Evolution of domain compositions in the metabolic networks of human and *Escherichia coli*

C.H. Yeang[1], **N. Baas**[2]
[1]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, R.O.C.
[2]Dept. of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

Abstract— *It is widely recognized that complexity of metabolic networks arises from duplication, recruitment and recombination of enzyme protein domains. However, variations of the domain evolution mechanisms among different organisms and metabolic subsystems are not well known. In this work we demonstrate strong heterogeneity of domain evolution mechanisms by comparing the domain compositions of the metabolic networks in human and Escherichia coli. While addition of novel domains and duplication of domains from enzymes catalyzing other reactions dominate the evolution in both species, incorporation of novel functions and domain duplication within the same reactions are much more frequent in human. Furthermore, different metabolic processes are driven by distinct evolution mechanisms. Energy utilization and biosynthesis of nucleotides, amino acids, and lipids retain more conserved domains and have more intra-reaction domain duplication events in human. In contrast, many alternative pathways of amino acid degradation are formed by incorporating novel functions to conserved enzymes. Furthermore, addition of novel domains occurs in signaling pathways of human and biosynthesis of species-specific compounds such as riboflavin (vitamin B2) and chorismate in E. coli and cholesterol in human. Our analysis provides insight about the functional constraints of domain evolution in metabolic systems.*

Keywords: protein domain, metabolic network, evolution, bipartite graph.

Introduction

Understanding the evolution of biomolecular systems is a central question in contemporary biology. Among various systems the evolution of metabolic networks is likely to be deciphered soon as great amount of information across many species is already available (e.g., Biocyc: Karp et al. 2005, KEGG: Kanehisa and Goto 2000, Recon 1: Duarte et al. 2007). Comparative studies suggest the metabolic networks of all species share a conserved central core (Morowitz 1999, Peegin-Alves et al. 2003). Yet how the highly specialized and complex metabolic system of an organism arises from a small set of core reactions remains unclear.

One aspect of studying network evolution is to find the principles governing the formation of specific network topologies. Various theoretical models are proposed to explain the emergence of the "scale-free" topologies of metabolic networks (e.g., Jeong et al. 2000, Barabasi et al. 1999, Carlson and Doyle 2002). Despite the fundamental insight, those theoretical models provide little information regarding the mechanisms of network evolution. On the other hand, a great number of researchers have investigated various mechanisms driving the evolution of metabolic networks, such as sequence substitution (e.g., Enattah et al. 2002), gene duplication (e.g., Ohno 1970), gene rearrangements (e.g., Vogel et al. 2004), fusion and fission (e.g., Fani et al. 2007), and horizontal gene transfer (e.g., Pal et al. 2005).

Proteins and genes may not be adequate subunits to study metabolic network evolution since many changes are involved in parts of the proteins instead of their entirety. Domains are polypeptide subunits of proteins that constitute similar molecular structures or sequences. Since many metabolic reactions share common substrates or chemical operations, it is sensible to expand and rewire the metabolic networks by duplicating domains and creating different combinations to form novel enzymes. Indeed, recruitment of domains from enzymes of different parts of the metabolic network is prevalent (Teichmann et al. 2001).

Domain evolution in metabolic networks has been investigated in a considerable number of previous studies. For instance, Chothia et al. showed the majority of protein domains appeared before the prokaryote-eukaryote split (Chothia et al. 2003). Teichmann et al. indicated frequent turnover of substrate binding domains and stability of catalytic domains in enzymes (Teichmann et al. 2001). Vogel et al. demonstrated the conservation of domain architectures in specific orders (Vogel et al. 2004). Caetano-Anolles et al. and Fukami-Kobayashi et al. inferred the phylogenetic relations of domain architectures (Caetano-Anolles et al. 2003, Fukami-Kobayashi et al. 2007). Caetano-Anolles et al. extended the analysis to infer the phylogeny of metabolic pathways (Caetano-Anolles et al. 2007).

Despite intensive studies in phylogeny and evolution mechanisms of domain compositions in metabolic networks, two questions are not yet actively pursued: do the mechanisms of domain evolution vary with organisms and specific metabolic processes? In this work, we provide positive evidence to answer both questions by comparing the domain

compositions of the metabolic networks in human and *Escherichia coli*. Our study demonstrates strong heterogeneity of the domain evolution mechanisms between human and *E. coli* and among various metabolic functional classes. Addition of novel domains and duplication of domains from enzymes of other reactions are common in both species. However, incorporation of novel functions for conserved enzymes and intra-reaction domain duplication occur much more frequently in human. Furthermore, different metabolic functional classes are driven by distinct domain evolution mechanisms.

The rest of the paper is organized as follows. We first describe an algorithm of reconstructing the consensus domain-metabolic network of *E. coli* and human, and define five types of mechanisms for domain-metabolic network evolution. We then compare the domain evolution mechanisms between *E. coli* and human and identify the distinct species-specific mechanisms. We also compare the differences of the domain evolution mechanisms among reactions of 15 functional classes, and relate these mechanisms with properties of the metabolic processes. We then list metabolic pathways enriched with each type of domain evolution mechanism and give plausible explanations for the occurrence of these domain changes. Finally we discuss the implications of domain evolution in shaping metabolic networks, extension and limitation of our method.

Reconstruction of domain-metabolic network evolution

To investigate the mechanisms of domain-metabolic network evolution we have to reconstruct the domain-metabolic networks of ancestral species from the data of contemporary species. In this work we use the data of human and *E. coli* for the poor annotations of reaction-enzyme associations in other organisms. About 13% and 25% *E. coli* and human reactions in Biocyc are not annotated with enzymes, while in *Saccharomyces cerevisiae* the fraction of unannotated reactions jumps to near 50%. Thus the results obtained from the current datasets of multiple species can be misleading. Nevertheless, the method can be extended to multiple species when more complete information becomes available.

We download human and *E. coli* subsets of the Biocyc database (Karp et al. 2005) as the version of October 2008. The database contains the substrates and enzymes of reactions and the metabolic pathways they belong to. 1620 reactions, 2862 enzyme proteins/complexes and 332 pathways from human and 1714 reactions, 2311 enzymes and 289 pathways from *E. coli* are extracted.

Domain architectures of human and *E. coli* enzyme proteins are extracted from the Pfam database (Bateman et al. 2002) as the version of August 2007. 5122 domain families appear in *E. coli* or human. We count the occurrence of each domain family in the proteome of each species. For simplicity the domain architecture of a protein is reduced to a "bag of domains" representation: we discard the order of domains in a protein and treat multiple occurrences of the same domain identical. 2838 human enzymes and 1658 *E. coli* enzymes constitute known domain architectures, and they catalyze 1201 and 1481 reactions respectively.

A domain-metabolic network $G = (V_D \cup V_R, E)$ is a bipartite graph of domain nodes V_D and reaction nodes V_R. An edge $e = (d, r) \in E$ denotes that domain d appears in the enzyme(s) catalyzing reaction r. Denote $A_D : V_D \rightarrow F_D \times P_D$ the annotation of Pfam family (F_D) and protein/complex (P_D) for each domain, and $A_R : V_R \rightarrow S_R$ the annotation of reactants for each reaction. A parsimonious reconstruction of two domain-metabolic networks is the consensus of the two networks and the proper mapping from the consensus to each domain-metabolic network. We adopt the algorithm in Figure 1 to find the consensus network. Briefly, it finds the identical reactions in the two species and the one-to-one correspondence between enzyme proteins/complexes on the identical reactions. The consensus domains of a consensus reaction are the common domains in the corresponding proteins/complexes. We identify the mechanisms of domain-metabolic network evolution by comparing the domain compositions between the consensus network and the network of human or *E. coli*. These mechanisms are categorized into five types and are illustrated in Figure 2.

1) Domain conservation. A reaction $r'_1 \in V_{R2}$ in species 2 contains conserved domains if it can be mapped from a consensus reaction $r_c = (r_1, r'_1) \in V_{Rc}$, and the domain neighbors of r'_1 in G_2 and those of r_c in G_c have a non-empty intersection. In other words, there exists conserved domains in the enzymes catalyzing the consensus reaction of both species. In contrast, a consensus reaction undergoes domain displacement if its catalyzing domains in the two species have no intersection (Chothia et al. 2003).
2) Incorporation of novel domain functions (abbreviated as incorporation). A reaction $r'_2 \in V_{R2}$ in species 2 undergoes incorporation of novel domain functions if there exists a domain $d'_1 \in V_{D2}$ which can be mapped from a consensus domain $d_c \in V_{Dc}$. (d'_1, r'_2) is an edge in G_2, but either r'_2 is not a consensus reaction or d_c is not adjacent to the consensus reaction of r'_2 in G_c. In other words, there exists homologous proteins/complexes catalyzing some consensus reactions of the two species, but the association of the homologous proteins/complexes with r'_2 is novel in species 2.
3) Intra-reaction domain duplication. A reaction $r'_1 \in V_{R2}$ in species 2 undergoes intra-reaction domain duplication if it can be mapped from a consensus reaction $r_c = (r_1, r'_1) \in V_{Rc}$, and multiple domain neighbors of r'_1 in G_2 are duplicated from one domain neighbor of r_c in G_c. In other words, multiple catalyzing domains

Fig. 1: Algorithm of finding a consensus domain-metabolic network

Inputs: Domain-metabolic networks $G_1 = (V_{D1} \cup V_{R1}, E_1), G_2 = (V_{D2} \cup V_{R2}, E_2)$ and the annotations of their nodes $A_{D1}, A_{D2}, A_{R1}, A_{R2}$.

Outputs: A consensus network $G_c = (V_{Dc} \cup V_{Rc}, E_c)$, mappings $M_{D1} : V_{Dc} \rightarrow V_{D1}, M_{D2} : V_{Dc} \rightarrow V_{D2}$ from the consensus domains to the domains in species 1 and 2 respectively, mappings $M_{R1} : V_{Rc} \rightarrow V_{R1}, M_{R2} : V_{Rc} \rightarrow V_{R2}$ from the consensus reactions to the reactions in species 1 and 2 respectively, annotation A_{Dc} of consensus domains.

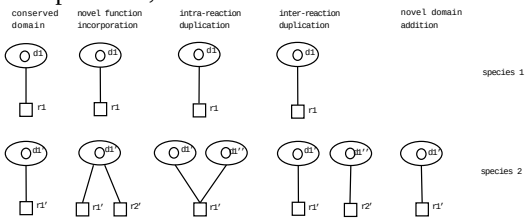
1) Find the set of reaction pairs $RP = \{(r_1, r_2) \in V_{R1} \times V_{R2} : A_{R1}(r_1) = A_{R2}(r_2)\}$. Each r_1 and r_2 constitute identical reactants. Construct the consensus reactions $V_{Rc} = RP$ and the corresponding mappings $M_{R1} : RP \rightarrow V_{R1}$ and $M_{R2} : RP \rightarrow V_{R2}$.
2) For each consensus reaction $(r_1, r_2) \in V_{Rc}$ where r_1 and r_2 are adjacent to some domains in G_1 and G_2, find the consensus domains $CD(r_1, r_2)$:
   a) Find the enzyme proteins/complexes P_1 catalyzing r_1. For each protein/complex $p_1 \in P_1$, find the domain composition $DC(p_1)$.
   b) Repeat the same step to find catalyzing proteins/complexes P_2 and their domain compositions $DC(p_2)$ for r_2.
   c) Find likely correspondence between elements in P_1 and P_2. Construct a bi-partite graph G_{P12} between P_1 and P_2. For each $p_1 \in P_1$, find the protein(s) $p_2 \in P_2$ such that $|DC(p_1) \cap DC(p_2)|$ is maximal among all the proteins in P_2. Connect p_1 to each such protein p_2 in G_{P12}. Repeat the same procedure to find likely corresponding proteins for each $p_2 \in P_2$, and connect these proteins to p_2 in G_{P12}.
   d) Apply the augmenting path algorithm (Cormen et al. 2001) on G_{P12} to find the optimal matching $MP = \{(p_1, p_2) \in P_1 \times P_2\}$ between P_1 and P_2.
   e) The consensus domains $CD(r_1, r_2)$ is the union of the consensus domains for each $(p_1, p_2) \in MP$.
   f) Annotate the domain family and protein/complex in species 1 and 2 for each consensus domain.

of r'_2 are duplicated from one catalyzing domain of the consensus reaction.

4) Inter-reaction domain duplication. A reaction $r'_2 \in V_{R2}$ in species 2 undergoes inter-reaction domain duplication if it does not undergo intra-reaction domain duplication, and some adjacent domain of r'_2 in G_2 is duplicated from a consensus domain $d_c \in V_{Dc}$ of other consensus reactions.
5) Novel domain addition. A reaction $r'_1 \in V_{R2}$ in species 2 undergoes novel domain addition if it is catalyzed by some domain d''_1 that only appears in species 2.

Notice these mechanisms are not mutually exclusive. Domain conservation and intra-reaction duplication occur only in conserved reactions, whereas other mechanisms can occur in both conserved and novel reactions.

Fig. 2: Mechanisms of domain evolution. Small circles: domains. Ellipses: enzymes. Squares: reactions. An edge denotes an enzyme catalyzes a reaction. d_1 and d'_1 are orthologous domains in species 1 and 2. d''_1 is a paralogous domain of d'_1. r_1 and r'_1 are identical reactions in species 1 and 2. r'_2 is another reaction in species 2. Each column represents a mechanism of domain evolution. From left to right: domain conservation, incorporation of novel domain functions, intra-reaction domain duplication, inter-reaction domain duplication, novel domain addition.



Evolution of human and *E. coli* metabolic networks is driven by distinct mechanisms

One critical question about the evolution of metabolic networks is whether various domain evolution mechanisms are homogeneous or species specific. To answer this question we count the number of human and *E. coli* reactions with each type of mechanism and report them in Table 1. The numbers demonstrate strong heterogeneity and suggest evolution of the metabolic network is driven by distinct mechanisms along the lineages of prokaryotes and eukaryotes.

Table 1: Domain evolution mechanisms in *E. coli* and human.

species	conserved	incorporation	intra duplication	inter duplication	novel
E. coli	281	54	20	284	554
Human	281	110	88	211	546

The number of reactions with conserved domains between the two species are identical by definition. Human has about twice the number of reactions with domain function incorporation compared to *E. coli* (110 vs 54). Incorporation has two possible implications. An enzyme of a conserved reaction may acquire an additional function to catalyze novel reactions. Multiple enzymes of distinct reactions in one species may be fused into one protein/complex in another species. Pleiotropy and consolidation of human enzymes caused by incorporation partially explain the previous observation that the fraction of enzyme numbers in the eukaryote proteome is lower than that of the prokaryotes (Freilich et al. 2005).

The number of human reactions undergoing intra-reaction domain duplication is about four times as *E. coli* reactions (88 vs 20). The numbers confirm a more prevalent redundancy in the eukaryotes metabolic network (Freilich et al. 2005) and the hypothesis that gene duplication is a dominant molecular mechanism for eukaryotic evolution (Ohno 1970).

In contrast, the *E. coli* metabolic network encounters more inter-reaction domain duplication than the human counterpart (284 vs 211). This mechanism recruits domains duplicated from existing enzymes to form new enzymes and to catalyze novel or conserved reactions. The large numbers of reactions undergoing inter-reaction domain duplication in both species suggest it is a dominant mechanism driving the network evolution. Domain recruitment models of metabolic network evolution have been proposed (Schmidt et al. 2003, Teichmann et al. 2001). However, our results indicate that domain duplication plays equally important roles in both *E. coli* and human but is utilized differently in the two species. In human, domain duplication both expands the size and connectivity of the metabolic network and adds redundancy of isozymes. In *E. coli*, domain duplication/recruitment is mainly used to expand the network.

Novel domain additions are the most common changes among the reactions of human and *E. coli* (546 vs 554). Both *E. coli* and human rely on novel domains to alter the metabolic networks. Many reactions involved in protein modification or the synthesis of species-specific compounds are catalyzed by novel domains. Detailed analysis of some of those reactions will appear in the subsequent sections.

Evolution of the subnetworks of different metabolic processes is driven by distinct mechanisms

Another important question is whether the evolution of the networks of different metabolic processes is driven by distinct mechanisms. We assign 15 labels to reactions according to their functional classes in Biocyc and count the occurrences of each type of mechanism in the reactions of each functional class. 882 *E. coli* reactions and 945 human reactions are labeled accordingly. Only a small fraction of reactions (61 *E. coli* reactions and 74 human reactions) are assigned to multiple labels. Table 2 shows the counts of each type domain evolution mechanism in each functional class.

Table 2: Domain evolution mechanisms in each functional class

functional class	total	conserved	incorporation
cofactor biosynthesis	172	29	11
signaling pathways	146	4	9
nucleotide biosynthesis	141	61	22
amino acid biosynthesis	127	28	9
lipid biosynthesis	121	16	12
amino acid degradation	107	21	16
carbohydrates degradation	60	6	3
energy metabolism	59	29	3
carbohydrates biosynthesis	53	14	2
protein biosynthesis	46	20	2
other degradation	45	6	3
secondary metabolite degrad.	37	1	0
amines degradation	31	4	0
lipid degradation	20	8	4
nucleotide degradation	18	7	1

functional class	intra duplication	inter duplication	novel
cofactor biosynthesis	6	49	63
signaling pathways	2	27	79
nucleotide biosynthesis	20	27	54
amino acid biosynthesis	13	35	58
lipid biosynthesis	11	22	68
amino acid degradation	4	39	31
carbohydrates degradation	2	18	20
energy metabolism	12	16	26
carbohydrates biosynthesis	3	9	17
protein biosynthesis	5	7	16
other degradation	3	5	29
secondary metabolite degrad.	0	14	13
amines degradation	0	15	6
lipid degradation	6	8	5
nucleotide degradation	0	7	5

Among various processes nucleotide biosynthesis, energy metabolism and protein biosynthesis contain the highest fractions of reactions with conserved domains (61/141, 29/59 and 20/46). These reactions constitute a highly conserved "central core" of metabolism (Morowitz 1999, Peegin-Alves et al. 2003). In contrast, reactions involved in signaling pathways and the degradation of carbohydrates and other compounds have fewer reactions with conserved domains. Intriguingly, degradation reactions tend to be less conserved than biosynthesis reactions.

Nucleotide biosynthesis, amino acid degradation, lipid and cofactor biosynthesis encounter more incorporation than other processes (22, 16, 12 and 11 reactions respectively). One consequence of functional incorporation in amino acid degradation is the emergence of alternative degradation pathways of identical or similar amino acids (see the analysis below). Several domain fusion events occur in the reactions of nucleotide and lipid biosynthesis (see the analysis below). Intriguingly, reactions involved in the carbon metabolism (energy metabolism, carbohydrates degradation and biosynthesis) have very few domain incorporation events (3, 3, and 2 reactions) despite the large size of each class.

Nucleotide and amino acid biosynthesis, energy metabolism and lipid biosynthesis encounter more intra-reaction domain duplication (20, 13, 12 and 11

reactions respectively). A considerable number of human reactions involved in these processes are catalyzed by multiple isozymes sharing common domain compositions. More biosynthesis reactions encounter intra-reaction domain duplication than degradation reactions. A plausible hypothesis is that energy metabolism and biosynthesis of essential compounds occur in every cell, whereas degradation of many compounds is undertaken by specialized tissues in human. The ubiquitous but differential demand for energy metabolism and biosynthesis may drive the evolution of tissue-specific isozymes.

Only two reactions of signaling pathways undergo intra-reaction domain duplication. This number is seriously underestimated because Biocyc collapses all protein phosphorylations into one reaction.

Most major functional classes are enriched with both inter-reaction domain duplication and novel domain addition. Novel domain addition is more prevalent than inter-reaction domain duplication in most functional classes. Specifically, in signaling pathways and lipid biosynthesis novel domain addition dominates domain evolution (79 and 68 reactions). Carbohydrates and protein biosynthesis and degradation of miscellaneous types of metabolites have much fewer inter-reaction domain duplication compared to novel domain addition (9 vs 17, 7 vs 16 and 5 vs 29). In contrast, reactions of amino acid and amines degradation undergo more domain duplication than novel domain addition (39 vs 31 and 15 vs 6). Moreover, the majority of the domains are duplicated from the enzymes in different functional classes (data not shown). This observation confirms the heterogeneous origins of domain duplication (Schmidt et al. 2003, Teichmann et al. 2001).

Specific cases of domain-metabolic network evolution

Tables 1 and 2 provide a general overview regarding the evolution of domain compositions of the entire metabolic network. To grasp more specific information of the domain-metabolic network evolution we identify the pathways enriched with each type of the evolution mechanism and scrutinize selected reactions undergoing these changes.

Table 3 lists selected Biocyc pathways of *E. coli* and human enriched with each type of the evolution mechanism. An enriched pathway is discarded if a superpathway containing it is also enriched with the same evolution mechanism.

Superpathways of the central carbon energy metabolism (glycolysis, pyruvate dehydrogenase, TCA and glyoxylate bypass) and histidine, purine and pyrimidine biosynthesis, as well as the pathway of fatty acid β oxidation I are highly enriched with conserved domains in both species. For instance, 19 of 24 reactions in the central carbon metabolism contain conserved domains (hyper-geometric p-value $< 10^{-8}$ for both species). The stability of these ancient core reactions imposes a strong selective pressure on domain compositions. In addition, pathways of gluconeogenesis and colanic acid building blocks biosynthesis are also enriched with conserved domains in both species.

Table 3: Selected pathways enriched with domain evolution mechanisms. N_0: # reactions. N_1: # reactions with the mechanism.

E. coli:

pathway	mechanism	N_0	N_1	p-value
central carbon metabolism	conserved	24	19	1.94E-010
histidine & nucleotide biosynthesis	conserved	52	31	4.19E-011
fatty acid β oxidation I	conserved	7	7	8.33E-006
glucoeogenesis	conserved	13	9	9.93E-005
colanic acid biosynthesis	conserved	10	7	5.88E-004
CDP-diacylglycerol biosynthesis II	conserved	4	4	1.27E-003
octaprenyl diphosphate biosynthesis	incorporation	5	4	7.71E-006
menaquinone-8 biosynthesis	incorporation	14	4	1.21E-003
proline degradation I	incorporation	2	2	1.31E-003
polyisoprenoid biosynthesis	incorporation	16	4	2.08E-003
isoleucine biosynthesis	intra duplication	5	2	1.69E-003
phenylacetate degradation I	inter duplication	5	5	2.52E-004
ppGpp biosynthesis	inter duplication	4	4	1.33E-003
ornithine biosynthesis	inter duplication	5	4	5.64E-003
ADP-L-β-D-heptose biosyn.	inter duplication	5	4	5.64E-003
peptidoglycan biosynthesis I	novel	10	9	9.18E-004
Phe, Tyr and Trp biosyn.	novel	19	14	1.36E-003
flavin biosynthesis	novel	9	8	2.24E-003
chorismate biosynthesis	novel	8	7	5.41E-003

Human:

pathway	mechanism	N_0	N_1	p-value
central carbon metabolism	conserved	24	19	8.89E-009
histidine & nucleotide biosynthesis	conserved	52	31	1.16E-008
fatty acid β oxidation I	conserved	7	7	3.72E-005
uridine-5'-monophosphate biosyn.	conserved	6	6	1.61E-004
gluconeogenesis	conserved	13	9	5.60E-004
colanic acid biosynthesis	conserved	10	7	2.28E-003
CDP-diacylglycerol biosynthesis II	conserved	4	4	2.99E-003
uridine-5'-monophosphate biosyn.	incorporation	6	5	3.30E-005
pyrimidine biosynthesis	incorporation	10	5	1.02E-003
fatty acid elongation	incorporation	5	3	6.53E-003
central carbon metabolism	intra duplication	24	10	2.26E-006
gluconeogenesis	intra duplication	13	7	1.07E-005
fatty acid β oxidation I	intra duplication	7	5	3.54E-005
histidine & nucleotide biosynthesis	intra duplication	52	12	2.06E-004
serine-isocitrate lyase	intra duplication	14	6	2.44E-004
CDP-diacylglycerol biosynthesis I	intra duplication	4	3	1.44E-003
phospholipid biosynthesis I	intra duplication	11	4	5.97E-003
N-acetylneuraminate degradation	intra duplication	18	5	7.53E-003
isoleucine degradation III	inter duplication	9	6	1.45E-003
valine degradation I	inter duplication	7	5	2.48E-003
methylmalonyl pathway	inter duplication	3	3	5.36E-003
cholesterol biosynthesis	novel	50	37	2.80E-005

Most conserved reactions retain at least one common domain between the two species. However, 64 reactions undergo domain displacement, i.e., domain compositions of the enzymes of the two species are completely different. One possible cause to completely replace the domain architectures is the function of the catalytic apparatus is changed. The two sets of enzymes may catalyze distinct reactions except the one where domain displacement occurs. We test this hypothesis by counting the overlap of co-catalytic partners of reactions with domain displacement and all conserved reactions. About half of reactions with domain displacement share no co-catalytic partners between the two species (35 in 64), whereas only one third of all conserved

reactions have this property (95 in 281). The enrichment is statistically significant (hyper-geometric p-value < 0.0017).

Table 4: Pathways of reactions undergoing domain fusion

pathway	species of fusion	# fused reactions
proline degradation	*E. coli*	2
Fatty acid elongation	Human	3
Fatty acid biosyn. initiation	Human	2
Uridine-5'-monophosphate de novo biosyn.	Human	2
Uridine-5'-monophosphate de novo biosyn.	Human	3
tRNA charging pathway	Human	2
Purine nucleotides de novo biosyn.	Human	3

In *E. coli* the biosynthesis involved in electron carriers (octaprenyl diphosphate, menaquinone-8 and polyisoprenoid) and proline degradation are enriched with domain novel function incorporation. In human the biosynthesis of uridine-5'-monophosphate, pyrimidine ribonucleotides and fatty acid elongation are enriched with incorporation. Most incorporation events in these pathways are domain fusion events. Table 4 lists the pathways of domain fusion detected in the metabolic networks of *E. coli* and human. For instance, three successive reactions in the pathway of fatty acid elongation are catalyzed by three distinct proteins in *E. coli* and one fused enzyme in human (Wakil 1989). All the domain fusion events occur in consecutive reactions in the same pathways, reaffirming the previous conclusion that physically and functionally coupled proteins tend to fuse together (Enright et al. 1999).

Table 5: Origins of selected novel reactions undergoing domain novel function incorporation

E. coli:

origin reaction id	pathway of origin reaction
OHACYL-COA-DEHYDROG-RXN	fatty acid β oxidation I
ACONITATEDEHYDR-RXN	central carbon metabolism

destination reaction id	pathway of destination reaction
RXN0-5391	oleate β oxidation
4.2.1.99-RXN	methylcitrate cycle I

Human:

origin reaction id	pathway of origin reaction
GLYCOPHOSPHORYL-RXN	glycogen degradation I
GLYC3PDEHYDROG-RXN	glycerol degradation I
PYRROLINECARBDEHYDROG	proline degradation I
SAICARSYN-RXN	purine biosynthesis I
OHACYL-COA-DEHYDROG-RXN	fatty acid β oxidation I

destination reaction id	pathway of destination reaction
RXN-9025	glycogen degradation II, III
RXN-6841	glycerol degradation IV
HYDROXYPYRROLINEDEH-RXN	hydroxyproline degradation I
AIRCARBOXY-RXN	purine biosynthesis II
TIGLYLCOA-HYDROXY-RXN	isoleucine degradation I

If a novel reaction is incorporated by a conserved enzyme, then the conserved function carried by this enzyme can be viewed as the origin of the novel reaction. Table 5 lists the origins of selected novel reactions according to domain incorporation. Most novel reactions arise from reactions of similar processes. For instance, a reaction in the second pathway of glycogen degradation (RXN-9025) is incorporated by PYGM, an enzyme catalyzing a reaction in the first pathway of glycogen degradation (GLYCOPHOSPHORYL-RXN). Similarly, reactions in the degradation or biosynthesis of various amino acids, glycerol and nucleotides are derived from reactions metabolizing identical or similar compounds.

Intra-domain duplication occurs primarily in human metabolic pathways. Specifically, about half of the reactions in the central carbon metabolism (10 of 24) and gluconeogenesis (7 of 13) undergo intra-reaction domain duplication. Fatty acid β oxidation I pathway (5 of 7) and the superpathway of nucleotide biosynthesis (12 of 52) are also enriched with intra-reaction domain duplication.

Despite the abundance of inter-reaction domain duplication in both species, fewer pathways are enriched with reactions of this kind. In *E. coli* most or all reactions in the pathways of phenylacetate degradation I (5 of 5), ppGpp biosynthesis (4 of 4), ornithine biosynthesis (4 of 5) and several others encounter domain duplication. In human isoleucine degradation III (6 of 9), valine degradation I (5 of 7) and methylmalonyl metabolism (3 of 3) are enriched with inter-reaction domain duplication.

Table 6: Top domains undergoing inter-reaction duplication. # dups: # duplications

E. coli:

accession	description	# dups
PF00070	pyridine nucleotide-disulphide oxidoreductase	15
PF07992	pyridine nucleotide-disulphide oxidoreductase	15
PF00171	aldehyde dehydrogenase family	13
PF00294	pfkB family carbohydrate kinase	11
PF00501	AMP-binding enzyme	10
PF01048	phosphorylase family	9
PF03099	biotin/lipoate A/B protein ligase family	9
PF00106	short chain dehydrogenase	8
PF00109	β-ketoacyl synthase, N-terminal domain	8
PF00291	pyridoxal-phosphate dependent enzyme	8
PF02801	β-ketoacyl synthase, C-terminal domain	8
PF00293	NUDIX domain	7

Human:

accession	description	# dups
PF00106	short chain dehydrogenase	21
PF00107	zinc-binding dehydrogenase	16
PF01266	FAD dependent oxidoreductase	10
PF02770	acyl-CoA dehydrogenase, middle domain	10
PF00248	aldo/keto reductase family	9
PF00441	acyl-CoA dehydrogenase, C-terminal domain	9
PF00171	aldehyde dehydrogenase family	8
PF01370	NAD dependent epimerase/dehydratase family	8
PF02771	acyl-CoA dehydrogenase, N-terminal domain	8
PF00378	enoyl-CoA hydratase/isomerase family	7
PF08240	alcohol dehydrogenase GroES-like domain	7
PF00111	2Fe-2S iron-sulfur cluster binding domain	6

Table 6 lists the top domains undergoing inter-reaction duplications. Most of the frequently duplicated domains are catalytic subunits carrying out common chemical operations, particularly oxidation/reduction and proton transfer. Oxidoreductases and dehydrogenases constitute 4 and 9 of the top 12 domains in *E. coli* and human respectively. Other top domains include kinases, phosphorylases and hydratases. In *E. coli* the AMP-binding domain appears in a variety of

ATP-dependent reactions.

Despite both species encounter frequent novel domain addition, human has much fewer enriched pathways than *E. coli*. Most pathways enriched with novel domain addition produce *E. coli* or human specific compounds: the biosynthesis of peptidoglycan (bacterial cellular surface coat), riboflavin (vitamin B2) and chorismate (a microbe-specific amino acid intermediate) in *E. coli*, and the biosynthesis of cholesterol in human. However, two pathways of amino acid biosynthesis (histidine; phenylalanine, tyrosine, and tryptophan) in *E. coli* are also enriched with novel domain addition.

Conclusion

One remarkable discovery from the recent development of genomics is that the complexity and diversity of biological systems do not only arise from genome size and sequence disparity but also from the combinations of common subunits such as protein domains and regulatory motifs. Alterations of domain architectures of enzymes are prevalent and are a major driving force for metabolic network evolution. In this study, we demonstrate these alterations exhibit strong heterogeneity and probably play distinct roles in different species and metabolic processes.

The distinction between *E. coli* and human domain compositions largely reflects the trends of systems evolution after their divergence. While inter-reaction domain duplication and novel domain addition are the dominant changes in both species, human encounters substantially more functional incorporation and intra-reaction duplication. These differences explain the higher level of pleiotropy and redundancy of human enzymes. Furthermore, distinct metabolic processes seem to have differential selective constraints on domain alterations. The central core of metabolism – nucleotide biosynthesis, energy utilization and lipid biosynthesis – retains more conserved domains and encounters more intra-reaction domain duplication in human. In addition, some alternative pathways of similar metabolic processes (such as amino acid degradation) are formed by incorporating conserved enzymes with novel functions. Furthermore, inter-reaction duplication tends to utilize paralogous domains from different functional processes, whereas novel domain addition facilitates complicated information processing in human (signaling pathways, hormone synthesis) and synthesis of species-specific compounds (chorismate and vitamin B2 for *E. coli*, cholesterol for human).

The analysis in this work focuses on the most prominent changes of metabolic network evolution after the prokaryotes-eukaryotes divergence. Investigation of more refined yet important metabolic alterations requires the data of more organisms. Some of the interesting questions include: How is a specific metabolic pathway derived and enriched from the recruitment of domains from other metabolic processes? What is the domain duplication history of tissue-specific isozymes? How are the metabolic networks of bacteria evolved to adapt diverse environment on earth? The parsimonious reconstruction algorithm can be extended to multiple species with minor modifications. However, the quality and completeness of the data in other species may limit the accuracy of inference results. Improved annotations of protein domain architectures and the enzyme-reaction associations of more species will be a critical factor to study the domain evolution of metabolic networks.

References

[1] 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.

[2] Morowitz H. A theory of biochemical organization, metabolic pathways, and evolution. 1999. Complexity 4:39-53.

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

[4] Jeong H, Tombor B, Albert R, Oltvai ZN, and Barbasi AL. 2000. The large-scale organization of metabolic networks. Nature 407:651-654.

[5] Barabasi AL, Albert R. 1999. Emergence of scaling in random networks. Science286:509-512.

[6] Carlson JM and Doyle J. 2002. Complexity and robustness. Proc. Natl. Acad. Sco. USA 99 supp 1:2538-2545.

[7] Enattah NS, Sahi T, Savilahti E, Terwilliger JD, PeltonenL, Jarvela I. 2002. Identification of a variant associated with adult-type hypolactasia. Nat. Genet. 30:233-237.

[8] Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

[9] Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. 2004. Structure, function and evolution of multidomain proteins. Curr. Opin. Struct. Biol. 14:208-216.

[10] Fani R, Brilli M, Fondi M, LiÃş P. 2007. The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case. BMC Evol. Biol. 7 Supp 2:S4.

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

[12] Teichmann SA, Rison SCG, Thornton JM, Riley M, Gough J, Chothia C. 2001. The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. J. Mol. Biol. 311:693-708.

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

[14] Vogel C, Berzuini C, Bashton M, Gough J, and Teichmann SA. 2004. Supra-domains: evolutionary units larger than single protein domains. J. Mol. Biol. 336:809-823.

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

[16] Fukami-Kobayashi K, Minezaki Y, Tateno Y, Nishikawa K. 2007. A tree of life based on protein domain organizations. Mol. Biol. Evol. 24:1181-1189.

[17] Caetano-AnollÃl's G, Kim HS, Mittenthal JE. 2007. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. Proc. Natl. Acad. Sci. USA. 104(22):9358-9363.

[18] Corman TH, Leiserson CE, Rivest RL, Stein C. 2001. Introduction to algorithms MIT Press & McGraw-Hill 643-700.

[19] Freilich S, Spriggs RV, George RA, Al-Lazikani B, Swindells M, Thornton JM. 2005. The complement of enzymatic sets in different species. J. Mol. Biol. 349(4):745-763.

[20] Schmidt S, Sunyaev S, Bork P, Dandekar T. 2003. Metabolites: a helping hand for pathway evolution? Trends Biochem. Sci. 28(6):336-341.

[21] Wakil SJ. 1989. Fatty acid synthase, a proficient multifunctional enzyme. Biochemistry 28(11):4523-4530.

[22] Enright AJ, Iliopoulos I, Krypides NG, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86-90.