ANALYSIS OF COMMON SUBSTRUCTURES OF METABOLIC COMPOUNDS WITHIN THE DIFFERENT ORGANISM GROUPS

AI MUTO
muto@kuicr.kyoto-u.ac.jp

MASAHIRO HATTORI
hattori@kuicr.kyoto-u.ac.jp

MINORU KANEHISA
kanehisa@kuicr.kyoto-u.ac.jp

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

With the increase in available post-genomic data and metabolic pathway information, we have been focusing on revealing the biological meaning of higher phenomena such as relationships of metabolic systems in different organisms. Metabolism plays an essential role in all cellular organisms, e.g. energy transportation, signal transduction and structural formation of cell components. The metabolic pathway of each organism has a different landscape from all others because of the different sets of enzymes encoded in the genome. The organisms that are incapable of producing their own essential chemical compounds should acquire them in some way from other organisms that can produce them. For example, several vitamins are required by animals to survive. In this manner we can assume that the different availabilities of metabolites may influence the relationship between organisms in nature. In this study, we focus on the differences in available metabolites among organisms. First, we divided 239 species with complete genomes into 9 organism groups in accordance with phylogeny and averaged out the annotation quality and the phylogenetic sparsity. Then, we calculated the commonly used chemical compounds between organism groups and the uniquely used chemical compounds in an organism group. The total number of metabolites we consider in this study is 1,074, which is about one-third of all metabolites that appear in the KEGG metabolic pathways. Finally we show the differences and the similarities between organism groups on every metabolic pathway map, illustrating the commonly observed substructures within the uniquely used metabolites. These results will help us to better comprehend the architecture of metabolic pathways and the relationships between organisms.

Keywords: metabolism, metabolite, organism group, KEGG.

1. Introduction

Using increasingly advanced experimental techniques available to analyze cellular functions, a massive amount of valuable data for biological functions is being assembled through transcriptome [13], proteome [8] and metabolome analyses [11]. The complex network information like metabolic systems and protein regulation under cellular processes has also been compiled in KEGG PATHWAY database [9]. Using such genome scale data we can now investigate the complexity of biological systems. In particular, metabolic systems are essential for all cellular organisms, and most enzymatic reactions and metabolic compounds that comprise their metabolisms play an important role such as energy transportation, signal transduction and structural transformation of

chemical compounds to construct cellular components. Every organism should produce or degrade chemical compounds through their metabolic activities to survive in the biosphere. This means that metabolic systems contribute to the maintenance of life for each organism and it is very worthwhile to understand them.

In the course of considering such a biological system, we have analyzed the chemical aspects of metabolic pathways to better elucidate the biological meaning of metabolism so far [6, 10]. Here, we are focusing on the differences in available metabolites among organisms. The metabolic pathway of each organism has a different look from all others because of the different sets of enzymes which are encoded in each genome. While a fraction of the available metabolites frequently overlaps between organisms, other compounds rarely overlap because of differences in their available enzymes. Therefore, the organisms which never metabolize their own essential chemical compounds should acquire them from other organisms that can produce them in order to survive. For example, animals should acquire several vitamins by predation activity from other organisms such as plants. In another case, we can discover that the metabolic difference is directly linked to a phenotypic difference because of different usage of metabolites. For instance, some prokaryotes can utilize hydrogen sulfide (H_2S) as nutrient, but it is well-known that this chemical compound is poisonous to animals in the vapor state [4]. In this manner we can assume that the different availabilities of metabolites may influence the relationship between organisms in nature.

To examine such a correlation we can use organism-specific metabolism information from the KEGG database, which is compiled from genome information, i.e., the set of enzymes encoded in each genome. However, levels of gene annotation in each organism vary widely, causing problems when iterating over organism pairs. In order to avoid such difficulties, we first divided 239 species with complete genomes into 9 organism groups in accordance with phylogeny. Then, we performed a comprehensive analysis on metabolite usage among different organism groups. For each pair of organism groups, we determined those chemical compounds that were common to both and those that were specific to each member of the pair. After that, we verified the differences and similarities between groups on every metabolic pathway map. Finally, we identified specific common substructures within the uniquely used compounds by surveying structural differences among metabolites. This information on preferred substructures will help us to know the utility and the limitations of each metabolic pathway in terms of chemical structure; thus, we will be able to understand the structural design of metabolic pathways and the relationship between organisms.

2. Results

2.1. *Extraction of the commonly possessed metabolites*

In the KEGG database, information on metabolic pathways has been manually curated and stored in the PATHWAY database on the basis of genome information, IUBMB enzyme information, textbook information and so forth. KEGG also provides the chemical information for metabolites in the KEGG LIGAND database. Using these databases, we can easily elucidate which components of metabolic pathways appear within each organism group. In this section, we first extracted the whole set of metabolic compounds on all metabolic pathways for each species, using the KGML information obtained from KEGG PATHWAY database. The number of metabolic compounds we used is 3,499.

After that, we identified 1,074 commonly used metabolites among all species in each organism group, as illustrated in Fig. 1. The results are shown in the "Total" column in Table 1. Here the commonly used compounds are defined as metabolites that over 80% of the species in each group possess in each metabolic system. We determined that the total number of identified compounds is about one-third of the metabolites that appear in metabolic pathway maps in KEGG. The organism group which has the highest number of metabolites is Fungi with 743 compounds, four times that of Spirochete, which has the fewest metabolites. The pair of groups sharing the highest number of common compounds is Fungi and gamma-Proteobacteria with 500 compounds in common.

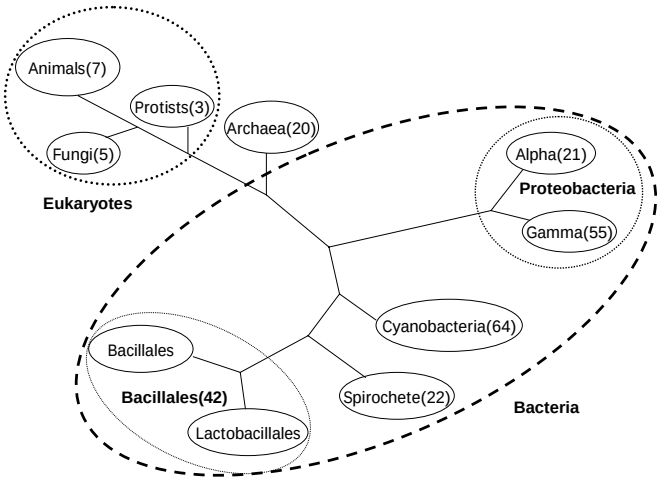

Fig. 1. Organism groups based on phylogeny
We obtained all phylogenetic information from the categorization used in KEGG database (http://www.genome.jp/kegg/catalog/org_list.html). The numbers in parentheses represent the numbers of organisms contained in each group.

The distribution of compounds in each organism group in each pathway category is shown in Table 2. Because some metabolites appear in more than one pathway, the sum of metabolites is not equal to the "Total" in Table 1.

Table 1. The number of commonly used compounds between organism groups
Each number represents the number of commonly observed chemical compounds between two organism groups. The "Total" in the last column means the unique number of compounds obtained from one organism group. Some labels are abbreviated as follows: alpha-Proteobacteria as Alpha, gamma-Proteobacteria as Gamma, and Cyanobacteria as Cyano.

	Animals	Fungi	Protists	Archaea	Alpha	Gamma	Bacillales	Spirochete	Cyano	Total
Animals	–	403	166	217	226	320	263	128	178	579
Fungi	403	–	214	354	316	500	404	177	237	743
Protists	166	214	–	143	149	171	179	118	150	220
Archaea	217	354	143	–	230	341	267	147	187	387
Alpha	226	316	149	230	–	351	283	146	207	408
Gamma	320	500	171	341	351	–	398	178	248	662
Bacillales	263	404	179	267	283	398	–	181	268	468
Spirochete	128	177	118	147	146	178	181	–	181	182
Cyano	178	237	150	187	207	248	268	181	–	268

Table 2. Distribution of compounds of each organism group in each pathway category
Each number represents the number of chemical compounds of each organism group that are observed in each pathway category. The pathway categories are obtained from the categorization used in the KEGG database (http://www.genome.jp/kegg/pathway.html).

	Animals	Fungi	Protists	Archaea	Alpha	Gamma	Bacillales	Spirochete	Cyano
Carbohydrate Metabolism	115	133	47	50	59	102	94	33	55
Energy Metabolism	41	60	23	35	35	55	35	14	20
Lipid Metabolism	122	126	31	55	54	80	62	8	15
Nucleotide Metabolism	66	86	36	62	58	79	74	54	59
Amino Acid Metabolism	143	226	51	127	108	185	145	45	78
Metabolism of Other Amino Acids	39	45	16	11	21	48	30	6	18
Glycan Biosynthesis and Metabolism	8	12	8	6	27	41	24	0	20
Biosynthesis of PK and NRP*	0	0	0	0	0	19	0	0	0
Metabolism of Cofactors and Vitamins	36	93	11	46	73	111	57	20	32
Biosynthesis of Secondary Metabolites	23	64	11	42	6	59	27	0	5
Xenobiotics Biodegradation and Metabolism	80	25	0	8	34	14	12	0	0

*PK and NRP: Polyketides and Nonribosomal Peptides

2.2. *Extraction of the different usage of metabolites*

In order to identify the different availability of chemical compounds within each organism group, we extracted the set of chemical compounds unique to one organism group, as shown in Table 3. In eukaryotes, the set of chemical compounds overlap greatly between Animals and Fungi, while the set of chemical compounds of Protists is almost completely included within the set of other eukaryotes, namely Animals and Fungi. This corresponds with the fact that parasitic protists have reduced metabolic systems. The pair of organism groups with the most similar set of chemical compounds is Fungi and gamma-Proteobacteria despite their more distant evolutionary relationship.

As for the "Specific" compounds that are uniquely observed in each organism group, many chemical compounds are extracted from eukaryotes. In contrast, bacteria have a very small number of intrinsic metabolites. Most of the frequently observed chemical compounds that are specifically possessed by eukaryotes are found to be lipid-

related metabolites whose chemical structures are relatively large (data not shown), indicating the existence of eukaryote-specific lipid-metabolisms.

Table 3. The number of uniquely used compounds in each organism group

Each number represents the number of uniquely used chemical compounds by the organism group in the row against the group in the column. The number in "Specific" is the number of specifically observed compounds in each organism group listed in each row.

	Animals	Fungi	Protists	Archaea	Alpha	Gamma	Bacillales	Spirochete	Cyano	Specific
Animals	–	176	413	362	353	259	316	451	401	111
Fungi	340	–	529	389	427	243	339	566	506	87
Protists	54	6	–	77	71	49	41	102	70	24
Archaea	170	33	244	–	157	46	120	240	200	6
Alpha	182	92	259	178	–	57	125	262	201	1
Gamma	342	162	491	321	311	–	264	484	414	1
Bacillales	205	64	289	201	185	70	–	287	200	3
Spirochete	54	5	64	35	36	4	1	–	1	0
Cyano	90	31	118	81	61	20	0	87	–	0

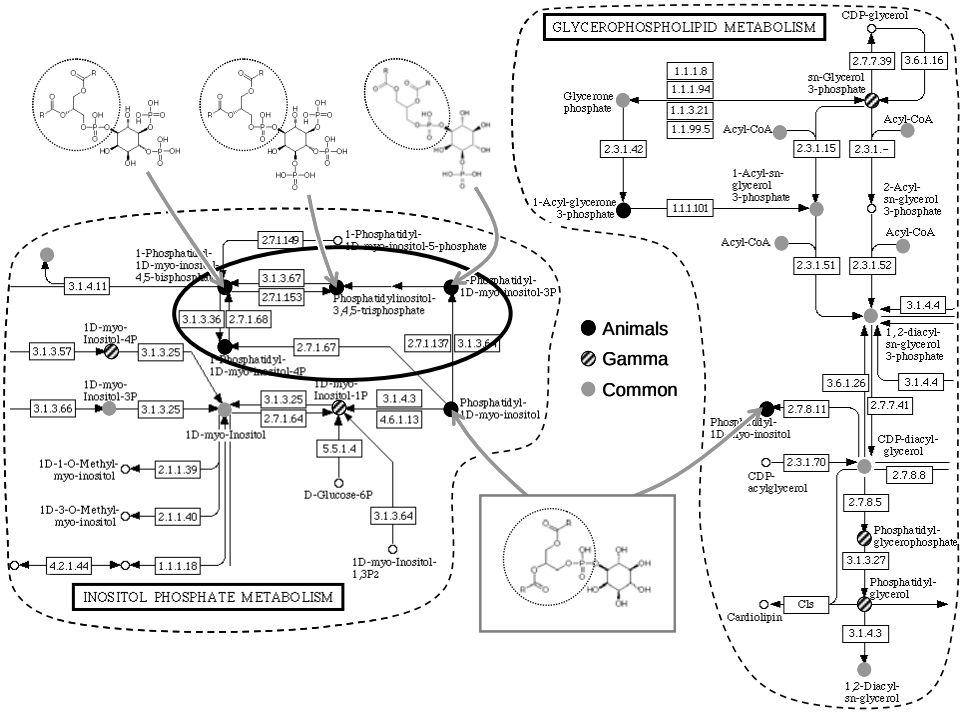

Fig. 2. An example of the set of structurally conserved compounds

The figure shows the crossover region between two metabolic pathways, the "Inositol phosphate metabolism" and the "Glycerophospholipid metabolism", represented by a dotted line. Here, Glycerophospholipid metabolism is a subcategory of Lipid metabolism and Inositol phosphate metabolism is that of Carbohydrate metabolism. These two pathways are connected by phosphatidyl-1D-myo-inositol, indicated in a square and arrows. The set of specifically used chemical compounds by Animals are also emphasized by a bold oval, indicating the observed common substructures on each chemical structure.

2.3. *Common substructures of organism-specific metabolites*

In order to know the biological meaning of organism-specific metabolic pathways, we demonstrated the set of differently used chemical compounds between two organism groups on metabolic pathway maps.

The phosphatidyl-1D-myo-inositol is one of the products of Glycerophospholipid metabolism and is specifically observed within three eukaryotes; Animals, Fungi, and Protists. It is also utilized in the Inositol phosphate metabolism (Fig. 2). In the subsequent path in the latter metabolism, we found that there was a common substructure containing a phosphatidyl group within the block of eukaryote-specific compounds, i.e., phosphatidyl-1D-myo-inositol derivatives. In contrast, the other derivatives, which have no phosphatidyl groups, have also been extracted from other organism groups, indicating that there may be other pathways for non-eukaryotic species.

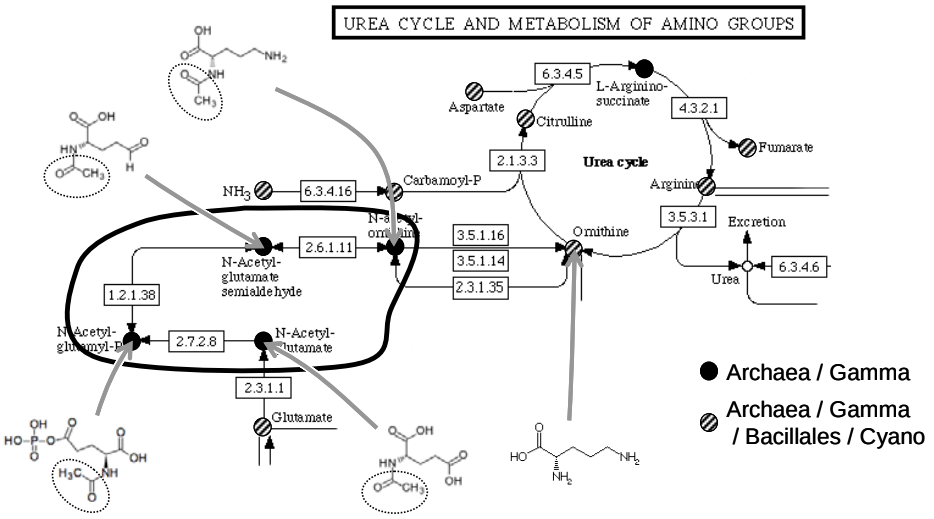

Fig. 3. Another example of the set of structurally conserved compounds
The figure shows a part of the metabolic pathway, "Urea Cycle and Metabolism of Amino Groups". The pathway is a part of Amino Acid metabolism. The set of specifically used chemical compounds by Archaea and gamma-Proteobacteria are surrounded by a bold line, and all acetyl groups are indicated by a dotted line.

In another case of the Urea cycle and metabolism of amino groups, illustrated in Fig. 3, we found a block of uniquely-used compounds. Within the block four glutamate derivatives comprise a series of paths. All of these compounds contain an acetyl group. The gamma-Proteobacteria, Fungi, alpha-Proteobacteria and Archaea possess all of these four chemical compounds, while Cyanobacteria, Protists, Bacillales and Spirochete have none of the four compounds. In this sub-pathway there are 2 terminals; N-acetylglutamate and N-acetylornithine. Animals have N-acetyl-ornithine; however, they do not have the other three compounds. The existence of a common substructure within the block suggests that the enzyme that adds an acetyl group to a precursor makes not only its product but also the following pathway components. In addition, the alpha-

Proteobacteria do not contain other components of a urea cycle, and are able to fill the gap of the pathway in Cyanobacteria or Bacillales.

3. Discussion

We demonstrated that there were specific common substructures within the uniquely used compounds on metabolic pathways. In this paper, two common substructures are closely depicted; one is the phosphatidyl group of phosphatidyl-1D-myo-inositol in the Inositol phosphate metabolism and the other is acetyl group of the N-acetylglutamate in the Urea cycle and metabolism of amino groups. In the former case, the organism groups are separated into two classes according to the possession of phosphatidyl derivatives. The first class contains three organism groups of eukaryotes, Animals, Fungi, and Protists, which possess all of phosphatidyl derivatives. The other organism groups, Bacteria and Archaea, comprise the second class and have no such derivatives. In the latter case, the organism groups are characterized by possession of N-acetylglutamate or N-acetylornithine, which are incorporated in the metabolisms of gamma-Proteobacteria, Fungi, alpha-Protebacteria and Archaea. On the other hand, neither chemical compound was found to be utilized by Cyanobacteria, Protists, Bacillales and Spirochete. These results suggest that there are metabolite-mediated relationships between two different organism groups.

Because of the restrictions of biological resources, metabolic systems of each species should have evolved in a way that utilizes their available metabolites more and more efficiently in diverse ways. Therefore, the metabolisms of each species differ significantly from each other, and we can assume that there are both available and unavailable chemical compounds that are specifically observed in certain species. We can also suppose that such a restriction of chemical compounds may cause the symbiosis or the predator-prey interaction between species. In this study, we found that several pairs of organism groups are highly correlated in terms of utilizing metabolites and the specific substructures are conserved within the uniquely used compounds. In particular, the lipid-related metabolites are specifically observed in eukaryotic groups. It corresponds to the experimental fact that there are some eukaryote-specific pathways of lipid biosynthesis. Thus, those findings may also help us to better understand the relationships between organisms. Of course, the annotation qualities of complete genomes are now completed in each species and the information on metabolisms has not been fulfilled. However, this type of research should be more important in this post-genomic era in order to understand the biological meaning of metabolites and metabolisms; hence we believe our current study may also contribute to solving such a grand challenge.

4. Materials and Methods

4.1. *Datasets on metabolic pathways and metabolites*

We obtained the information on enzymatic reactions and their reactants belonging to each organism from the KEGG PATHWAY database (version 42.0 + update 2007/04/20), which is comprised of a series of XML files written in KGML (KEGG Markup Language) at the KEGG FTP site; ftp://ftp.genome.jp/pub/kegg/xml/. We also obtained the chemical compound structures from the COMPOUND section of the KEGG LIGAND database (version 42.0 + update 2007/04/20), which contains information on 12,338 chemical compounds with 2D graph representations of structures in MDL/MOL format. However, many of the chemical compounds in COMPOUND are not always assigned to metabolic pathways. In this study, we used only the 3,499 metabolites that appear on metabolic pathway maps.

4.2. *Phylogenetic groups of organisms and their specific metabolites*

All 239 species that have complete genomes have been collected in the KEGG GENES database; however we decided to first classify all organisms into several organism groups and treat each group as one collective organism because of the limitations of available gene annotation data. When compiling the set of metabolites that each organism group possesses we estimate that a chemical compound should be included when over 80% of the genomes in the group possess a gene for the relevant enzyme. Also, we estimated that the chemical compound should be specific to the organism group when less than 20% of the genomes in other groups possess it. Here, the total number of organism groups is set to 9, and each group is identified using the following labels; Animals, Fungi, Protists, Archaea, alpha-Proteobacteria, gamma-Proteobacteria, Bacillales+Lactobacillales, Spirochete, Cyanobacteria, as illustrated in Fig. 1. These categorizations come originally from the definition of species categorization used in KEGG.

4.3. *Commonly used or uniquely used compounds*

After obtaining the organism group-specific metabolites, we performed a correlation to check the commonly used chemical compounds between two organism groups as well as the metabolites that appear only within one of the groups. This information is represented on each pathway map, by coloring the commonly used compounds in green and the uniquely possessed compounds in blue or yellow, to easily visualize their distribution on pathway maps. Then we carefully checked all the colored pathways for interesting regions and elucidated the common substructures specifically observed within those regions of pathway maps, using SIMCOMP. SIMCOMP can compute the pairwise atom alignments between two chemical compound structures. Here,

we calculated the alignments for all possible pairs among the concerned set of chemical compounds and then extracted the commonly aligned area as the common substructures.

Acknowledgments
This work was supported by the 21st Century COE Program "Genome Science" and a grant-in-aid for scientific research on the priority area "Comprehensive Genomics", both from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and by the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency. The computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Aguilar, D., Aviles, F.X., Querol, E., Sternberg, and M.J., Analysis of phenetic trees based on metabolic capabilities across the three domains of life, *J. Mol. Biol.*, 340(3):491-512, 2004.
[2] Brooksbank, C., Cameron, G., and Thornton, J., The European Bioinformatics Institute's data resources: towards systems biology, *Nucleic Acids Res.*, 33(Database issue):D46-53, 2005.
[3] Feldman, H.J., Dumontier, M., Ling, S., Haider, N., and Hogue, C.W., CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules, *FEBS Lett.*, 579(21):4685-4691, 2005.
[4] Friedrich, C.G., Physiology and genetics of sulfur-oxidizing bacteria, *Adv. Microb. Physiol.*, 39:235-289, 1998.
[5] Goto, S., Nishioka, T., and Kanehisa, M., LIGAND: chemical database for enzyme reactions, *Bioinformatics*, 14(7):591-599, 1998.
[6] Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M., Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways, *J. Am. Chem. Soc.*, 125:11853-11865, 2003.
[7] Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M., Heuristics for chemical compound matching, *Genome Inform.*, 14:144-153, 2003.
[8] Humphery-Smith, I. and Blackstock, W., Proteome analysis: genomics via the output rather than the input code, *J. Protein. Chem.*, 16(5):537-544, 1997.
[9] Kanehisa, M. and Goto, S., KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, 28(1):27-30, 2000.
[10] Kotera, M., Okuno, Y., Hattori, M., Goto, S., and Kanehisa, M., Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions, *J. Am. Chem. Soc.*, 126(50):16487-16498, 2004.
[11] Oliver, S.G., Winson, M.K., Kell, D.B., and Baganz, F., Systematic functional analysis of the yeast genome, *Trends Biotechnol.*, 16(9):373-378, 1998.
[12] Rison, S.C. and Thornton, J.M., Pathway evolution, structurally speaking, *Curr. Opin. Struct. Biol.*, 12(3):374-382, 2002.
[13] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P.,

Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci. USA*, 102(43):15545-15550, 2005.