Analysis of the bipartite networks of domain compositions and metabolic reactions

(Invited Paper)

Chen-Hsiang Yeang
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan.

Abstract

It is widely accepted that complexity of biological systems arises from combinations of common subunits. In this work we investigate the combinatorial patterns of protein domains in the metabolic networks and find several general rules in the patterns of domain combinations and their evolution. First, the reactions catalyzed by a domain subunit carrying specialized or accessory functions are often subsumed to the reactions catalyzed by a domain subunit carrying generic operations. Second, some reactions contain multiple domains in their enzymes because they require multiple chemical operations carried by distinct domains. Third, pleiotropy (multi-functionality) of enzymes either results from the similarity of the catalyzed reactions or is achieved by merging domains with distinct functions. Fourth, comparison of domain compositions and metabolic reactions between human and Escherichia coli suggests that requirements for novel reactions, redundancy and pleiotropy are the dominant driving factors for domain evolution. The methods and results provide a framework to study the combinatorial complexity of a biological system.

Introduction

One remarkable discovery from the recent development of genomics is that the complexity and diversity of biological systems do not arise from genome size and sequence disparity alone. It is widely accepted that most differences result from the combinations of a set of common subunits. Combinatorial complexity has been demonstrated in multiple biomolecular systems such as the transcription regulatory circuitry (Carroll 2005), alternative splicing of exons (House and Lynch 2008; Ben-Dor et al. 2008), and domain compositions of proteins (Apic, Gough and Teichmann 2001; Chothia et al. 2003).

Domains are polypeptide subunits of proteins that constitute similar molecular structures or sequences. Combinatorial complexity is manifested in the domain architectures of proteins, as similar domains can be utilized in different proteins and combinations of domains yield proteins of diverse functions. Comparison of domain sequences and compositions has already led to insightful discoveries such as the enriched patterns of domain combinations (Apic, Gough and Teichmann 2001; Chothia et al. 2003), power-law distribution of the number of co-occurring domain partners (Wuchty 2001; Ravasz et al. 2002), conserved linear orders of domains (Vogel et al. 2004), and possible mechanisms of domain formation and recombination (Vogel et al. 2005; Schmidt and Davies 2007; Kaessmann et al. 2007).

There are three key questions regarding domain combinations. What are the general patterns of domain combinations? How are the domain combinations related to the functions of proteins? How are the domain combinations evolved? Metabolic networks are an ideal system to answer these questions for two reasons. First, many metabolic reactions share common substrates and chemical operations. It is reasonable to test whether modularity of metabolic reactions is achieved by domain combinations. Second, comprehensive information about protein functions and domain compositions of metabolic enzymes are available in many species. There are a rich collection of databases about metabolic reactions (e.g., Biocyc, Karp et al. 2005; KEGG, Kanehisa and Goto 2000; Recon 1, Duarte et al. 2007), protein functional annotations and sequences (e.g., RefSeq, Pruitt, Tatusova and Maglott 2007; Uniprot, Leinonen et al. 2004), domain sequences and architectures (e.g., Pfam, Bateman et al. 2002; ProDom, Corpet et al. 2000; SCOP, Murzin et al. 1995). Comparative studies of these data revealed properties regarding the evolution of the metabolic system (e.g., Chothia et al. 2003; Dandekar et al. 1999; Pal, Papp and Lercher 2005; Bowers et al. 2004). However, a comprehensive investigation of metabolic networks to address the three questions above has not been pursued yet.

In this study we address these three questions by investigating the domain compositions of enzyme proteins in the metabolomes of *Escherichia coli* and human. First, by inspecting the the reactions catalyzed by each domain family, we find most inclusion relations coincide with the functional dependencies of the corresponding domains. Second, we test the modularity hypothesis by identifying the reactions which require the chemical operations carried by distinct domains in their enzymes. Third, we explain the pleiotropy of enzyme proteins/complexes catalyzing multiple reactions by the operational similarity of the catalyzed reactions and domain fusions. Fourth, by examining metabolic reactions and domain compositions between *E. coli* and human, we find the need for new reactions, redundancy and pleiotropy

are the major factors driving the evolution of novel domain combinations.

Databases and data processing

We downloaded the human and *E. coli* subsets of the Biocyc database (Karp et al. 2005). Each dataset contains the substrates and enzymes of reactions and the metabolic pathways they belong to. Components of the same enzyme complex are treated as co-occurring proteins in the same reaction, while distinct proteins catalyzing the same reaction are treated as alternative enzymes. 1661 *E. coli* reactions and 1313 human reactions were extracted from the datasets.

Domain architectures of 2238 *E. coli* enzyme proteins and 2635 human enzyme proteins were extracted from the Pfam database (Bateman et al. 2002). 5122 domain families appear in *E. coli* or human. For simplicity the domain architecture of a protein is reduced to a "bag of domains" representation: we discarded the order of domains in a protein and treated multiple occurrences of the same domain identical. 8042 distinct domain compositions were extracted from *E. coli* and human proteins.

1460 *E. coli* and 996 human reactions are catalyzed by enzymes with known domain compositions. For simplicity we collapsed the homologous reactions with identical substrates into one reaction. There are 1883 reduced reactions in *E. coli* and/or human. 1776 of them are catalyzed by enzymes of known domain compositions, and 479 reactions contain homologous reactions in both species.

Bipartite networks of domain compositions and enzymatic reactions

The relations of protein domains and metabolic reactions are represented as a bipartite graph G_1 constituting nodes of protein domain families and metabolic reactions. An edge in G_1 from domain A to reaction B denotes that A appears in enzymes catalyzing B. Similar to other biological networks, G_1 exhibits a power-law distribution in their connectivity. Table 1 lists the top 10 highly connected domains and reactions in G_1. The hub domains are involved in transport, NAD and NADP dependent oxidation/reduction, and transfer of amino groups. Highly connected reactions are catalyzed by either large protein complexes (such as cytochrome c oxidoreductase, RNA polymerase and ATP synthetase) or a diverse family of proteins (such as protein kinases). We also counted the number of domain compositions containing each domain and found the distribution of the membership counts also followed a power-law distribution. The results resemble the analysis in Wuchty 2001 of the domain membership network. Highly connected domains appear in many signaling proteins (EGF-like domain, SH3 domain), transcription factors (Zinc finger, C2H2 type), and tandem repeats (Ankyrin repeat). These domains occur almost exclusively in human, suggesting complexity of domain compositions along eukaryotes lies on cell-cell and cell-environment communications.

Table 1. Top 10 highly connected domains and reactions

Domains:

Description	# reactions
ABC transporter (PF00005)	58
Short chain dehydrogenase (PF00106)	36
Binding-protein-dependent transport system (PF00528)	35
Major Facilitator Superfamily (PF07690)	31
Aldehyde dehydrogenase family (PF00171)	30
Pyridine nucleotide-disulphide oxidoreductase (PF07992)	26
Cytochrome P450 (PF00067)	25
Haloacid dehalogenase-like hydrolase (PF00702)	24
Pyridine nucleotide-disulphide oxidoreductase (PF00070)	23
Sugar (and other) transporter (PF00083)	20

Reactions:

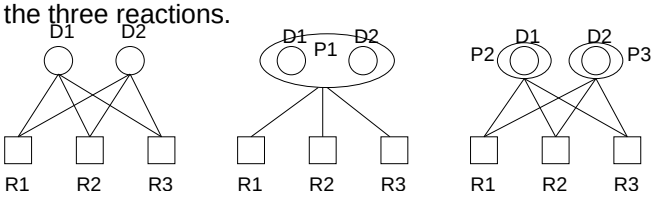
Description	# domains
protein phosphorylation	49
protein tyrosine phosphorylation	40
NADH dehydrogenase	37
cytochromoe c oxidation	34
NADH dehydrogenase	34
RNA polymerization	32
protein tyrosine dephosphorylation	28
ATP dependent proton transport	26
protein serine/threonine phosphorylation	25
DNA polymerization	18

Identification of domain and reaction subunits

Domains and reactions may not be the basic subunits of the network. Some domains always co-appear in the same enzyme proteins/complexes, and some reactions are catalyzed by the same set of domains. We grouped these domains and reactions together and termed them domain and reaction subunits. Members of a domain subunit need not to co-appear in the same protein. For instance, a subunit of 5 domains are involved in the oxidation of xanthine. In human these domains aggregate in the same protein (xanthine dehydrogenase) (Kimiyoshi et al. 1993), whereas in *E. coli* they split into several proteins in a complex (xdhA, xdhB, xdhC of xanthine dehydrogenase) (Xi, Schneider and Reitzer 2000). Figure 1 illustrates the domain and reaction subunits.

To identify domain subunits we constructed an undirected graph G_{du} of domains. Two domains are adjacent in G_{du} if they co-occur in the same enzyme proteins/complexes of all the reactions they catalyze. Isolated cliques in G_{du} are domain subunits. Similarly, a reaction subunit consists of reactions catalyzed by the same set of domain subunits. To find reaction subunits we constructed an undirected graph G_{ru} of reactions. Two reactions are adjacent in G_{ru} if the sets of all domain subunits catalyzing them are identical. Isolated cliques in G_{ru} are reaction subunits.

Figure 1. Domain and reaction subunits. Circles: domains. Squares: reactions. (1)Two domains D_1, D_2 are involved in three reactions R_1, R_2, R_3. (2)D_1 and D_2 appear in the same protein P_1, and P_1 catalyzes the three reactions. (3)D_1 and D_2 appear in distinct isozymes of the three reactions.



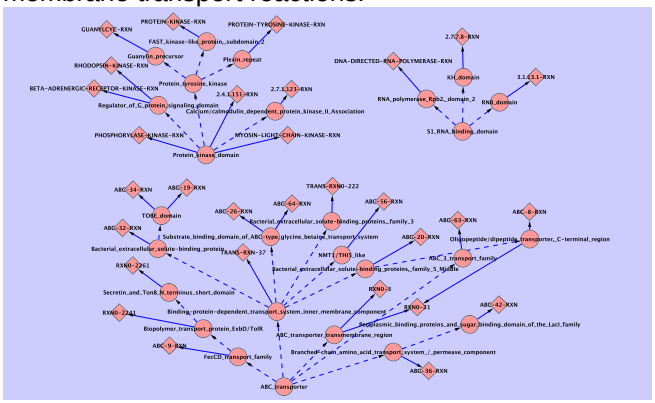
A hierarchical network of domain compositions and reactions

The functional relations of domain subunits are revealed by the reactions they catalyze – effect reactions. Two sets of reactions may be identical, included, overlapped but not included, or disjoint. These relations are implicit in G_1. To better represent this information we applied parsimony and transitivity of inclusion relations and transformed G_1 into a directed, hierarchical network G_h:

1) For each domain subunit d identify its effect reactions $R(d)$.
2) Construct a graph $G_d = (V_d, E_d)$ of domain subunits. A directed edge (d_1, d_2) in G_d connects d_1 to d_2 if $R(d_2) \subseteq R(d_1)$, or $|R(d_1) \cap R(d_2)| \geq 5$ and $|R(d_1) \cap R(d_2)| \geq |R(d_2)| - 1$. Bi-directed edges are allowed.
3) For each domain subunit $d \in V_d$, find all ancestor domain subunits whose effect reactions cover those of d: $An(d) = \{d' | (d', d) \in E_d\}$.
4) For each domain subunit $d \in V_d$, prune edges from indirect ancestors. A uni-directional edge (d, d') is removed from E_d if $\exists d''$ such that $d'' \in An(d')$ and $d \in An(d'')$. Proceed until there is no edge from indirect ancestors.
5) Build a bi-partite graph $G_h = (V_h, E_h)$ of domain and reaction subunits. A directed edge (d, r) in G_h connects d to r if $r \in R(d)$. Copy all edges of G_d to G_h.
6) For each reaction subunit $r \in V_h$, prune edges from indirect ancestors. An edge $(d, r) \in E_h$ is removed from E_h if $\exists d'$ such that $d \in An(d')$ and $(d', r) \in E_h$. Proceed until there is no domain-reaction edge from indirect ancestors.
7) Return G_h.

Similar methods have been applied to reconstruct the causal order of genes in a regulatory network from knock-out experiments (Wagner 2001; Markowetz, Bloch and Spang 2005). The resulting network G_h preserves the information of G_1 and makes the inclusion relations explicit. A directed edge denotes an inclusion or catalytic relation irreducible from other inclusion and catalytic relations. G_h is hierarchical as it conveys the nested relations of inclusion.

Figure 2. Three examples of hierarchies. Circles: domains. Diamonds: reactions. Dashed lines: effect reactions of one domain contain the effect reactions of another. Solid lines: a domain and its inclusion closure catalyze the reaction. From top left to bottom right. (1)Protein kinase domain pairs with other domains in various phosphorylation reactions. (2)S1 RNA binding domain pairs with other domains in RNA polymerization, ribonuclease and mRNA processing. (3)ABC transporter domain pairs with other domains in various membrane transport reactions.

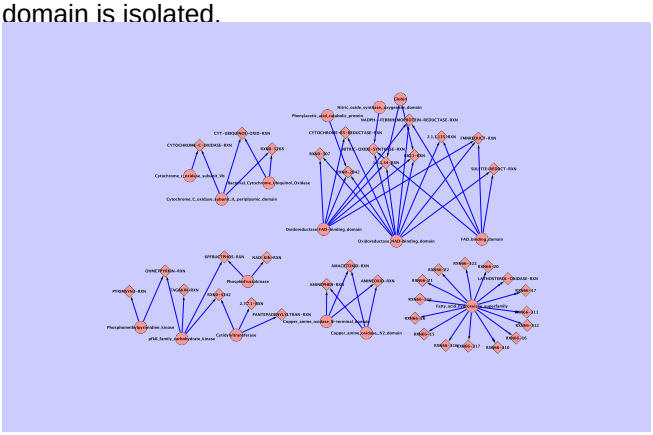


To simplify analysis we decompose G_h into subnetworks called hierarchies. A hierarchy is a tree rooted at a generic (parentless) domain subunit. 662 nontrivial hierarchies with more than one domain subunit were extracted. We show three major hierarchies in Figure 2. The protein kinase domain (PF00069) is part of the conserved catalytic core shared by serine/threonine and tyrosine kinases in eukaryotes (Hanks and Hunter 1995). Three specialized domains – protein tyrosine kinase (PF07714), regulator of G protein signaling domain (PF00615), and calcium/calmodulin dependent protein kinase (PF08332) – appear in non-overlapping subsets of protein kinases respectively.

S1 RNA binding domain (PF00575) occurs in many RNA-associated proteins such as ribosome, RNA polymerase, ribonuclease and polynucleotide phosphorylase (Bycroft et al. 1997). This domain family is involved in diverse functions with distinct partners – ribosome domains, RNA polymerase domains, RNB domain (PF00773), KH domain (PF00013) respectively.

ABC transporter domain (PF00005) is a large family responsible for translocating a variety of compounds across membranes (Higgins 2001). In addition to the generic ABC transporters, many compounds are also transported by specific domains. For example, PF00005 and a domain of

Figure 3. Connection types of domain subunits. Circles: domains. Diamonds: reactions. An edge designates a domain appears in the enzyme of a reaction. From upper left to lower right. (1)Cytochrome C oxidase subunit II, periplasmic domain belongs to a generic domain subunit. (2)Two FAD binding domains are specialized domain subunits. (3)pfkB domain cooperates with three domains in distinct reactions. (4)Two copper amine oxidase domains are equivalent. (5)Fatty acid hydroxylase domain is isolated.



branched-chain amino acid transport system (PF02653) are both involved in the transport of L-valine and D-allose (Adams et al. 1990).

Inclusion relations of domain subunits reflect their functional dependency

We categorize domain subunits into 5 classes according to their intersection relations and illustrate an example in each class in Figure 3.

Generic. A domain subunit is generic if its effect reactions are not subsumed to the effect reactions of any other domain subunits. In other words, it has no parents in G_h. In the upper left graph of Figure 3, the periplasmic domain of cytochrome c oxidase (PF00116) catalyzes electron transfer in both human and *E. coli* (Tsukihara et al. 1996). Subunit Vb of cytochrome c oxidase (PF01215) and bacterial cytochrome ubiquinol oxidase (PF01654) exist only in eukaryotes and bacteria respectively (Tsukihara et al. 1996; Sturr et al. 1996).

Specialized. A domain subunit is specialized if its effect reactions are subsumed to the effect reactions of only one generic domain subunit. In other words, a specialized domain subunit d has one parent or a single equivalent class of parents in G_d. In the top right graph of Figure 3, all the NAD/NADP dependent oxidoreductases contain an oxidoreductase NAD binding domain (PF00175). Many of them also contain FAD binding domains (Dym and Eisenberg 2001). The two FAD binding domains (PF00667 and PF00970) are subsumed to the NAD binding domain.

Cooperative. A domain subunit is cooperative if its effect reactions are also catalyzed by multiple generic domain subunits and their specialized domain subunits. In other words, a cooperative domain subunit d has multiple non-equivalent parents in G_h. Alternatively, a cooperative domain subunit co-catalyzes some reactions with other domain subunits, and these co-catalytic partners are neither ancestors nor descendants of d in G_h. In the lower left graph of Figure 3.3, pfkB family carbohydrate kinase family (PF00294) constitutes a variety of carbohydrate and pyrimidine kinases (Sigrell et al. 1998). It co-catalyzes with three domains in distinct reactions. Each of these domains also catalyzes the reactions not covered by pfkB.

Equivalent. Two domain subunits are equivalent if they have identical neighbors in G_h. In other words, domain subunits in a clique of bi-directional edges in G_h constitute an equivalent class. In the lower middle graph of Figure 3, two domains in copper amine oxidase – N terminal domain (PF07833) and N2 domain (PF02727) – catalyze the oxidative deamination of primary amines (Parsons et al. 1995).

Isolated. A domain subunit is isolated if it does not co-catalyze with other domain subunits in any effect reaction. In other words, an isolated domain subunit is not connected to any other domain subunit in G_h. In the lower right graph of Figure 3, the fatty acid hydroxylase superfamily (PF04116) alone catalyzes the hydrolysis of fatty acids in human (Li and Kaplan 1996).

An inclusion relation such as domain subunit B is subsumed to domain subunit A may imply that A is required for B to take effect. To verify this hypothesis we examine the functions 415 domain subunit pairs with inclusion relations. The majority of the pairs (243 of 415) exhibit asymmetric functional dependency. Table 2 reports selected functionally dependent pairs. Roughly the functional dependency can be categorized into the following types. (1)Domain A manipulates a chemical group and domain B participates in the reactions of specific substrates. For instance, the domains of glutamine amidotransferases transfer an ammonia group from glutamine (van den Heuvel et al. 2002). These domains co-occur with domains involved in glutamate synthesis (Filetici et al. 1996) and asparagine synthesis (Larsen et al. 1999) respectively. (2)Domain A is involved in a generic operation and domain B participates in a specific reaction requiring the generic operation. For instance, the OB-fold nucleic acid binding domain binds to nucleic acids (Ruff et al. 1991). It co-occurs with the domains of tRNA synthetase (Perona et al. 1993) and DNA polymerase III (Aravind et al. 1998) respectively. (3)Domain A carries the essential function of an enzyme (catalytic domain, ligand binding domain, active site) and domain B carries an accessory function (regulatory domain,

Table 2. Functional dependency of selected domains with inclusion relations

Generic domains	Specialized domains
ABC transporter ATP hydrolysis	Branched-chain amino acid transport
PEP utilizing enzyme	Transfer of phosphoryl groups
Aminotransferase class I and II	Regulation of aminotransferase
4Fe-4S binding domain	NADH ubiquinone oxidoreductase
Glutamine amidotransferase	Carbamoyl-phosphate synthase
Protein kinase domain	Regulator of G protein signaling
ATP hydrolysis	K+-transporting ATPase
Acetyltransferase family	Citrate lyase ligase C-terminal
SH3 domain	Signal transduction
Binding to 4Fe-S	Electron transfer from NADH
Alpha amylase, catalytic domain	Alpha amylase C-terminal
S1 RNA binding domain	Ribonuclease, RNA polymerase
B12 binding domain	B12 dependent methionine synthase

dimerisation domain) which may be dispensable in some reactions. For instance, aminotransferase domains transfer an amino group between substrates. They co-occur with domains containing aminotransferase ubiquitination sites that regulate the activities of the enzymes (Gross-Mesilaty et al. 1997), while some enzymes contain the aminotransferase domains but no ubiquitination sites. About one third of the pairs without evident functional dependency (54 of 172) are in the enzymes of protein modification (phosphorylation, acetylation, ubiquitination, etc.). There are diverse families of domains participating in protein modification but do not confer functional dependencies in any of the three types above.

Combinations of domains with no apparent functional dependency synthesize functions of enzymes.

633 of 1883 reactions are catalyzed by multiple domain subunits. It is of interest to know why certain reactions require multiple domains in their enzymes. One obvious explanation is that each domain subunit constitutes a distinct isozyme of the reaction. 72 reactions fall into that category.

An alternative explanation consistent with the modularity hypothesis is that a reaction requires multiple chemical operations and each domain subunit is responsible for a specific operation. Homologous domains can be used in other reactions requiring the same operations. We termed the reactions of this type "synthetic" since they are catalyzed by the synthesis of distinct domain functions. We implemented the following procedures to identify reactions containing domains of distinct functional labels. In metabolism the functional labels of domains are determined by the EC numbers of the reactions they catalyze. To reduce redundant information we only considered the domain subunits with no inclusion relations.

1) For each reaction identify its functional domain subunits according to the criteria described above.
2) Remove the domain subunits with inclusion relations.
3) For each enzyme of the reaction, mark the functional labels of the remaining active domain subunits.
4) If the functional labels of the enzymes of the reaction are not all identical, then mark the reaction synthetic.

Table 3. Selected reactions with synthetic operations

reaction EC	reaction	domain subunit 1
6.3.5.8	glutamine → glutamate	glutamine amidotransferase class
6.1.1.3	tRNA synthesis	tRNA synthetase core domain
2.7.7.8	RNA synthesis	S1 RNA binding domain
2.1.3.2	carbamoyl transfer to aspartate	aspartate carbamoyltransferase
operation 1	domain subunit 2	operation 2
amino group transfer	chorismate binding enzyme	chorismate binding
catalytic domain	TGS domain	regulatory domain
RNA binding	Exoribonuclease family	Nucleic acid cleavage
carbamoyl group transfer	MGS-like domain	regulatory domain

100 reactions are synthetic according to these criteria. Table 3 shows an excerpt of these reactions. Functional synthesis is evident in these reactions. For instance, the transfer of an amino group from glutamine to chorismate in *E. coli* (EC # 6.3.5.8) requires the domains of amidotransferase and chorismate binding.

Pleiotropy of enzymes is achieved by reaction similarity and domain fusion

A reciprocol question of the functional synthesis of reactions is the pleiotropy of enzyme proteins/complexes. The number of reactions catalyzed by each protein/complex again follows a power-law distribution. Despite the majority of enzymes are specific to one reaction, 498 of 2238 *E. coli* proteins/complexes and 372 of 2635 human proteins/complexes catalyze multiple reactions. It is of interest to understand how pleiotropy of these enzymes is achieved.

One obvious explanation is that multiple reactions catalyzed by the same enzyme are similar. For instance, in *E. coli* the ProP osmosensory MFS transporter transports a variety of substrates across the cellular membrane to adapt the change of osmotic pressure (Mileykovskaya 2007). For each pleiotropic enzyme, we checked whether its effect reactions belonged to the same category according to the EC number hierarchies. As expected, 458 of 498 *E. coli* enzymes and 327 of 372 human enzymes catalyze similar reactions.

Despite that reaction similarity is the dominant factor for enzyme pleiotropy, certain enzymes do catalyze heterogeneous reactions. The pleiotropy of some of these enzymes can be explained by domain fusion: The protein/complex is an aggregation of domains responsible for distinct reactions. We developed a method to detect the functional domains responsible for each reaction and identify the pleiotropic enzymes containing distinct functional domains in their effect reactions. The following procedures were implemented.

1) Extract all domain subunits and effect reactions of the enzyme.

Table 4. Selected pleiotropic enzymes containing aggregation of domains of distinct reactions

species	protein
E. coli	aspartate kinase
E. coli	LYSU-CPLX
human	H6PD
human	purine biosynthetic protein
human	uridine 5'-monophosphate synthase
reaction 1	reaction 2
lysine biosyn.	homoserine biosyn.
tRNA synthesis	ATP hydrolysis
pentose phosphate reac 1	pentose phosphate reac 2
purines biosyn.	phosphoribosylglycinamide biosyn.
biosyn. of pyrimidines	synthesis of uridine-5'-monophosphate

2) Build a bi-partite graph G of these domain subunits and effect reactions. Domain subunit A is adjacent to reaction B if A is assigned a functional domain subunit of B.
3) Build an unconnected graph G' of the same nodes as G.
4) Find isolated bicliques of G and add the edges of these bicliques to G'.
5) Identify the nodes in G with one neighbor. Copy the edges in G connecting these nodes to G'.
6) Incrementally copy the maximal edges in G that connect the unconnected nodes in G'.
7) Stop when all connected nodes in G are also connected in G'.

G' is a parsimonious assignment of domains to the effect reactions of an enzyme. The neighbors of each reaction in G' correspond to the functional domains assigned to the reaction. A pleiotropic enzyme is labeled as case 2 if its effect reactions possess distinct sets of functional domains.

18 *E. coli* enzymes and 13 human enzymes exhibit aggregation of functional domains. Table 4 reports excerpts of these enzymes. We present two examples of domain aggregation. In *E. coli*, *metL* encodes a bifunctional enzyme that catalyzes the first step of lysine biosynthesis (ASPARTATEKIN-RXN, 2.7.2.4) and the last step of homoserine biosynthesis. (HOMOSERDEHYDROG-RXN, 1.1.1.3) (Falcoz-Kelly et al. 1969). It contains a domain of amino acid kinase (PF00696) for the first reaction and domains of homoserine dehydrogenase (PF00742, PF03447) for the second reaction. In human, H6PD encodes a bifunctional protein that catalyzes two reactions in the pentose phosphate pathway – 6PGLUCONOLACT-RXN, 3.1.1.31 and GLUCOSE-1-DEHYDROGENASE-RXN, 1.1.1.47 (Beutler and Morrison 1967). It contains a domain of 6-phosphogluconolactonase (PF01182) responsible for the first reaction and domains of glucose-6-phosphate dehydrogenase (PF00479, PF02781) for the second reaction.

Table 5. Counts of domain and reaction conservation between *E. coli* and human

	Conserved reactions	Novel reactions
Conserved domains	237	686
Novel domains	85	757

Requirements for novel reactions, redundancy and pleiotropy are the main driving factors for domain evolution

One billion years of separation between human and *E. coli* substantially alters their metabolic networks and protein domain architectures. We compared the domain compositions and reactions between *E. coli* and human, reported the general properties about their network evolution, and gave possible explanations for the driving forces of their evolution. There are four possible configurations in terms of the conservation of reactions and domains.

- Conserved reaction, conserved domain. Reactions with identical substrates appear in both species. Enzyme proteins/complexes between the conserved reactions of the two species share at least one common domain.
- Novel reaction, novel domain composition. The reaction is specific in *E. coli* or human. There exists at least one *E. coli* or human specific domain in the enzyme.
- Conserved reaction, novel domain. Reactions with identical substrates appear in both species. Some domains appear in the enzymes of only one species.
- Novel reaction, conserved domain. The reaction is specific in *E. coli* or human. The domains in the enzymes for the novel reactions of one species also appear in enzymes of another species.

We counted the number of reactions in each configuration and report them in Table 5. The sum of the four numbers is not equal to the total number of reactions because (1)(3) and (2)(4) are not mutually exclusive.

Conserved reaction, conserved domain. Conserved reactions are likely to retain conserved domains and domain compositions in their enzymes. Among the 287 conserved reactions, 237 (82.5%) have overlapping domains between the two species. 171 (59.6%) have identical compositions in at least one enzyme protein/complex, and 138 (48.4%) have identical domain compositions in all enzyme proteins/complexes.

Novel reaction, novel domain. There are 1404 *E. coli* or human specific reactions. 849 reactions contain enzyme proteins/complexes with novel domains, and the two sets intersect in 757 reactions. A great majority (89.2%) of reactions with novel domains are *E. coli* or human specific, suggesting that the need to catalyze novel reactions is a main driving force for the emergence of novel domains. In contrast, only about half (53.9%) species-specific reactions are catalyzed by novel domains, suggesting the emergence of

Table 6. Top metabolic pathways in E coli and human enriched with novel domains

E. coli:

pathway	# reactions	# novel domains	# novel reactions
chorismate biosyn.	50	22	42
His, Pur, and Pyr biosyn.	52	19	12
lipopolysaccharide biosyn.	24	15	24
peptidoglycan biosyn. I	10	9	0
KDO_2-lipid A biosyn.	16	9	16
threonine metabolism	15	9	8
aspartate biosyn.	25	8	0
central carbon metabolism	24	8	0
flavin biosyn.	9	8	8
respiration (anaerobic)	13	7	3

Human:

pathway	# reactions	# novel domains	# novel reactions
cholesterol biosyn.	50	37	46
nicotine degradation III	18	5	6
phenylalanine degradation I	9	5	6
reductive acetyl coenzyme A	10	4	1
methionine degradation	15	4	5
nucleotides salvage	16	4	1
formylTHF biosyn. I	12	4	0
central carbon metabolism	24	3	0
glycine degradation I	6	3	1

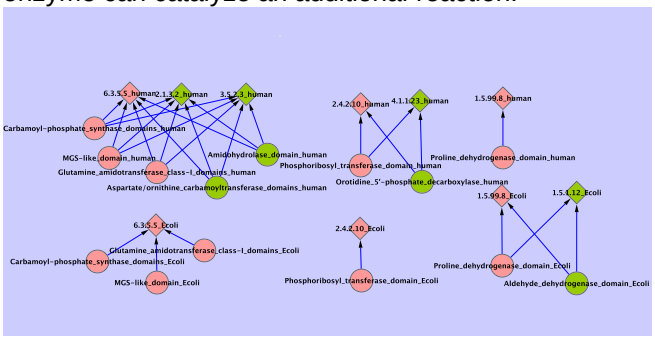
novel domains is not necessary to generate a novel reaction.

We counted the numbers of novel reactions and reactions with novel domains in 270 *E. coli* and 229 human metabolic pathways. The number of novel domains appeared in a pathway is proportional to the number of novel reactions (Pearson correlation coefficients 0.72 for *E. coli* data, 0.93 for human data), confirming the correlation between novel reactions and novel domains. Table 6 lists the top 10 *E. coli* and human pathways enriched with novel domains. In *E. coli* the chorismate superpathway has the highest number of novel reactions and domains. Chorismate is a precursor for many reactants in plants and microbes. Most reactions and domains involved in chorismate metabolism do not exist in human. In human the superpathway of cholesterol biosynthesis has the highest numbers of novel reactions and domains.

Conserved reaction, novel domain. 85 conserved reactions contain domains which appear in the enzymes of only one species. The appearance of novel domains in conserved reactions can be explained by redundancy and pleiotropy. 35 of 85 reactions contain overlapping domains in the enzymes of the two species. In these reactions there are conserved enzymes in the two species. The novel domains either constitute an alternative isozyme or are incorporated to the conserved enzyme(s). Furthermore, the new domains may expand the function of a conserved enzyme to catalyze additional reactions. Alteration of enzymatic functions is revealed by the change of co-catalytic partner reactions between *E. coli* and human. Among the 85 reactions 45 exhibit the change of co-catalytic partners. Overall redundancy and pleiotropy explain 57 of 85 reactions.

Augmentation of species-specific domains to enzymes of conserved reactions is a common way to alter the

Figure 4. Examples of domain incorporation. Circles: domains. Diamonds: reactions. Solid lines: a domain appears in the enzyme of a reaction. Pink nodes: conserved domains and reactions between *E. coli* and human. Green nodes: species specific domains and reactions. Top networks are in human, bottom networks are in *E. coli*. From left to right, (1)Two additional domains in human are incorporated in the enzyme for carbamoyl phosphate synthesis from glutamine, and the enzyme can catalyze two additional reactions. (2)One addition domain in human is incorporated in the enzyme for orotate phosphorylation, and the enzyme can catalyze an addition reaction. (3)One additional domain in *E. coli* is added to the enzyme for proline degradation and the enzyme can catalyze an additional reaction.



metabolic network. Three examples are illustrated in Figure 4. (1)6 domains of carbamoyl phosphate synthetase and glutamine amidotransferase appear in the enzymes of carbamoyl phosphate synthesis (CARBPSYN-RXN, EC # 6.3.5.5) in both species (Guillou et al. 1992). In human three additional domains of aspartate/ornithine carbamoyltransferase are incorporated in the enzyme, and they catalyze carbmoyl-aspartate synthesis (EC # 3.5.2.3 and 2.1.3.2). These three reactions constitute the first three steps of pyrimidine biosynthesis (Chen et al. 1989). (2)The phosphoribosyl transferase domain appears in the enzymes of phosphryl group transfer to orotidine-5'-phosphate (OROPRIBTRANS-RXN, EC # 2.4.2.10) in both species. In human the phosphoribosyl transferase incorporates an orotidine 5'-phosphate decarboxylase domain, and also catalyzes carboxylation of orotidine-5'-phosphate (OROTPDECARB-RXN, 4.1.1.23). The two reactions constitute the last two steps of de novo pyrimidine biosynthesis (Suchi et al. 1997). (3)The proline dehydrogenase domain appears in the enzymes of proline reduction (RXN-821, EC # 1.5.99.8) in both species. In *E. coli* the enzyme incorporates an additional domain of aldehyde dehydrogenase, and catalyzes conversion of pyrroline 5-carboxylate to glutamate (PYRROLINECARBDEHYDROG-RXN, EC # 1.5.1.12) (Ling et al. 1994).

Novel reaction, conserved domain. 686 of 1404 species-specific reactions contain domains utilized in other reactions of another species. A likely explanation is that these "novel reactions" are similar to some conserved reactions and therefore can be catalyzed by conserved domains. Indeed, 591 of 686 reactions (86.2%) have the same EC labels or share the same substrates with reactions in another species.

Conclusions

In this work we demonstrate that the combinatorial interactions of metabolic enzyme protein domains follow certain general rules. First, the effect reactions of many domain subunits have nested inclusion relations. We have shown that the majority of domains with inclusion relations also exhibit asymmetric functional dependency. A domain subunit whose effect reactions cover those of another domain subunit typically carries generic operations such as transferring an amino group or nucleic acid binding. In contrast, a domain subunit whose effect reactions are subsumed to those of another domain subunit often carries specialized or accessory functions such as interactions with a specific substrate or regulation of enzyme activities.

Second, about one third of reactions in human or *E. coli* are catalyzed by multiple domain subunits. Some of these reactions need multiple domain subunits because their operational requirements are the synthesis of the functions provided by each domain subunit. For instance, a reaction may require one domain to transfer amino groups and another domain to bind to a specific substrate. The combinations of these domain subunits can in principle create great complexity of the metabolic network. However, since the number of partners is small for most domain subunits, the actual complexity of the network is restricted.

Third, many enzyme proteins/complexes are able to catalyze multiple reactions. In most cases pleiotropy of enzymes results from the similarity of the effect reactions. For instance, a transporter protein can transport multiple substrates across the cellular membrane. However, in some enzymes pleiotropy is achieved by merging domains with distinct functions. Many of these merged enzymes have a selective advantage for catalyzing reactions in the same metabolic pathway (e.g., Chen et al. 1989; Suchi et al. 1997).

Comparison of domain compositions and metabolic reactions between human and *E. coli* provides insights regarding the evolution of the metabolic network. *E. coli* and human share a substantial number of common reactions. Over 80% of conserved reactions retain either the entire domain compositions or overlapping domains between the two species, suggesting some part of the metabolic network is conserved even between the two distant species. Nearly 90% of the reactions containing novel domains are specific to either human or *E. coli*, suggesting that the need to catalyze novel reactions is a major driving force to create novel domains.

A considerable number of conserved reactions are catalyzed by enzymes with species-specific domains. Two possible assumptions may explain why novel domains are included in the enzymes of conserved reactions. A novel domain may be added to a conserved enzyme to improve its catalytic function in human or *E. coli*. Alternatively, it may expand the function of the enzyme to catalyze other reactions. About two thirds of the conserved reactions with novel domains are consistent with these hypotheses. Complete replacement of domains (non-orthologous displacement, Chothia et al. 2003) occurs in the remaining reactions.

Many species-specific reactions are catalyzed by enzymes with conserved domains. Indeed, the majority of these reactions are not "novel", as they have the same EC labels or share the same substrates with the conserved reactions.

The methodological framework and data in this study have several limitations. The completeness and accuracy of data are one major concern. The current version of Biocyc has many "holes" in reactions where information about their catalytic enzymes is missing. As the two species in this analysis are among the most well studied model organisms, problems of data sparsity will be more prominent as comparative analysis extends to other species. Also, demarcation of domains and domain families in Pfam may not precisely match the structural and evolutionary properties of proteins. Some domain families contain members with several distinct functions, and some domain families may have evolutionary relations. The current version of Pfam groups evolutionarily related domain families into clans. Incorporation of clan information and comparison with other domain databases (e.g., ProDom and SCOP) will be a useful extension of the current work.

Our analysis does not tackle the sequence substitution among the domain family members. Sequence substitution is a major evolutionary mechanism between closely related species. Incorporation of sequence substitution models between the sites of the same or distinct domains will be a useful extension. We do not incorporate the information of allosteric and transcriptional regulation of enzymes in this analysis, either. It would be of great interest to study the evolution of the (allosteric and transcriptional) regulatory networks of the metabolic system. However, current knowledge about regulation may not be sufficient for a comprehensive analysis like the metabolic network.

References

[1] Carroll SB. 2005. Evolution at two levels: on genes and form. PLoS Biology 3(7):1159-1166.

[2] House AE, Lynch KW. 2008. Regulation of alternative splicing: more than just the ABCs. Journal of Biological Chemistry 283(3):1217-1221.

[3] Ben-Dov C, Hartmann B, Lundgren J, Valcrcel J. 2008. Genome-wide analysis of alternative pre-mRNA splicing. Journal of Biological Chemistry 283(3):1229-1233.

[4] Apic G, Gough J, and Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. Journal of Molecular Biology 310:311-325.

[5] Chothia C, Gough J, Vogel C, and Tecichmann SA. 2003. Evolution of the protein repertoire. Science 300: 1701-1703.

[6] Wuchty S. 2001. Scale-free behavior in protein domain networks. Molecular Biology and Evolution 18(9):1694-1702.

[7] Ravasz E, Somera AL, Mongru DA, Oltvai Z, Barabasi AL. 2002. Hierarchical organization of modularity in metabolic networks. Science 297:1551-1555.

[8] Vogel C, Berzuini C, Bashton M, Gough J, and Teichmann SA. 2004. Supra-domains: evolutionary units larger than single protein domains. Journal of Molecular Biology 336:809-823.

[9] Vogel C, Teichmann SA and Pereira-Leal J. 2005. The relationship between domain duplication and recombination. Journal of Molecular Biology 346:355-365.

[10] Schmidt EE and Davies CJ. 2007. The origins of polypeptide domains. BioEssays 29:262-270.

[11] Kaessmann H, Zollner S, Nekrutenko A, and Li WH. 2007. Signatures of domain shuffling in the human genome. Genome Research 12:1642-1650.

[12] Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, and Lopez-Bigas N. 2005. BioCyc: Expansion of the BioCyc collection of pathway/genome databases to 160 genomes, Nucleic Acids Research 19:6083-89 2005.

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

[14] Duarte N, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, and Palsson BO. 2007. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of National Academy of Science, USA, 104(6):1777-1782.

[15] Pruitt KD, Tatusova, T, Maglott DR. 2007. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 35(Database issue):D61-65.

[16] Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R. 2004. UniProt Archive. Bioinformatics 20:3236-3237.

[17] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffith-Jones S, Howe KL, Marshall M, Sonnhammer ELL. 2002. The Pfam protein families database. Nucleic Acids Res., 30:276-280. http://www.sanger.ac.uk/Software/Pfam/

[18] Corpet F, Servant F, Gouzy J, Kahn D. 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Research 28:267-269.

[19] Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 247:536-540.

[20] Dandekar T, Schuster S, Snel B, Huynen M, Bork P. 1999. Pathway alignment: application to the comparative analysis of glycolytic enzymes. Biochemistry Journal 343:115-124.

[21] Pal C, Papp B, and Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nature Genetics 37(12):1372-1375.

[22] Bowers P, Cokus SJ, Eisenberg D, and Yeates TO. 2004. Use of logic relationships to decipher protein network organization. Science 306:2246-2249.

[23] Kimiyoshi I, Yoshihiro A, Kumi N, Shinsei M, Tatsuo H, Osamu S, Nobuyoshi S, and Takeshi N. 1993. Cloning of the cDNA encoding human xanthine dehydrogenase (oxidase): Structural analysis of the protein and chromosomal location of the gene. Gene 133(2):279-284.

[24] Xi H, Schneider BL, and Reitzer L. 2000. Purine catabolism in *Escherichia coli* and function of xanthine dehydrogenase in purine salvage. Journal of Bacteriology 182(19):5332-5341.

[25] Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S. 1996. The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. Science 272(5265):1136-1144.

[26] Sturr MG, Krulwich TA, Hicks DB. 1996. Purification of a cytochrome bd terminal oxidase encoded by the Escherichia coli app locus from a delta cyo delta cyd strain complemented by genes from Bacillus firmus OF4. Journal of Bacteriology 178(6):1742-1749.

[27] Dym O and Eisenberg D. 2001. Sequence-structure analysis of FAD-containing proteins. Protein Science 10(9):1712-1728.

[28] Sigrell JA, Cameron AD, Jones TA, Mowbray SL. 1998. Structure of Escherichia coli ribokinase in complex with ribose and dinucleotide determined to 1.8 A resolution: insights into a new family of kinase structures. Structure 6(2):183-193.

[29] Parsons MR, Convery MA, Wilmot CM, Yadav KD, Blakeley V, Corner AS, Phillips SE, McPherson MJ, Knowles PF. 1995. Crystal structure of a quinoenzyme: copper amine oxidase of Escherichia coli at 2 A resolution. Structure 3:1171-1184.

[30] Li L and Kaplan J. 1996. Characterization of yeast methyl sterol oxidase (ERG25) and identification of a human homologue. Journal of Biological Chemistry 271(28):16927-16933.

[31] van den Heuvel RH, Ferrari D, Bossi RT, Ravasio S, Curti B, Vanoni MA, Florencio FJ, Mattevi A. 2002. Structural studies on the synchronization of catalytic centers in glutamate synthase. Journal of Biological Chemistry 277:24579-24583.

[32] Filetici P, Martegani MP, Valenzuela L, Gonzalez A, Ballario P. 1996. Sequence of the GLT1 gene from Saccharomyces cerevisiae reveals the domain structure of yeast glutamate synthase. Yeast 12:1359-1366.

[33] Larsen TM, Boehlein SK, Schuster SM, Richards NG, Thoden JB, Holden HM, Rayment I. 1999. Larsen TM, Boehlein SK, Schuster SM, Richards NG, Thoden JB, Holden HM, Rayment I. Biochemistry 38(49):16146-16157.

[34] Ruff M, Krishnaswamy S, Boeglin M, Poterszman A, Mitschler A, Podjarny A, Rees B, Thierry JC, Moras D. 1991. Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp). Science 252:1682-1689.

[35] Perona JJ, Rould MA, Steitz TA. Structural basis for transfer RNA aminoacylation by Escherichia coli glutaminyl-tRNA synthetase. 1993. Biochemistry 32(34):8758-8771.

[36] Aravind L, Koonin EV. Phosphoesterase domains associated with DNA polymerases of diverse origins. 1998. Nucleic Acids Research 26:3746-3752.

[37] Gross-Mesilaty S, Hargrove JL, Ciechanover A. 1997. Degradation of tyrosine aminotransferase (TAT) via the ubiquitin-proteasome pathway. FEBS Lett 405:175-180.

[38] Wagner A. 2001. How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. Bioinformatics 17:1183-1197.

[39] Markowetz F, Bloch J, Spang R. 2005. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. Bioinformatics 21: 4026-4032.

[40] Hanks SK, Hunter T. 1995. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB Journal 9(8):576-596.

[41] Bycroft M, Hubbard TJ, Proctor M, Freund SM, Murzin AG. 1997. The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. Cell 88(2):235-242.

[42] Higgins CF. 2001. ABC transporters: physiology, structure and mechanism–an overview. Research in Microbiology 152(3-4):205-210.

[43] Adams MD, Wagner LM, Graddis TJ, Landick R, Antonucci TK, Gibson AL, Oxender DL. 1990. Nucleotide sequence and genetic characterization reveal six essential genes for the LIV-I and LS transport systems of Escherichia coli. Journal of Biological Chemistry 265(20):11436-11443.

[44] Mileykovskaya E. 2007. Subcellular localization of Escherichia coli osmosensory transporter ProP: focus on cardiolipin membrane domains. Molecular Microbiology 64(6):1419-1422.

[45] Falcoz-Kelly F, van Rapenbusch R, Cohen GN. 1969. The methionine-repressible homoserine dehydrogenase and aspartokinase activities of Escherichia coli K 12. Preparation of the homogeneous protein catalyzing the two activities. Molecular weight of the native enzyme and of its subunits. European Journal of Biochemistry 8(1):146-152.

[46] Beutler E, Morrison M. 1967. Localization and characteristics of hexose 6-phosphate dehydrogenase (glucose dehydrogenase). Journal of Biological Chemistry. 242(22):5289-5293.

[47] Guillou F, Liao M, Garcia-Espana A, Lusty CJ. 1992. Mutational analysis of carbamyl phosphate synthetase. Substitution of Glu841 leads to loss of functional coupling between the two catalytic domains of the synthetase subunit. Biochemistry 31(6):1656-1664.

[48] Chen KC, Vannais DB, Jones C, Patterson D, Davidson JN. 1989. Mapping of the gene encoding the multifunctional protein carrying out the first three steps of pyrimidine biosynthesis to human chromosome 2. Human Genetics 82(1):40-44.

[49] Suchi M, Mizuno H, Kawai Y, Tsuboi T, Sumi S, Okajima K, Hodgson ME, Ogawa H, Wada Y. 1997. Molecular cloning of the human UMP synthase gene and characterization of point mutations in two hereditary orotic aciduria families. American Journal of Human Genetics. 60(3):525-539.

[50] Ling M, Allen SW, Wood JM. 1994. Sequence analysis identifies the proline dehydrogenase and delta 1-pyrroline-5-carboxylate dehydrogenase domains of the multifunctional Escherichia coli PutA protein. Journal of Molecular Biology 243(5):950-956.