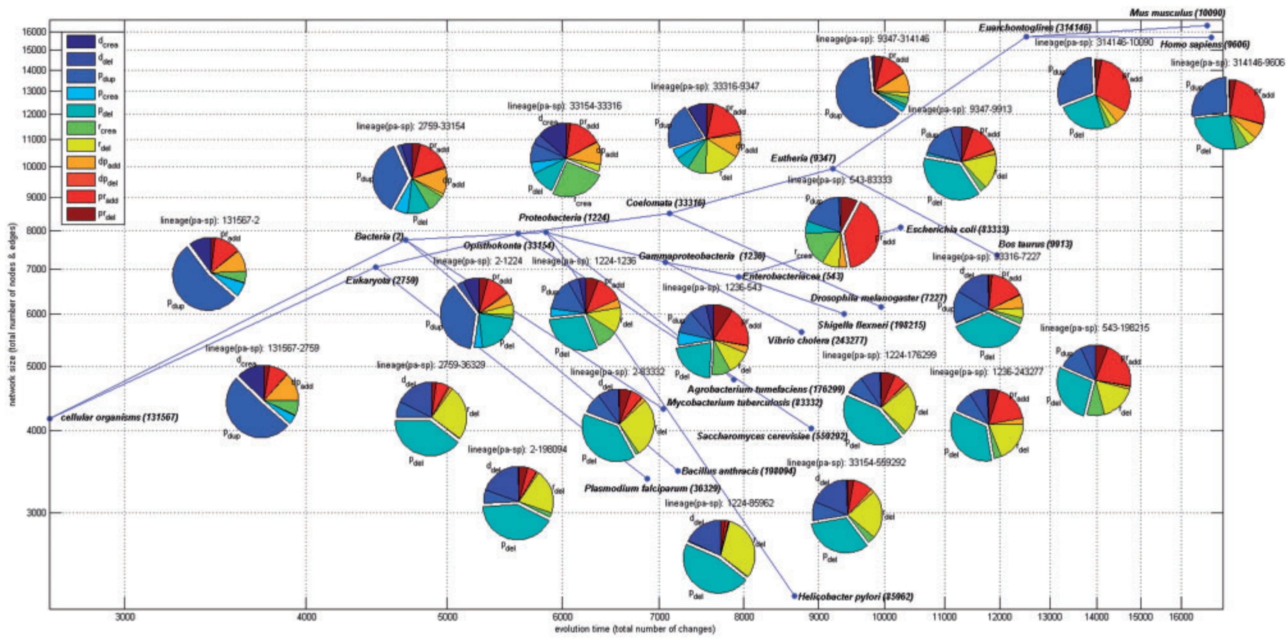

FIG. 5.—Summary of metabolic network evolution of 13 species. The topology of the phylogenetic tree (shown by blue lines) is extracted from the National Center for Biotechnology Information (NCBI) taxonomy. Each node represents a contemporary or ancestral species marked with its taxonomy name and ID. Vertical positions of nodes denote the total sizes of their DPR networks (node number + edge number). Horizontal distances between two adjacent nodes denote the total numbers of network change events between the adjacent species pairs. The compositions (contributions) of network change types between each pair of adjacent species nodes are visualized as pie charts and placed along their edges. Prominent network change mechanisms include protein duplications (medium blue), protein deletions (light blue), protein-reaction edge additions (red), domain-protein edge additions (orange), reaction creations (light green), and domain creations (dark blue).

creations of new domain compositions and reactions, play more important roles in the early stage of metabolic network evolution.

Fifth, after the eukaryote–prokaryote split metabolic network evolution follows divergent paths. Along the eukaryotic lineage, protein duplications remain prominent throughout most branches except the branch of opisthokonta (the animal–fungus group, taxid 33,154)-coelomata (the common group of mammals and insects, taxid 33,316) (7%). In the branch of eutheria (placental mammals, taxid 9,347)–euarchontoglires (the common group of primates and rodents, taxid 314,146), protein duplications even comprise 63% of network changes. Reaction creations expand substantially in the branch of opisthokonta–coelomata (26%) and remain marginal along other branches. Domain-protein edge additions constitute a moderate but stable fraction of network change mechanisms along most branches of the eukaryotic lineage. Intriguingly, from euarchontoglires to human and mouse, there are considerable numbers of protein duplications (26% in human and 24% in mouse) and protein-reaction edge additions (26% in human and 30% in mouse), but very few creations of novel domains (0.05% in both human and mouse), proteins (0.02% in human and 0% in mouse), or reactions (6% in human and 3% in mouse) in these two branches. Along the prokaryotic lineage, protein duplications are dominant only in the branch of prokaryotes–proteobacteria (taxid 1,224) (37%) and remain moderate or low throughout other branches. Protein-reaction edge additions are prominent in the branches of enterobacteriacea (taxid 543)–*E. coli* (taxid 83,333) (39%), enterobacteriacea–*S. flexneri* (taxid 198,215) (22%), gammaproteobacteria (taxid 1,236)–*V. cholera* (taxid 243,277) (17%), gammaproteobacteria–enterobacteriacea (18%), proteobacteria–gammaproteobacteria (13%), and prokaryote–proteobacteria (11%). Reaction creations concentrate primarily in the branches of enterobacteriacea–*E. coli* (16%) and proteobacteria–gammaproteobcteria (10%). Intriguingly, domain, protein creations, and domain-protein edge additions have consistently lower contributions along the prokaryotic lineage compared with the eukaryotic counterparts. The distinct patterns of network changes imply that eukaryotes and prokaryotes expand their metabolic networks with different mechanisms. Although both species acquire novel reactions and encounter frequent protein duplications along their lineages, eukaryotes expand their catalytic repertoire by creating new proteins and incorporating new or old domains to duplicated proteins. In contrast, prokaryotes tend to reassign novel catalytic functions to inherited or duplicated proteins with fewer alterations on their domain architectures. A similar observation was reported by Freilich et al. (2005). Furthermore, diverse patterns of network changes are also present along the eukaryotic lineage. The majority of novel domains and reactions arise in the common ancestors of the animal–fungus group (opisthokonta) or animals with body cavity (coelomata). Additions of domain-protein edges are frequent throughout the lineage from the cellular organisms to the primate–rodent group. Within the primate–rodent group, DPR networks are altered primarily by protein duplications and protein-reaction edge additions.

The initial protein trees generated by a phylogeny inference algorithm can in principle affect the exact wirings of the reconstructed DPR network evolutionary history but probably do not alter the aforementioned generic trends of network evolution. To verify this assumption, we compared the reconstructed network evolutionary histories using two phylogeny inference programs to generate the initial protein trees: MUSCLE (applies multiple sequence alignment and the neighbor-joining method to build the gene trees) and PhyloBayes 3.3 (applies MCMC to sample trees from aligned amino acid sequences and reports the consensus trees). Figure 5 and supplementary table S2, Supplementary Material online, summarize the reconstructed network evolutionary history using MUSCLE, and supplementary figure S4 and table S3, Supplementary Material online, summarize the reconstructed network evolutionary history using PhyloBayes 3.3. Despite the differences on the numbers of each type of network changes, their relative contributions along each branch (pie charts in figure 5 and supplementary figure S4, Supplementary Material online) are strikingly close. The largest difference of their relative contributions is 9.3%. Furthermore, all the five observations regarding the generic trends of DPR network evolution still sustain in the reconstructed network evolutionary history using PhyloBayes 3.3. These results indicate the robustness of the reconstructed network evolutionary history relative to the initial protein trees.

Mechanisms of Network Evolution in Distinct Metabolic Processes

In addition to the global summary on the evolutionary history of the entire metabolic networks, it is also important to understand the evolution of the DPR networks of individual metabolic pathways. We are interested in finding the enriched types of network change mechanisms for individual pathways, conditioned on the global trend from the entire networks. To fulfill this goal, we developed two methods to quantify the significance of evolutionary mechanism enrichments on specific pathways. The first method evaluates the contribution of each type of network change mechanisms. By assuming each event is sampled from a multinomial distribution, we calculate the probability of each type of network changes that maximizes the likelihood over the species tree. The second method calculates the reweighting factors relative to the contributions derived from the entire metabolic networks. The multinomial probability for each type of network change mechanisms is the product of the global contribution and the pathway-specific reweighting factor. The *P* values are calculated by comparing the contributions or reweighting factors of the empirical data with the results generated by randomly sampled reactions. The consensus of enriched network change mechanisms obtained by both methods are reported. Quantification of evolutionary mechanism enrichment is described in the supplementary text, Supplementary Material online.

A total of 386 and 750 metabolic pathways were downloaded from KEGG (Kanehisa and Goto 2000) and BioCyc (Krummenacker et al. 2005) databases, respectively, and their DPR subnetworks on the species tree were extracted. We identified the enriched evolutionary mechanisms for each pathway and categorized them into four disjoint classes:

1. Protein duplication: Protein duplications are enriched, and protein-reaction edge additions are possibly enriched. Other expansion mechanisms—domain creations, protein and reaction creations, and domain-protein edge additions—are not enriched.
2. Reaction creation: Reaction creations are enriched, and protein-reaction edge additions are possibly enriched. Other expansion mechanisms (including protein duplications) are not enriched.
3. Novel domain or protein generation: Domain creations, protein creations, or domain-protein edge additions are enriched. Reaction creations are not enriched.
4. Novel domain or protein generation and reaction creation: Domain creations, protein creations, or domain-protein edge additions are enriched. Reaction creations are enriched.

This categorization reflects a decreasing level of conservation in metabolic pathway evolution. In class 1, the rates of increasing the repertoire of domain combinations and reactions do not exceed the background values of the entire networks. In class 2, reaction creations are accelerated relative to the background values, but the expansion of proteins is not. In class 3, there is an increasing rate of generating novel domain compositions, but the rate of reaction creations does not exceed the background values. In class 4, generations of novel domain compositions and reactions both exceed the background rates.

Table 1 reports the summarized pathways belonging to each category. Complete information about the network change patterns and classes of metabolic pathways are reported in supplementary table S2, Supplementary Material online. Class 1 consists of the highest number (26) of pathways. The reactions and enzyme domain compositions of these pathways are largely conserved. The majority of pathways in this category synthesize and degrade the metabolites essential for all cellular organisms: fatty acids, amino acids, energy, and carbohydrates. This "metabolic core" arises in the early stage of life and remains highly conserved (Morowitz 1999; Peregrin-Alves et al. 2003). Isozymes are thought to

Table 1
Four Categories of Metabolic Network Evolution and Their Constituent Pathways

Category	Pathway	Pathway
Protein duplication	Fatty acid elongation	Fatty acid metabolism
	Alanine, aspartate, and glutamate metabolism	Valine, leucine, and isoleucine biosynthesis
	Biosynthesis of alkaloids	Trna metabolism
	Starch degradation	Folate transformation
	Pyrimidine metabolism	Methionine degradation
	Nicotine degradation	Lysine, threonine, and methionine biosynthesis
	Glycolysis	Tca cycle
	Entner–Doudoroff pathway	Gluconeogenesis
	Serine-isocitrate lyase pathway	Peptidoglycan biosynthesis
	Aspartate superpathway	Ribose and deoxyribose phosphate degradation
	Arginine, ornithine, and proline metabolism	Udp-sugar interconversion
	Formaldehyde assimilation	Formyithf biosynthesis
	Folate transformation	Heterolactic fermentation
Reaction creation	Steroid biosynthesis	Bile acid metabolism
	DDT degradation	Chlorocyclohexane and chlorobenzene degradation
	Benzene metabolism	Polyketide metabolism
	Peptidoglycan biosynthesis	Nitrotoleucine degradation
	Indole alkaloid biosynthesis	Monoterpenoid biosynthesis
	Insect hormone biosynthesis	Palmitate biosynthesis
	Noradrenaline and adrenaline degradation	Actinorhodin biosynthesis
	Tryptophan degradation	Myristate biosynthesis
Protein generation	Oxidative phosphorylation	Purine metabolism
	Alanine metabolism	C5-branched dibasic acid metabolism
	Carbon fixation in prokaryotes	Thiamine metabolism
	Kinases	Phosphatidylinositol signaling system
	mTOR signaling pathway	KDO2-lipid a biosynthesis
	Lipopolysaccharide biosynthesis	Arginine and polyamine biosynthesis
	Aromatic compound degradation	Foraldehyde assimilation
	Phospholipid biosynthesis	Ascorbate biosynthesis
	Acetyl-CoA assimilation	Bifidum pathway
	PIP metabolism	Atrazine degradation
Novel domain/protein and reaction creation	Glycan biosynthesis	Glycosaminoglycan biosynthesis
	Glycosaminoglycan degradation	Inositol phosphate metabolism
	Glycosylphosphatidulinositol-anchor biosynthesis	Sphingolipid metabolism
	Glycosphingolipid biosynthesis	Biotin metabolism
	Dolichyl-diphosphooligosaccharide biosynthesis	Heparan sulfate biosynthesis
	Zymosterol biosynthesis	Cholesterol biosynthesis
	Thyronamine and iodothyronamine metabolism	Thyroid hormone metabolism
	BMP signaling pathway	Adenosylcobalamin biosynthesis
	NAD biosynthesis	Ergosterol biosynthesis

optimize temporal/tissue-specific metabolic demands in multicellular organisms (Wilson 2003). However, contrary to this notion, we found abundant protein duplications in both eukaryote and prokaryote clades. The pathways of fatty acid metabolisms—fatty acid elongation in mitochondria (KEGG ID 00062) and fatty acid metabolism (KEGG ID 00071)—possess high rates of protein duplications along the branches cellular organism–eukaryote, cellular organism–prokaryote, and eutheria–euarchontoglires. Their network evolutionary patterns conform with the global patterns of network changes in figure 5, with stronger enrichment along those branches. In contrast, the pathways of amino acid metabolism possess high rates of protein duplications along several other branches, in addition to the three branches with high global duplication rates. For instance, the pathway of alanine, aspartate, and glutamate metabolism (KEGG ID 00250) encounters 37 and 30 protein duplications along the branches of cellular organism to prokaryote and eukaryote, 18 protein duplications along the branches of eutheria–euarchontoglires and prokaryote–proteobacteria, and 10 protein duplications along the branch of euarchontoglires–mouse. Similarly, the pathway of valine, leucine, and isoleucine biosynthesis

(KEGG ID 00290) encounters 37 and 14 protein duplications along the branches of cellular organism to prokaryote and eukaryote and 6 protein duplications along the branch of euarchontoglires–mouse.

Supplementary figure S4, Supplementary Material online, visualizes the DPR subnetwork evolution of the fatty acid elongation pathway. Both domains and reactions are relatively conserved. Eleven of 14 domains and 6 of 9 reactions in human already appear in the common ancestor of cellular organisms. From the common ancestor of cellular organisms, there are three domain creations, three protein creations, and seven reaction creations. Protein duplications occur primarily along the branches cellular organism–eukaryote (10), cellular organism–prokaryote (29), and eutheria–euarchontoglires (10).

Class 2 consists of 16 pathways. They have excessive numbers of reaction creations, whereas the rates of changes on enzyme proteins and domain compositions do not exceed the background rates. Neofunctionization—novel reactions are catalyzed by conserved proteins—may explain the patterns on these pathways. Class 2 includes pathways involved in steroid biosynthesis, bile acid biosynthesis, xenobiotics metabolism, secondary metabolite biosynthesis, and metabolism of terpenoids and polyketides. Among them, steroid biosynthesis is almost restricted to eukaryotes (Ourisson et al. 1994).

The steroid biosynthesis pathway in supplementary figure S5, Supplementary Material online, presents a remarkable example of class 2 patterns. Steroids are precursors of many signaling molecules in animals, plants, and fungi. On the one hand, the majority of reactions are created along the branches of eukaryote–opisthokonta (18 reaction creations), opisthokonta–coelomata (32 reaction creations), and coelomata–eutheria (8 reaction creations). On the other hand, half of the enzymes in the human subnetwork (10 of 20) have origins in the common ancestor of cellular organisms. From the root to human, there are only 6 domain creations, 6 protein creations, and 24 protein duplications. By examining the enzymes of steroid biosynthesis in human, we found most of their domains arise before the eukaryote/prokaryote split. For instance, squalene monooxygenase catalyzes the first oxygenation step in sterol biosynthesis (Squalene $\rightarrow$ (S)-Squalene-2,3-epoxide, EC number 1.14.99.7) (Laden et al. 2000). It consists of two domains: PF01266 (FAD-dependent oxidoreductase) and PF08491 (Squalene epoxidase). PF01266 also appears in cholesterol oxidase of *M. tuberculosis* (Bryzostek et al. 2007), suggesting that squalene monooxygenase may have an ancient origin.

Class 3 consists of 20 pathways and exhibit an opposite pattern from class 2. The rates of expanding the repertoire of domain compositions—domain and protein creations and domain-protein edge additions—are higher than the background values, yet those of reaction creations do not exceed the background rates. The members in class 3 cover a variety of biological processes including purine metabolism, signal transduction (e.g., protein kinases, phosphatidylinositol signaling system, and mTOR signaling pathway), and lipid metabolism (e.g., KDO_2 lipid A biosynthesis and phospholipid biosynthesis).

The evolutionary history of protein kinases in supplementary figure S6, Supplementary Material online, provides a remarkable example of the way to achieve systems complexity. In eukaryotes, regulatory signals are propagated by transfers of phosphate groups on tyrosines or serines/threonines of proteins (Gu and Gu 2003). Operating on a small number of reactions of the same type (phosphorylations/dephohsphorylations of amino acid residues), a diverse family of protein kinases have been evolved in eukaryotes. These kinases respond to different environmental stimulations and regulate distinct biological processes in a wide variety of cell types (Gu and Gu 2003). There are 172 domains, 482 proteins, and 26 reactions in human, and 50 domains, 17 proteins, and 3 reactions in the common ancestor of cellular organisms. The majority of human kinases are evolved from protein duplications and incorporations of novel domains to duplicated proteins (domain-protein edge additions). Along the branch of cellular organisms–eukaryote, there are 37 domain creations, 151 protein duplications, 8 protein creations, and 128 domain-protein edge additions but no reaction creations. Along the branch of opisthokonta–coelomata, there are 52 domain creations, 4 protein duplications, 6 protein creations, 54 domain-protein edge additions, and 17 reaction creations. Along the branch of coelomata–eutheria, there are 22 domain creations, 89 protein duplications, 2 protein creations, 69 domain-protein edge additions, and no reaction creations.

Class 4 consists of 18 pathways and have high rates of expansions of both domain compositions (domain and protein creations and domain-protein edge additions) and reactions (reaction creations) relative to the background rates. Strikingly, most pathways involved in glycan and glycosaminoglycan metabolism fall into this category. The domains, proteins, and reactions of glycan and glycosaminoglycan biosynthesis pathways arise primarily in the coelomata common ancestor. However, the pathway of glycosaminoglycan degradation (KEGG ID 00531) is conserved between human and *E. coli*.

Supplementary figure S7, Supplementary Material online, visualizes the DPR network evolution of glycosaminoglycan biosynthesis pathway (KEGG ID 00532). The majority of domains, proteins, and reactions arise in the common ancestors of coelomata and eutheria. There are 2 domains, 1 protein, and no reaction in the common ancestor of cellular organisms; 11 domains, 23 proteins, and 15 reactions in human; and no domains, proteins, or reactions in *E. coli*. From opisthokonta to coelomata, there are six domain creations, three protein creations, and five reaction creations. From coelomata to eutheria, there are two domain creations, three protein duplications, four protein creations, and five reaction creations. There are frequent protein duplications from euarchontoglires to human and mouse (17 and 12, respectively) but no creations of domains, proteins, or reactions. Therefore, the entire apparatus of glycosaminoglycan biosynthesis is probably established after the emergence of animals with body cavity (coelomata). Glycosaminoglycans form an essential component of connective tissues and may bind to proteins to form proteoglycans, which play important roles in cell adhesion and cellular matrix formation, signal transduction, and immune response (Gabius 1997). Diverse varieties of protein kinases and proteoglycans are both hallmarks of complex multicellular organisms. Metabolism of both families is catalyzed by novel domain compositions. However, protein phosphorylations/dephosphorylations involve in a small number of highly conserved reactions, whereas glycan and glycosaminoglycan biosynthesis requires novel reactions arising after the emergence of animals with body cavity.

Principles Underlying the Evolution of DPR Networks

A DPR network consists of two types of information pertaining to enzyme proteins: their domain architectures and catalytic functions. Intuitively, these two aspects should be tightly coupled. However, it remains unclear whether this intuition can be systematically substantiated from the evolutionary history of DPR networks. Here, we describe two quantitative principles linking the evolution of domain architectures and catalytic functions of enzymes.

Protein duplications are a dominant mechanism of network changes. Many proteins have homologous counterparts with similar or identical domain compositions in multiple species. It is sensible to assume that these duplicated proteins perform similar functions. To verify this hypothesis, we examined the reactions catalyzed by groups of homologous proteins. We divided the 15,052 enzymes in the 13 species into 1,146 families. Each family has a disjoint protein tree from the inferred evolutionary history of the DPR networks, thus consists of orthologous and paralogous proteins. We extracted the EC numbers of reactions catalyzed by proteins in each family. Each EC number consisted of four digits representing numerical classes of more refined levels. Thus, reactions sharing more EC digits are functionally more similar. For each protein family, we examined the first two EC digits of the catalyzed reactions and identified the dominant EC subclass containing the largest number of reactions. We then calculated the fraction of reactions belonging to the dominant EC subclass in each protein family and showed its distribution in the left part of figure 6. As anticipated, the majority of the protein families are dominated by one EC subclass. Among the 1,024 protein families with known EC numbers, 762 of them (74.41%) are dominated by one EC subclass (>90% of the reactions belonged to the same EC subclass). The results confirm functional similarity of duplicated proteins.

Domain creations are less frequent than protein duplications but still comprise a conspicuous portion of DPR network evolution. Although there are novel domains evolved to catalyze conserved reactions (e.g., kinases along the eukaryotes lineage), we suspect that the majority of novel domains arise to satisfy new catalytic demands. To verify this hypothesis, we performed two tests. First, we counted the occurrences of domain creations and reaction creations along each branch of the species tree. The right part of figure 6 shows the scatter plot of occurrences of domain and reaction creations along the 22 branches. The two quantities are weakly correlated (correlation coefficient 0.33). A close examination indicates that the correlation coefficient is compromised by outliers on the branches from the euarchontoglires common ancestor to human and mouse and from the enterobacteriaceae common ancestor to *E. coli*. These branches encounter smaller numbers of domain creations and much larger numbers of reaction creations. By removing this outlier, the correlation coefficient between domain and reaction creations becomes 0.64. The results suggest that the rates of reaction creations are proportional to the rates of domain creations.

Second, we counted the total numbers of novel and conserved reactions catalyzed by novel or conserved domains. Conserved domains catalyze far more reactions than novel domains (40,047 vs. 868), but a disproportionally higher fraction of reactions catalyzed by novel domains are novel reactions. Only 7.6% (3,039 of 40,047) of reactions catalyzed by conserved domains are novel, whereas 26.5% (230 of 868) of reactions catalyzed by novel domains are also novel. The χ^2 P value $< 10^{-16}$ and the hypergeometric P value $\leq 1.42 \times 10^{-61}$. The results further confirm the strong relations between domain and reaction creations.

Discussion

From the reconstructed DPR networks, we are able to deduce a high-level history of metabolic network evolution; 44% of domains and 34% of reactions appeared in the common ancestor of cellular organisms. After the prokaryote/eukaryote split, the ancestors of both kingdoms underwent frequent "innovation events": creations of new domains, reactions, and domain recruitments/reshufflings. Creations of novel domains and domain combinations in prokaryotes slow down after their early ancestor. Instead, existing domain combinations are assigned to more catalytic functions. In contrast, a large number of innovation events occur in the ancestors of animal/fungus group (opisthokonta), animals with body cavity (coelomata), and placental mammals (eutheria). These innovations pertain to multicellular physiology: signal transduction, cell adhesion, tissue-specific metabolic demands, and so on.

Protein duplications are considered as a major mechanism to expand gene repertoires (Ohno 1970). In metabolic network evolution, protein duplications comprise the largest portion of network change events. They serve two roles: 1) increase the redundancy (isozymes) of enzymes and 2) acquire additional catalytic functions on reactions with

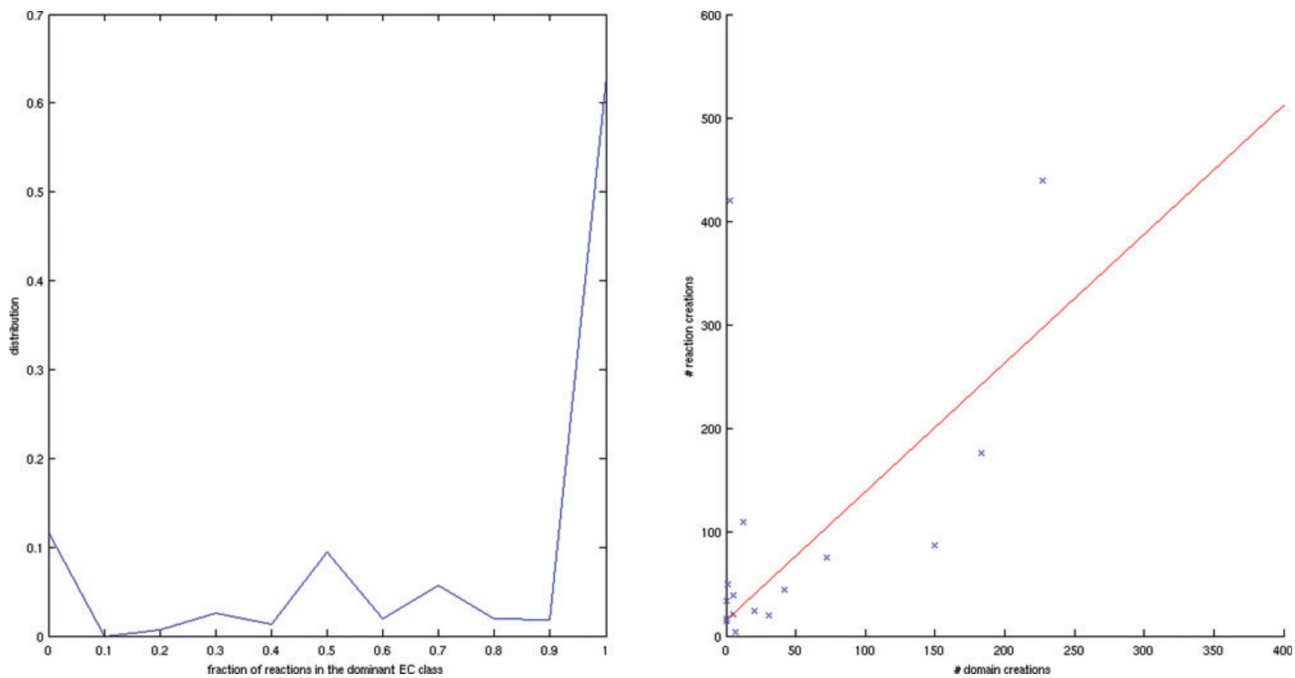

Fig. 6.—General rules relating domains and reactions in the DPR networks. Left: Distribution of the fraction of reactions in the dominant EC class among all protein family. Horizontal axis: fraction of reactions in the dominant EC class. Vertical axis: distribution of the reaction fraction among all protein families. Right: Scattered plot of the numbers of domain creations and reaction creations along each branch of the species tree.

similar chemical operations. Surprisingly, protein duplications dominate the network change mechanisms in the early evolution of both eukaryotes and prokaryotes. Despite the prominent differences in their proteome sizes, the number of protein duplications in human is about twice as the number of those in *E. coli* (5,111 vs. 2,230).

The pathways involved in metabolism of primary metabolites—energy, carbohydrates, amino acids, and nucleotides—are highly conserved among all living organisms. These pathways constitute the backbone of metabolic systems that dated back to primitive organisms 3.8 billion years ago. In contrast, the pathways involved in metabolism of phospholipids, peptidoglycans, glycoproteins, steroids, glycolipids, phosphorylated proteins, and environmental toxins all exhibit lineage-specific variations. Most of these metabolites are required for multicellular physiology such as signal transduction, cell–cell communication, and extracellular matrix formation. Intriguingly, multiple pathways involved in cell–cell communication and signaling possess distinct patterns of evolutionary mechanisms. Steroid biosynthesis has enriched reaction creations but retains a relatively conserved repertoire of domain compositions and proteins. Protein kinases possess many domain and protein creations but retain conserved reactions. Glycosaminoglycan and glycan metabolism has a high rate of protein and reaction creations and domain recruitments.

Simultaneous reconstruction of domain architectures and catalytic functions of enzyme proteins can both provide more complete characterization about metabolic systems and give more accurate predictions of one type of features with information of another type of features. To justify the benefits of joint optimization in prediction, we compare our simultaneous reconstruction algorithm with a benchmark method of separate reconstructions of protein phylogenies and domain-protein and protein-reaction edge presence. For protein phylogeny reconstruction, we incurred simulation studies by introducing perturbations (errors) to the initial protein phylogenies. Simultaneous reconstruction outperforms the benchmark when the perturbation rate is high (≥ 0.1), indicating the benefit of including protein function information when sequence-based phylogenetic reconstruction is erroneous. For domain-protein and protein-reaction edge inference, we incur cross validations to predict the presence/absence of edges in a test set giving a separate training set. Simultaneous reconstruction yields poorer specificity (accuracy on nonedges) but better sensitivity (accuracy on edges) and overall accuracy than the benchmark.

Every model has to balance the tradeoff between tractability of the problem and fitness to the phenomena. Our reconstruction algorithm is based on several simplifying assumptions thus cannot handle the following events of DPR network evolution. First, domain fusions are not considered as they violate the assumption of a single parent of each protein. Second, HGTs are also discarded as they cause the parent species of a protein to differ from that of its host genome and would drastically increase model complexity.

Third, the model assigns a unique domain composition to each protein thus ignores its multiple splice variants. Fourth, permutations of domain orders are also ignored due to problem tractability and scarcity of the events. Fifth, frequent emergence of the same domain architecture in multiple lineages may distort the initial protein trees and thus affect the reconstructed network evolutionary history. Despite these limitations, our reconstruction algorithm sheds lights on general patterns of evolutionary mechanisms of the DPR networks. More detailed and precise reconstruction of DPR network evolution requires more accurate protein phylogenies, complete information of protein functions and domain architectures in multiple species, and a more comprehensive and complex model to characterize the aforementioned mechanisms of network changes.

Supplementary Material

Supplementary text, figures S1–S7, and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Ming Chen, Alex Yu, and Manyuan Long for the inputs and advice during the preparation of the manuscript, Nils Baas for the discussion on hierarchical network models, and Tsung-Ju Lee for the technical support on data processing. This work was supported by the Institute of Statistical Science of Academia Sinica and the National Science Council grants of the Republic of China, Taiwan to C.H.Y. and S.S. (grant numbers 99-2221-E-001-021 and 100-2118-M-001-008-MY2).

Literature Cited

Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial, and eukaryotic proteomes. J Mol Biol. 310:311–325.

Bateman A, et al. 2002. The Pfam protein families database. Nucleic Acids Res. 30(1):276–280.

Behzadi B, Vingron M. 2006. Reconstructing domain compositions of ancestral multi-domain proteins. Comparative Genomics, Lecture Notes in Computer Science. RECOMB-CG 2006, LNBI 4205, p. 1–10.

Bjorklund AK, et al. 2005. Domain rearrangements in protein evolution. J Mol Biol. 353:911–923.

Borenstein E, et al. 2008. Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc Natl Acad Sci U S A. 105: 14482–14487.

Bornberg-Bauer E, et al. 2005. The evolution of domain arrangements in proteins and interaction networks. CMLS Cell Mol Life Sci. 62: 435–445.

Bryzostek A, Dziadek B, Rumijowska–Galewicz A, Pawelczyk J, Dziadek J. 2007. Cholesterol oxidase is required for virulence of *Mycobacterium tuberculosis*. FEMS Microbiol Lett. 275(1):106–112.

Caetano-Anolles G, Caetano-Anolles D. 2003. An evolutionarily structured universe of protein architecture. Genome Res. 13:1563–1571.

Caetano-Anolles G, et al. 2009. The origin and evolution of modern metabolism. Int J Biochem Cell Biol. 41:285–297.

Chothia C, et al. 2003. Evolution of the protein repertoire. Science 5626: 1701–1703.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acid Res. 32(5): 1792–1797.

Fong JH, et al. 2007. Modeling the evolution of protein domain architectures using maximum parsimony. J Mol Biol. 366:307–315.

Forslund K, et al. 2007. Domain tree-based analysis of protein architecture evolution. Mol Biol Evol. 25(2):254–264.

Freilich S, et al. 2005. The complement of enzymatic sets in different species. J Mol Biol. 349(4):745–763.

Freilich S, et al. 2008. Metabolic innovations towards the human lineage. BMC Evol Biol. 8:247.

Gabius HL. 1997. Glycosciences: status and perspectives. London: Chapman & Hall.

Gough J. 2005. Convergent evolution of domain architectures (is rare). Bioinformatics 21:1464–1471.

Gu J, Gu X. 2003. Natural history and functional divergence of protein tyrosine kinases. Gene 317:49–57.

Hein J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. Math Biosci. 98:185–200.

Kaessmann H, et al. 2002. Signatures of domain shuffling in the human genome. Genome Res. 12:1642–1650.

Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27–30.

Krummenacker M, et al. 2005. Querying and computing with BioCyc databases. Bioinformatics 21:3454–3455.

Kschischang F, Frey B, Loeliger H. 2001. Factor graphs and the sum-product algorithm. IEEE Trans Inform Theory. 47:498–519.

Kunin V, Ouzounis CA. 2003. The balance of driving forces during genome evolution in prokaryotes. Genome Res. 13:1589–1594.

Laden BP, Tang Y, Porter TD. 2000. Cloning, heterologous expression, and enzymological characterization of human squalene monooxygenase. Arch Biochem Biophys. 374(2):381–388.

Lartillo N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25(17):2286–2288.

Ma J, et al. 2008. The infinite sites model of genome evolution. Proc Natl Acad Sci U S A. 105:14254–14261.

Mithani A, Preston GM, Hein J. 2009. A stochastic model for the evolution of metabolic networks with neighbor dependence. Bioinformatics 25(12):1528–1535.

Mithani A, Preston GM, Hein J. 2010. A Bayesian approach to the evolution of metabolic networks on a phylogeny. PLoS Comp Biol. 6(8):e1000868.

Morowitz HJ. 1999. A theory of biochemical organization, metabolic pathways, and evolution. Complexity 4:39–53.

Ohno S. 1970. Evolution by gene duplication. New York: Springer.

Ourisson G, Nakatani Y. 1994. The terpenoid theory of the origin of cellular life: the evolution of terpenoids to cholesterol. Chem Biol. 1(1):11–23.

Pal C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet. 37(12):1372–1375.

Peregrin-Alves J, Tsoka S, Ouzounis CA. 2003. The phylogenetic extent of metabolic enzymes in pathways. Genome Res. 13:422–427.

Pfeiffer T, Soyer OS, Bonhoeffer S. 2005. The evolution of connectivity in metabolic networks. PLoS Biol. 3:1269–1275.

Pinney JW, et al. 2007. Reconstruction of ancestral protein interaction netwroks for the bZIP transcription factors. Proc Natl Acad Sci U S A. 104:20449–20453.

Przytycka T, et al. 2006. Graph theoretical insights into Dollo parsimony and evolution of multidomain proteins. J Comp Biol. 13(2):351–363.

Schmidt EE, Davies CJ. 2007. The origins of polypeptide domains. Bioessays 29:262–270.

Schmidt S, et al. 2003. Metabolites: a helping hand for pathway evolution? Trends Biochem Sci. 28(6):336–341.

Teichmann SA, et al. 2001. The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. J Mol Biol. 311(4):693–708.

UniProt Consortium. 2011. Ongoing and future developments at the universal protein resource. Nucleic Acids Res. 39:D214–D219.

Vaske CJ, et al. 2009. A factor graph nested effects model to identify networks from genetic perturbations. PLoS Comp Biol. 9(5):e1000274.

Vogel C, et al. 2005. The relationship between domain duplication and recombination. J Mol Biol. 346(1):355–365.

Wiedenhoeft J, Krause R, Eulenstein O. 2011. The plexus model for the inference of ancestral multi-domain proteins. IEEE/ACM Trans Comp Biol Bioinform. 8(4):890–901.

Wilson JE. 2003. Isozymes of mammalian hexokinase: structure, subcellular localization, and metabolic function. J Exp Biol. 206: 2049–2057.

Wuchty S. 2001. Scale-free behavior in protein domain networks. Mol Biol Evol. 18(9):1694–1702.

Yeang CH, Baas N. Evolution of domain compositions in the metabolic networks of human and *Escherichia coli*. Proceedings of the World Congress in Computer Science, Computer Engineering and Applied Computing (WORLDCOMP), 2009 Jul. 13–16; Las Vegas, NV.

Yeang CH, Ideker T, Jaakkola T. 2004. Physical network models. J Comp Biol. 11(2–3):243–262.

Zmasek CM, Eddy SR. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17:821–828.

Associate editor: Shu-Miaw Chaw