Extraction of Phylogenetic Network Modules from Prokayrote Metabolic Pathways

Takuji Yamada takuji@kuicr.kyoto-u.ac.jp
Susumu Goto* goto@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Abstract

In the post-genomic era, it is important to analyze interaction networks that include genes, proteins, enzymes and compounds such as a metabolic pathway. Every organism has such networks individually. However, several parts of them are conserved in different organisms. The purpose of this analysis is to extract sub-networks composed of these common elements through the phylogenetic analysis. We extracted network modules from metabolic pathways using phylogenetic profile and cluster analysis. The enzymes of these modules are related by evolutionary and functional correlation. Our results give a valuable insight into the evolution of metabolic pathways.

Keywords: metabolic pathway, phylogenetic profile, pathway module

1 Introduction

The interaction between molecules such as proteins, genes and compounds plays a central role in the complicated cellular system. Many experimental methods have been developed for detecting these interactions extensively [5, 13]. One of the most important issues in recent studies is how to deal with the cellular function assumed by qualitatively different interactions integratively. There are a lot of comprehensive analyses using multi-interactions for finding and characterizing unknown functions [12]. In general, those interactions are described as an integral network where metabolic and regulatory pathways are representative examples. It is known that a structure of the network is scale free, and recent studies revealed that the topology also contained modularity [10].

As the database containing the interactions described above, the KEGG database presents various relations between genes and metabolites. Metabolic and regulatory pathways are also depicted there [6]. In this database, metabolic pathways are composed of compounds and enzymes which catalyze chemical reaction. In the current version of KEGG, 133 completely sequenced organisms (7 eukaryotes, 110 bacteria, 16 archaea) are described and stored.

In the case of network analysis usually only one or a few organisms are contemplated. Each organism has its own particular genome, so it includes a certain gene set. Although this gene set forms a particular network in each organism, in many cases, there are homologous parts among the networks from several organisms. Experimental analysis revealed the common part of some metabolic pathways, and these were often found to be essential parts. Phylogenetic profiles are used in order to establish relationships between orthologs especially for the large number of genomes. The phylogenetic profile of a gene is represented by a string that encodes the presence or absence of the gene in every fully sequenced genome [9](Fig. 1(A)). In the case where two genes have similar profiles, they are assumed to be evolutionary and functionally correlated.

Using the similarity between profiles and connectivity on the metabolic pathway, we extracted conserved sub-networks. Those "pathway modules" are comprised of the enzymes having the same phylogenetic pattern, and also close to each other in the metabolic pathway.

*Correspondence: Susumu Goto (goto@kuicr.kyoto-u.ac.jp)

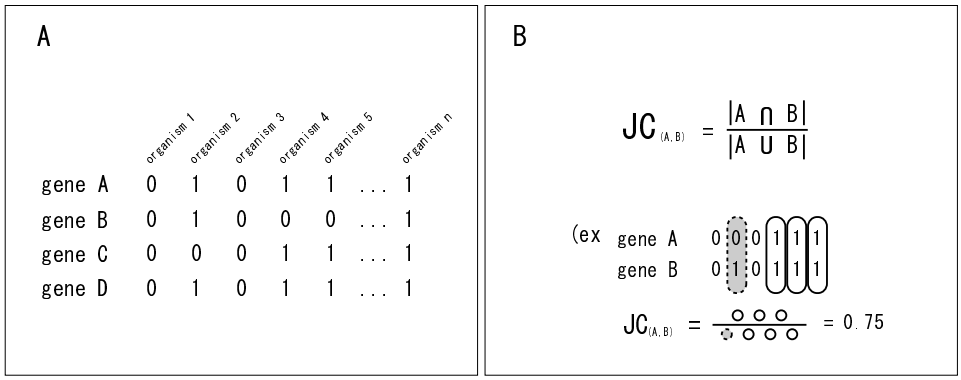

Figure 1: Phylogenetic profile and Jaccard coefficient: (A)The phylogenetic profile of a gene is the string that encodes its presence or absence in different genomes. We define the number of organisms that have the gene as "phylogenetic extent", and use it as an index for conservation among genomes. In this analysis, we constructed phylogenetic profiles corresponding to EC numbers. (B)Jaccard coefficient(JC) as a measurement of similarity between profiles. A∪B means the number of organisms that have either gene A or gene B, or both. A∩B means the number of organisms that have both genes A and B.

2 Data Set and Method

2.1 Data Set

We constructed phylogenetic profiles by using the KEGG/GENES database where 3611 ortholog groups are defined according to sequence similarity, functional relationship and gene location. From the ortholog groups which contain 104 fully sequenced prokaryotic genomes, 871 phylogenetic profiles of single enzymes were constructed, ortholog groups absent from metabolic pathways were eliminated. The organisms included in these profiles are listed in Table 1. Although the ortholog groups have been constructed by using 104 organisms, some are closely related to each other. Therefore, we merged several organisms into 38 taxa according to the NCBI taxonomy [19].

The metabolic pathway was extracted from the KEGG/PATHWAY database. Binary relations between adjacent enzymes were obtained from the pathway map [18], and adjacent enzymes in this case mean that a product of one enzyme is a substrate of another. One or more ortholog groups correspond to a particular enzyme which is defined by the enzyme classification system (EC) [17]. When multiple ortholog groups correspond to a unique enzyme, we merged them into a single profile and assigned it to the enzyme in the pathway map.

2.2 Profile Similarity and Pathway Distance

As a similarity measure for phylogenetic profiles, we adopted the Jaccard coefficient(JC) which is defined by (A∩B)/(A∪B) (Fig. 1(B)). The JC is simple and biologically intuitive [15] compared to other proposed measurements such as correlation coefficient, humming distance, dot-product, or kernel method [8, 14, 16]. When two profiles are same, JC between them becomes 1.0, on the other hand, 0.0 indicates that the two profiles have no relationships.

To investigate the relationship between function and similarity of the phylogenetic profile(JC), We used the pathway distance, which is defined by the shortest path between two enzymes on the pathway represented by a graph consisting of the binary relations between adjacent enzymes [12]. We assessed the relationship between pathway distance and similarity of phylogenetic profiles by measuring an average of phylogenetic profile similarities for each distance.

Table 1: List of organisms used in this analysis. We merged 104 organisms into 38 taxa according to the NCBI taxonomy. The left most column shows the index for taxa, the second is the index for organisms, the third is abbreviation used in KEGG, and the right most column represents simplified names of the organisms.

1	1	eco	E.coli
	2	ecj	E.coli_J
	3	ece	E.coli_O157
	4	ecs	E.coli_O157J
	5	ecc	E.coli_CFT073
	6	sty	S.typhi
	7	stm	S.typhimurium
	8	ype	Y.pestis
	9	ypk	Y.pestis_KIM
	10	sfl	S.flexneri
	11	buc	Buchnera
	12	bas	B.aphidicola_Sg
	13	bab	B.aphidicola_Bp
	14	wbr	W.brevipalpis
2	15	hin	H.influenzae
	16	pmu	P.multocida
3	17	xfa	X.fastidiosa
	18	xft	X.fastidiosa_T
	19	xcc	X.campestris
	20	xac	X.axonopodis
4	21	vch	V.cholerae
	22	vvu	V.vulnificus
5	23	pae	P.aeruginosa
	24	ppu	P.putida
6	25	son	S.oneidensis
7	26	nme	N.meningitidis
	27	nma	N.meningitidis_A
8	28	rso	R.solanacearum
9	29	rpr	R.prowazekii
	30	rco	R.conorii
10	31	mlo	M.loti
	32	sme	S.meliloti
	33	atu	A.tumefaciens
	34	atc	A.tumefaciens_C
	35	bme	B.melitensis
	36	bms	B.suis
	37	bja	B.japonicum
11	38	ccr	C.crescentus
12	39	bsu	B.subtilis
	40	bha	B.halodurans
	41	oih	O.iheyensis
	42	sau	S.aureus_N315
	43	sav	S.aureus_Mu50
	44	sam	S.aureus_MW2
	45	sep	S.epidermidis
	46	lmo	L.monocytogenes
	47	lin	L.innocua
13	48	lla	L.lactis
	49	spy	S.pyogenes
	50	spm	S.pyogenes_M18
	51	spg	S.pyogenes_M3
	52	spn	S.pneumoniae
	53	spr	S.pneumoniae_R6
	54	sag	S.agalactiae
	55	san	S.agalactiae_NEM316
	56	smu	S.mutans
14	57	cac	C.acetobutylicum
	58	cpe	C.perfringens
	59	ctc	C.tetani
15	60	tte	T.tengcongensis
16	61	mge	M.genitalium
	62	mpn	M.pneumoniae
	63	mpu	M.pulmonis
	64	mpe	M.penetrans
	65	uur	U.urealyticum
17	66	mtu	M.tuberculosis
	67	mtc	M.tuberculosis_CDC1551
	68	mle	M.leprae
	69	cgl	C.glutamicum
	70	cef	C.efficiens
	71	sco	S.coelicolor
18	72	blo	B.longum
19	73	fnu	F.nucleatum
20	74	ctr	C.trachomatis
	75	cmu	C.muridarum
	76	cpn	C.pneumoniae
	77	cpa	C.pneumoniae_AR39
	78	cpj	C.pneumoniae_J138
21	79	bbu	B.burgdorferi
	80	tpa	T.pallidum
	81	lil	L.interrogans
22	82	syn	Synechocystis
	83	tel	T.elongatus
23	84	ana	Anabaena
24	85	cte	C.tepidum
25	86	dra	D.radiodurans
26	87	aae	A.aeolicus
27	88	tma	T.maritima
28	89	mja	M.jannaschii
29	90	mac	M.acetivorans
	91	mma	M.mazei
30	92	mth	M.thermoautotrophicum
31	93	mka	M.kandleri
32	94	afu	A.fulgidus
33	95	hal	Halobacterium
34	96	tac	T.acidophilum
	97	tvo	T.volcanium
35	98	pho	P.horikoshii
	99	pab	P.abyssi
	100	pfu	P.furiosus
36	101	ape	A.pernix
37	102	sso	S.solfataricus
	103	sto	S.tokodaii
38	104	pai	P.aerophilum

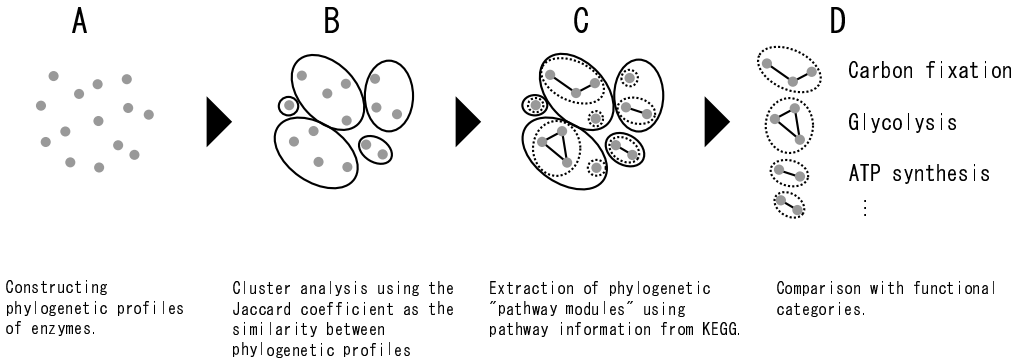

Figure 2: Outline of the clustering method: Each dot represents an enzyme. In B and C, the ovals with solid line represent clusters using Jaccord coefficient as the similarity score. The edge between nodes represents the binary relation between two enzymes which are defined as being located adjacently on the metabolic pathway map. Dotted ovals in C and D represent the "pathway modules".

2.3 Cluster Analysis Using Similarity of Phylogenetic Profile (JC)

We obtained enzyme clusters using the hierarchical complete linkage clustering [4]. The similarity or the distance measure between the elements is required in the hierarchical clustering, and the result obtained from this method depends on the measure. In this analysis, the Jaccard coefficient was used as the similarity measure of phylogenetic profiles. The threshold was defined by calculating all-against-all similarities among the profiles for all enzyme genes, and detecting significant point P(0.05).

2.4 Extracting Modules Using Pathway Information

Each one of obtained clusters was divided into functionally related subgroups by iteratively connecting two enzymes which have a linkage on the metabolic pathway. Information of enzyme linkages was collected from binary relation of enzymes in the KEGG database. Each enzyme group obtained in this way was defined as a "pathway module". We mapped each module on the KEGG/PATHWAY, and investigated the location and function of each module according to the KEGG functional categories. Using linkages between enzymes belonging to different modules, we constructed a global network of pathway modules. An outline of the method is shown in Fig. 2.

3 Result and Discussion

3.1 Pathway Distance and the Clustering Result

Fig. 3(A) represents a distribution of average similarities for each pathway distance. In the case that pathway distance between two enzymes is small, they are expected to have a close functional relationship. Negative correlation was observed between the average of similarities and pathway distance, indicating that, the nearer two enzymes are located in the pathway, the higher the similarity score of their phylogenetic profiles is. This fact indicates close relationship between phylogenetic pattern (described by the phylogenetic profile) and the cellular function. Fig. 3(B) shows the distribution of Jaccard-coefficient. Significant point P(0.05) of the distribution was found to be 0.69, and we set this value as a threshold of JC for the clustering analysis. The number of enzyme pairs with the JC values above the threshold was 17373.

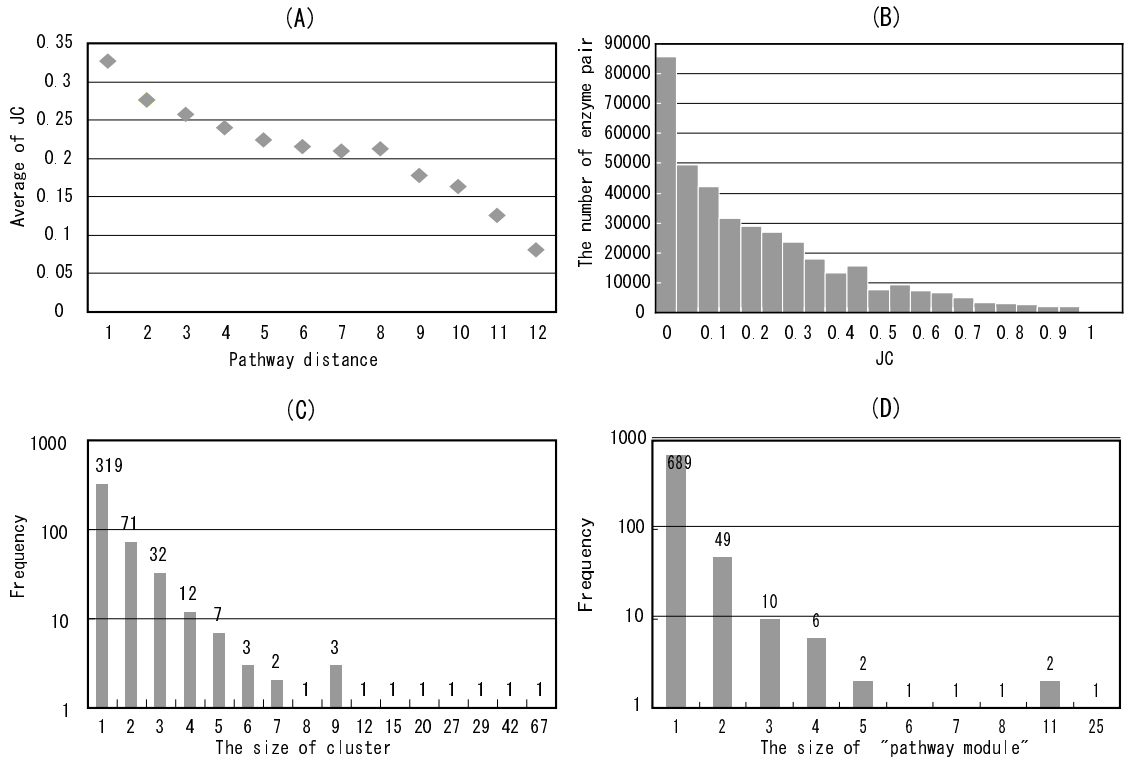

Figure 3: Distribution of (A)average of JC in each pathway distance, (B)Jaccard-coefficient from all-against-all enzyme pairs, P(0.05)=0.69, (C)size of the clusters using similarity between phylogenetic profile, and (D)size of the "pathway modules"

Fig. 3(C) indicates the distribution of cluster size. Although 457 clusters were obtained, more than 60% among the clusters contained only one enzyme. On the contrary, the largest cluster contained 67 enzymes, which spread across the several categories of the metabolic pathways. Enzymes in the same cluster have similar phylogenetic profiles, meaning that they have similar evolutionary histories, such as gene-loss or gene-duplication. Fig. 3(D) indicates the frequency of enzymes included in the same "pathway module" (state D in Fig. 2). Although 762 modules were obtained from 871 enzymes, as in the previous case, most modules contained only one enzyme, and only 8 modules contained 5 or more enzymes. Functional relationship between enzymes is often clear when those enzymes belong to the same pathway category, such as glycolysis and purine metabolism. However, the number of enzymes is different in each category. Therefore, to assess the functional relationship between enzymes on the pathway map, we should consider relative location of elements as a whole (see Data set and method). Ma *et al.* utilized such information, and reconstructed the pathway from binary relations of chemical reactions [7], other recent studies also used binary relation of enzymes or compounds to identify functions [1].

3.2 Phylogenetic Pathway Modules and Molecular Functions

Although most of pathway modules contain one or a few enzymes, several modules contain a large number of them. Some of the large modules expand across several pathway categories, and others are distributed in a single category. The modules indicate a characteristic feature which includes phylogenetic and functional information. Fig. 4 represents the largest module including 25 enzymes. Intermediate compounds corresponding to linkages between enzymes were extracted from the KEGG/LIGAND

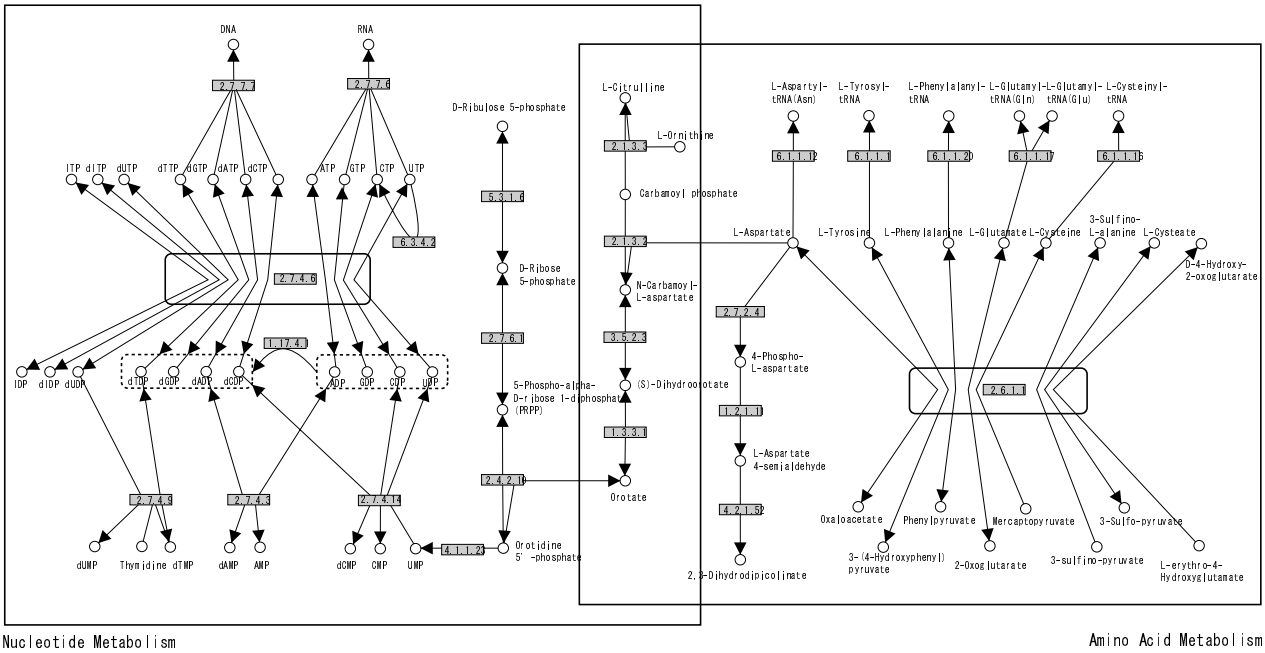

Figure 4: The largest pathway module. Circles, gray rectangles and arcs indicate chemical compounds, enzymes and direction of reaction, respectively. The enzymes in this modules belong to Amino acid metabolism and Nucleotide metabolism, and several enzymes are also included in Carbohydrate metabolism and Energy metabolism.

database and also shown. The phylogenetic extent shows that almost all organisms have all enzymes in this module. Although they extend across more than one pathways, all appear in Nucleotide and Amino acid metabolisms. Because the two categories play a critical role for cell life, this module seems to be one of the core structures of the metabolic pathway. In fact, there are several modules which have high phylogenetic extent similar to this module, and those modules are in the same cluster(Fig. 2(B)). This cluster includes 67 enzymes whose phylogenetic profiles are highly conserved, and the enzymes appear in the glycolysis, tryptophan biosynthesis, several amino acid metabolisms among others. Hence the cluster seems to construct a basic framework of the metabolic pathway. Specifically, aspartate transaminase (EC 2.6.1.1) and nucleotide diphosphokinase (EC 2.7.6.4) are linked to a lot of enzymes in this module.

Fig. 5 shows other example of large pathway modules mapped onto the pathway diagrams in the KEGG database (Fig. 5(A): Peptidoglycan biosynthesis, (B): Folate biosynthesis, (C): Valine, Leucine and Isoleucine biosynthesis, (D): Histidine metabolism). Most enzymes in the Peptidoglycan biosynthesis pathway(A) is covered by an eubacterial specific module. Methanogenesis, a part of Folate biosynthesis(B), is covered by an archaeal specific module. Several prokaryotes have operon like structures, and they are often observed in the Amino acid biosynthesis pathway. Modules in the (C) and (D) correspond to operon like structures. However, (D) indicates that phylogenetic pattern of enzymes is not uniform even though they are distributed in single pathway map in KEGG.

3.3 Global Network of Pathway Modules

Fig. 6 shows the global network comprised of the all pathway modules. A circle represents one module. The size of the circle indicates the number of enzymes in the module, and the gradation indicates the phylogenetic extent. The network shows that phylogenetic extent is high if the number of enzymes in a

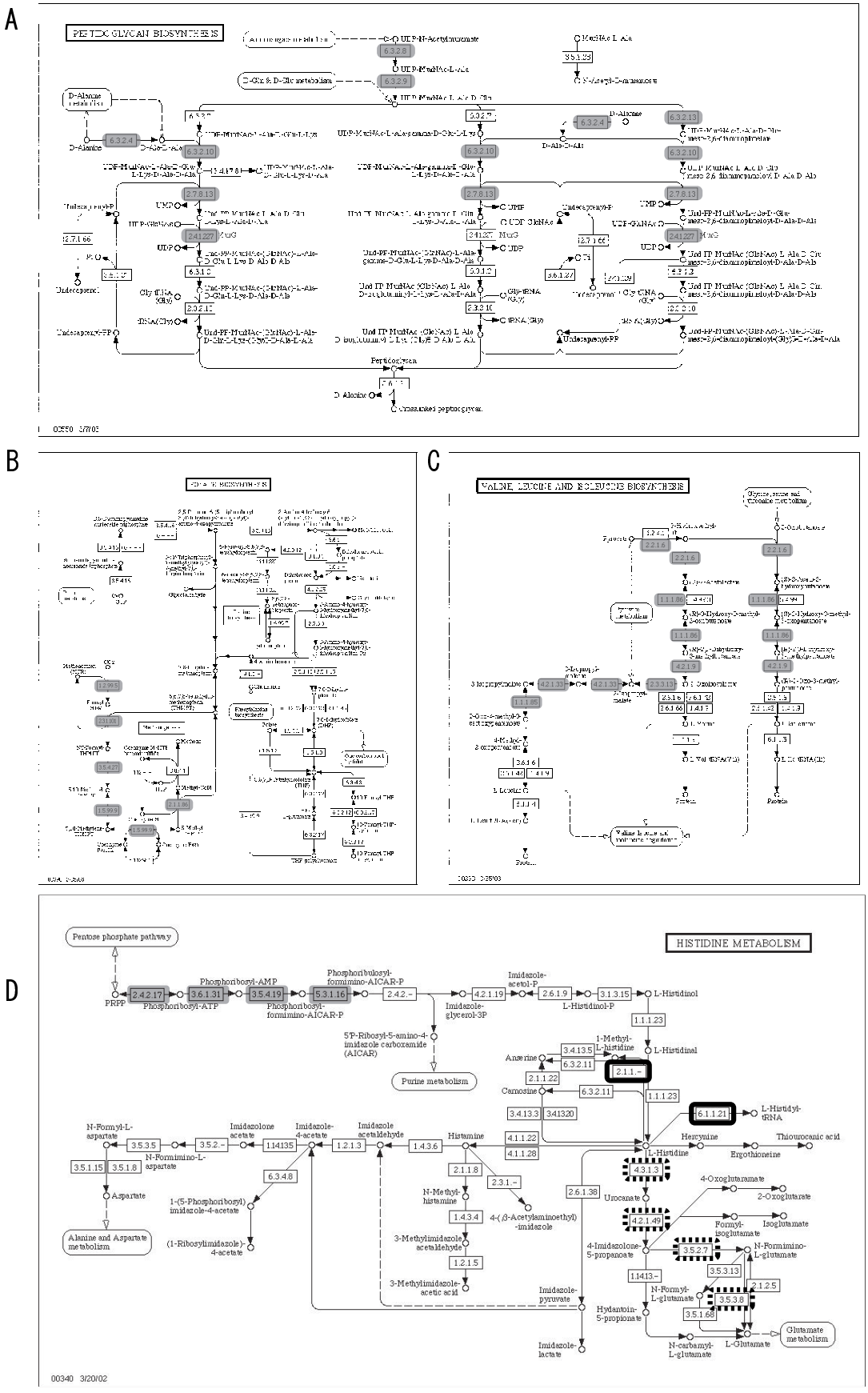

Figure 5: Examples of modules on the pathway. (A,B,C) Gray nodes represent elements in the same module in each map. (D) Nodes with gray, heavy line and dotted line correspond to respective modules.

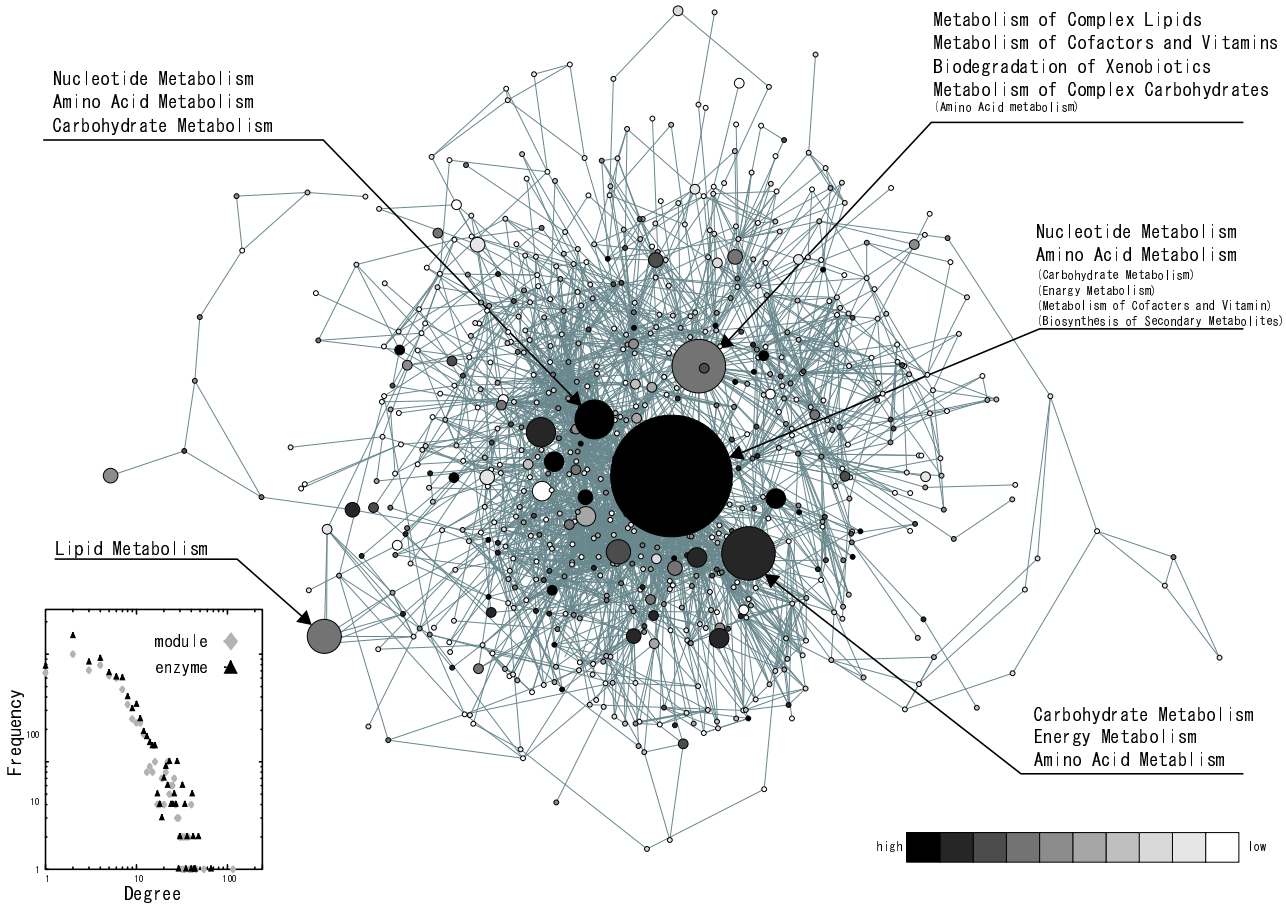

Figure 6: Global network of pathway modules. Nodes and edges correspond to the pathway modules and compounds linking the modules, respectively. The gradation indicates the phylogenetic extent. The graph plot in lower left shows the scale free topologies of both the modules and enzyme network.

module is large. In short, more larger, more conserved. There have been two models for the evolution of the metabolic pathway [11]. Both models handle the enzymes or compounds as a single nodes from the viewpoint of the graph theory, and do not consider a group of enzymes as an evolutionary unit. As the knowledge of the complete genomes is accumulated, we can now reevaluate these models with the large amount of genomic data [2, 11]. The result of our analysis suggests that enzymes which have similar phylogenetic profiles make a module in the metabolic pathway, so the module should be applied as a node rather than the individual enzyme as a node.

It is well known that the graph topology of the metabolic pathway shows a scale-free property. The topology of the graph comprised of modules also shows the similar property(lower left in Fig. 6). This fact is consistent with the hierarchical model of the metabolic pathway which was reported in a recent study [10]. It also suggested that the hierarchical model contains the modularity derived from the graph topology [3]. Phylogenetic and functional modules derived from this analysis are also consistent with the topological theory of network evolution. In addition, it is said that the network "hub" is critical because the number of interacting partners is large. In the graph whose nodes are pathway modules, large modules tend to be "hub" because the degree of modules depends on their size in part, although some archaea specific small modules play a role of "hub" as well as large modules.

4 Conclusion

We extracted phylogenetic modules from metabolic pathways, and found them to be composed of enzymes which have similar phylogenetic profiles. The pathway module represents basic "building block" of the metabolic pathways and some of them have a close correlation with functional modules. By developing the scoring system for relationship between pathway modules, the global system of pathway evolution will be clearer. Although eukaryotic pathways were not included in this analysis, because the number of eukaryotic genomes is not enough to construct phylogenetic profiles, they will be considered for future work.

5 Acknowledgments

Authors are grateful to Dr. R.Jauregui for helpful comments on an earlier draft of the manuscript. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeasts, T.O., Protein function in the post-genomic era, *Nature*, 405:823–826, 2000.

[2] Forst, C.V. and Schulten, K., Phylogenetic Analysis of Metabolic Pathways, *J. Mol. Evol.*, 52:471–489, 2001.

[3] Gagneur, J., Jackson, D.B., and Casari, G., Hierarchical analysis of dependency in metabolic networks, *Bioinformatics*, 19(8):1027–1034, 2003.

[4] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York Springer, 2001.

[5] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci. USA*, 98(8):4569–4574, 2001.

[6] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M., The KEGG resource for deciphering the genome, *Nucl. Acids. Res.*, 32:D277–D280, 2004.

[7] Ma, H. and Zeng, A.P., Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms, *Bioinformatics*, 19(2):270–277, 2003.

[8] McGuire, A.M. and Church, G.M., Predicting regulons and their cis-regulatory motifs by comparative genomics, *Nucl. Acids. Res.*, 28: 4523–4530, 2000.

[9] Pellegrini, M., Marcotte, E.M. and Thompson, M.J., Eisenberg, D., and Yeates, T.O., Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proc. Natl. Acad. Sci. USA*, 96:4285–4288, 1999.

[10] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabasi, A.L., Hierarchical organization of modularity in metabolic networks, *Science*, 297(5586):1551–1555, 2002.

[11] Rison, S.C. and Thornton, J.M. Pathway evolution, structurally speaking, *Current Opinion in Structural Biology*, 12:374–382, 2002.

[12] Rison, S.C., Teichmann, S.A., and Thornton, J.M., Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*, *J. Mol. Biol.*, 318(3):911–932, 2002.

[13] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi, E.A., Li. Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J.M., A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature*, 403(6770):623–627, 2000.

[14] Vert, J.P., A tree kernel to analyze phylogenetic profiles, *Bioinformatics*, 18:S276–284, 2002.

[15] Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., and Koonin, E.V., Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context, *Genome Res.*, 11: 356–372, 2001.

[16] Yamanishi, Y., Itoh, M., and Kanehisa, M. Extraction of organism groups from phylogenetic profiles using independent component analysis, *Genome Informatics*, 13:61–70, 2002.

[17] `http://www.chem.qmul.ac.uk/iubmb/enzyme/`

[18] `http://www.genome.ad.jp/kegg`

[19] `http://www.ncbi.nih.gov/Taxonomy/`