Quantitative Biology 2013, 1(3): 192–200
DOI 10.1007/s40484-013-0019-x

MINI REVIEW

Automated interpretation of metabolic capacity from genome and metagenome sequences

Minoru Kanehisa*

Institute for Chemical Research, Kyoto University, Uji Kyoto 611-0011, Japan
* Correspondence: kanehisa@kuicr.kyoto-u.ac.jp

Received August 15, 2013; Accepted August 20, 2013

The KEGG pathway maps are widely used as a reference data set for inferring high-level functions of the organism or the ecosystem from its genome or metagenome sequence data. The KEGG modules, which are tighter functional units often corresponding to subpathways in the KEGG pathway maps, are designed for better automation of genome interpretation. Each KEGG module is represented by a simple Boolean expression of KEGG Orthology (KO) identifiers (K numbers), enabling automatic evaluation of the completeness of genes in the genome. Here we focus on metabolic functions and introduce reaction modules for improving annotation and signature modules for inferring metabolic capacity. We also describe how genome annotation is performed in KEGG using the manually created KO database and the computationally generated SSDB database. The resulting KEGG GENES database with KO (K number) annotation is a reference sequence database to be compared for automated annotation and interpretation of newly determined genomes.

Keywords: metabolic pathway; functional module; genome annotation; genome interpretation; KEGG database

INTRODUCTION

The hierarchy from data to information to knowledge, and sometimes further to wisdom, has been used as a conceptual framework for understanding structural and functional relationships among them in computer sciences. This concept is relevant in life sciences as well except the top level, which may be modified to "principle" of life that is beyond human wisdom. Thanks to the continuous development of high-throughput experimental technologies, big data in genomics and other areas of life sciences are routinely generated. There are increasing needs for better bioinformatics technologies to process and interpret the data, which can be considered in the hierarchy of data, information, and knowledge. Computational methods play a major role in the initial processing of big data to extract information, but they become less effective to compile knowledge from information. In our definition, a gene-disease association extracted from data are information, while a molecular mechanism uncovered for this disease is knowledge. The latter process requires more detailed analysis including low-throughput but high-quality experiments, manual works, and human intuition. Therefore, it is extremely important to capture this type of knowledge accumulated in literature and develop a bioinformatics resource that can be integrated into the pipeline of big data analysis.

Notable examples of such resources include Kyoto Encyclopedia of Genes and Genomes (KEGG) [1] and Gene Ontology (GO) [2]. The KEGG resource (http://www.kegg.jp/) that we have been developing since 1995 contains, among others, accumulated knowledge on metabolism, other cellular processes, organismal systems, human diseases and drugs represented as networks of molecular interactions, reactions, and relations. This knowledge base is then used as a reference for biological interpretation of genome sequences and other large-scale data. For example, a set of genes identified in the genome is matched against KEGG pathway maps, which would reveal metabolic and other features of the organism. This KEGG pathway mapping analysis as well as the similar

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2013

GO enrichment analysis usually require manual inspection of the results to determine which KEGG pathways are present and which GO categories are enriched. Here we focus on metabolism and introduce better automated methods using KEGG modules and reaction modules.

MODULAR ARCHITECTURE OF THE METABOLIC NETWORK

Genetic and chemical units

It has been suggested that the metabolic network has a modular architecture containing functional modules [3–6]. Here we do not attempt to characterize the entire metabolic network; rather we identify and characterize individual functional modules for the specific purpose of genome annotation and interpretation. There are two types of modules, which are defined as genetic units and chemical units. The metabolic network may be viewed as a network of enzymes or a network of enzyme genes that encode enzymes. The genetic unit was first identified as a set of enzyme genes encoded in an operon-like structure on the genome that corresponds to a set of enzymes catalyzing consecutive reaction steps in the metabolic pathway [7]. The current collection of more extended genetic units, called KEGG modules, is stored in the KEGG MODULE database. KEGG modules are manually defined by considering not only the operon-like structures but also how well certain pathways are conserved among certain organism groups, together with knowledge of specific pathways.

The chemical units in the metabolic network are called reaction modules, which are identified from purely chemical analysis without using any information about enzymes and enzyme genes, namely, as conserved sequences of chemical structure transformation patterns of small molecules in consecutive reaction steps [8,9]. Reaction modules are extracted both computationally and manually, and they are found to represent logical units of organic reactions, such as a sequence of reactions for the carboxylic acid chain elongation. Although KEGG modules and reaction modules are extracted from different types of information, they tend to coincide on the metabolic pathway suggesting inherent relationships between the genetic network (network of enzymes) and the chemical network (network of small molecules). Let us first examine some examples of KEGG modules and reaction modules before describing how they are defined and utilized.

KEGG modules for citrate cycle

Organism-specific metabolic pathways in KEGG are computationally generated by matching the genomic content of enzyme genes against the manually drawn reference pathway maps. Figure 1 shows examples of citrate cycle (TCA cycle) for *Escherichia coli* (KEGG organism code: eco), *Helicobacter pylori* (KEGG organism code: hpy), and *Hemophilus influenzae* (KEGG organism code: hin), where green boxes represent enzymes whose genes are identified in the genome. The cycle is complete for *E. coli* (Figure 1A) because green boxes cover all necessary reaction steps, but it is incomplete for *H. pylori* (Figure 1B) and *H. influenzae* (Figure 1C), which lack certain enzymes. It is interesting to note that the pattern of missing enzymes is complementary, *H. pylori* having only the first segment from oxaloacetate to 2-oxoglutarate and *H. influenzae* having only the second segment from 2-oxoglutarate to oxaloacetate. Therefore, the cycle appears to be formed by combining these two segments. Furthermore, genes that encode the first segment are sometimes found in an operon-like structure as in the case for *Pyrococcus furiosus* (KEGG organism code: pfu) (Figure 1D). The genes are found next each other on the genome for citrate synthase (PF0203) for catalyzing oxaloacetate and acetyl-CoA to form citrate, aconitate hydratase (PF0201) for a two-step reaction from citrate to isocitrate, and isocitrate dehydrogenase (PF0202) for another two-step reaction from isocitrate to 2-oxoglutarate [10]. These observations led us to define three KEGG modules for citrate cycle: M00009 for the entire cycle, M00010 for the first segment, and M00011 for the second segment. The completeness ratio of these modules was 32% for M00009, 72% for M00010, and 33% for M00011 according to the annotation of 2800 genomes in KEGG (as of August 2013). Thus, the first segment M00010 is by far the most frequently found functional unit.

Reaction module for 2-oxocarboxylic acid chain elongation

The first segment M00010 of citrate cycle corresponds to a characteristic reaction sequence involving tricarboxylic acids (TCAs), which effectively increases the chain length of 2-oxocarboxylic acids by one. This is defined as reaction module RM001, 2-oxocarboxylic acid chain extension by tricarboxylic acid pathway. 2-oxocarboxylic acids are an important class of precursor metabolites including pyruvate (2-oxopropanoate), oxaloacetate (2-oxosuccinate), and 2-oxoglutarate. As shown in Figure 2 reaction module RM001 is found in different pathways. The sequence of chemical structure transformation patterns is identical from oxaloacetate (four-carbon or C4) to 2-oxoglutarate (C5) in citrate cycle (Figure 2A) and from 2-oxoglutarate (C5) to 2-oxoadipate (C6) in

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2013

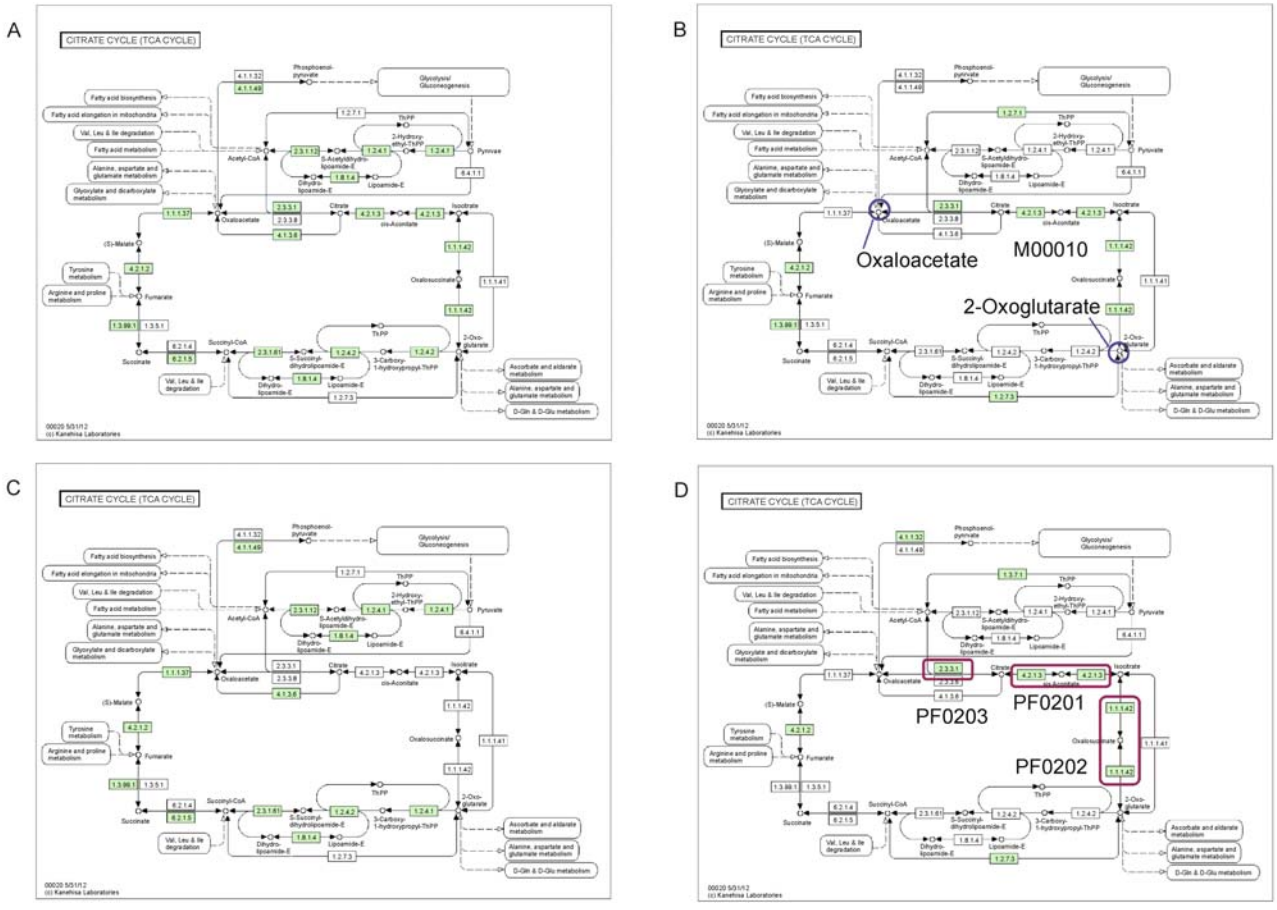

Figure 1. KEGG module M00010. The pathway maps of citrate cycle for (A) *Escherichia coli*, (B) *Helicobacter pylori*, (C) *Hemophilus influenzae*, and (D) *Pyrococcus furiosus* are generated by matching the genomic content of enzymes genes against the KEGG reference pathway map (http://www.kegg.jp/pathway/map00020). The KEGG module M00010 is defined for the first segment of citrate cycle from oxaloacetate to 2-oxoglutarate, because *H. pylori* has only this segment, *H. influenzae* lacks this segment, and *P. furiosus* has only this segment encoded by adjacent genes in an operon-like structure.

lysine biosynthesis (Figure 2B). The sequences are not identical but very similar from pyruvate (C3) to 2-oxobutanoate (C4) and from 2-oxoisovalerate (C5) to 2-oxoisocaproate (C6) in valine, leucine and isoleucine biosynthesis (Figure 2C).

The same reaction modules in different pathways are found to correspond to different KEGG modules. RM001 corresponds to M00010 from oxaloacetate to 2-oxoglutarate in citrate cycle, M00433 from 2-oxoglutarate to 2-oxoadipate in lysine biosynthesis, M00535 from pyruvate to 2-oxobutanoate in isoleucine biosynthesis, and M00432 from 2-oxoisovalerate to 2-oxoisocaproate in leucine biosynthesis. An obvious question would then be whether any relationship of genes exists among these KEGG modules matching the same reaction module. The answer is yes. There is a tendency that paralogous genes play roles of forming these related KEGG modules [8,9].

GENOME ANNOTATION AND INTERPRETATION

Overall procedure

Figure 3 illustrates an overall procedure of linking genomic information to higher-level knowledge of molecular networks for inferring metabolic capacities. The collection of 2800 KEGG organisms (as of August 2013) represents all high-quality, completely sequenced genomes available in RefSeq [11]. Starting from the gene set provided by RefSeq (or the authors) the KEGG Orthology (KO) identifiers (K numbers) are assigned to individual genes in the KEGG GENES database (currently 42% of 11 million genes in 2800 genomes are annotated). The K number annotated gene set is then compared against the collection of KEGG modules to identify the existence of specific pathways and further to

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2013

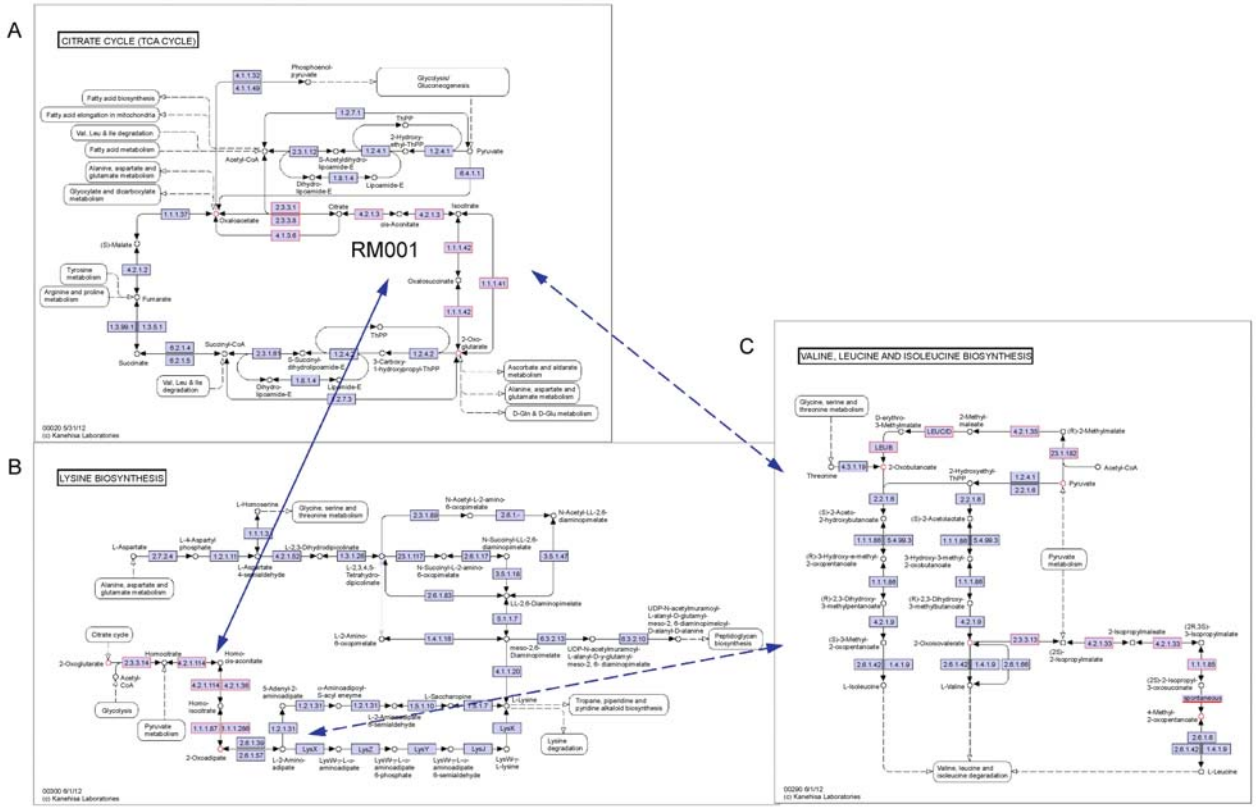

Figure 2. Reaction module RM001. The reaction module RM001 is a characteristic reaction sequence involving tricarboxylic acids for extension of 2-oxocarboxylic chain using acetyl-CoA derived carbon. It is shown in pink boxes (A) from oxaloacetate to 2-oxoglutaratein in citrate cycle, (B) from 2-oxoglutarate to 2-oxoadipate in lysine biosynthesis, (C) from pyruvate to 2-oxobutanoate and from 2-oxoisovalerate to 2-oxoisocaproate in valine, leucine and isoleucine biosynthesis.

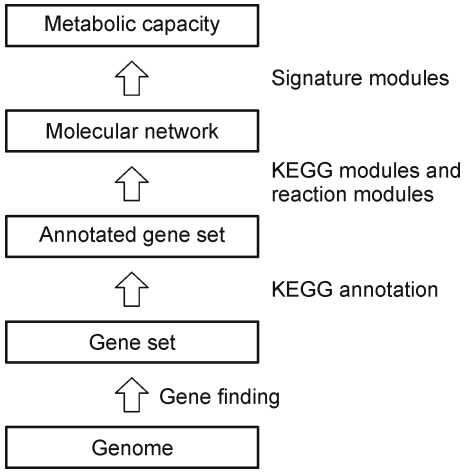

Figure 3. Linking genomes to molecular networks for interpretation of metabolic capacities.

characterize metabolic capacities. This is an ortholog based approach and there may be unannotated paralogs left in the genome, which can be examined with reaction modules to improve annotation. There is a hierarchy from individual genes of KO entries (K numbers) to molecular networks of KEGG modules (M numbers) to phenotypes of metabolic capacities. To improve automation of the last step, signature modules are under development. A signature module is defined by a combination of M numbers and associated with a text definition of phenotype. Thus, once a newly determined genome is automatically annotated (K number assigned) by comparing against the KEGG GENES database, subsequent mapping procedures against KEGG modules and signature modules will enable automated interpretation of metabolic capacities.

KEGG Orthology (KO)

The KEGG pathway maps, as well as the BRITE functional hierarchies (ontologies) and the KEGG modules, are represented in a generic way to be applicable to all organisms. This allows manually created reference pathways to be computationally expanded to organism-

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2013

specific pathways by combining with genome annotation data. The KEGG Orthology (KO) system is the basis for this representation and expansion, consisting of manually defined ortholog groups (KO entries) for all proteins and functional RNAs that appear in the KEGG pathway maps, BRITE functional hierarchies, and KEGG modules. Each KO entry (identified by K number) also corresponds to a sequence similarity group in order to allow computational assignment in newly determined genomes. When, for example, a pathway map is drawn based on experimental observations in specific organisms, an additional work is performed for generalizing gene information from those specific organisms to other organisms. This is done by associating map objects (boxes) to KO entries and, when necessary, by defining a new KO entry and creating a corresponding set of orthologous genes from available genomes. This process ensures that the degree of sequence similarity is defined in a context (pathway) dependent manner for each KO entry.

Genome annotation

There are two key databases for the genome annotation (K number assignment) in KEGG. One is the manually created KEGG ORTHOLOGY (KO) database for ortholog grouping and for accumulating experimental evidence on known functions of genes and proteins. The other is computationally generated KEGG SSDB database, which contains sequence similarity scores and best-hit relations computed from the KEGG GENES database by pair wise genome comparisons using the SSEARCH program. SSDB can thus be viewed as a huge weighted, directed graph of genes. The sequence similarity group of each KO entry corresponds to a subgraph in the SSDB graph, and the genome annotation involves extending and modifying this subgraph. Therefore, the genome annotation in KEGG is essentially cross-species annotation finding members (orthologous genes) of each KO entry in all available genomes. As the result, K numbers are assigned to individual genes in each genome.

This annotation procedure is highly computerized as follows. For each gene in a genome the GFIT (Gene Function Identification Tool) table is created from SSDB detailing the information about best-hit genes, including paralogs, in all other genomes. The KOALA (KEGG Orthology And Links Annotation) tool, which contains human annotators' knowledge, processes all the GFIT tables at a time and makes computational K number assignments for all genes. KOALA's computational assignments are automatically reflected for a well-curated set of K numbers (currently 73%) in a newly determined genome, and also in the existing genomes that meet various other criteria. GFIT tables and KOALA's computational assignments are continuously updated (currently three times a week). There are facilities that alert human intervention. Discrepancies between KOALA's assignments and current annotations are examined by annotators with the manual version of KOALA and GFIT tools. Discrepancies may also lead to regrouping of KO entries, including split, merge, and new addition.

KEGG module

The KEGG module is a tighter functional unit of molecules often corresponding to a subpathway of the KEGG pathway map, which depicts a large network of molecular interactions and reactions. Each KEGG module (identified by M number) is manually defined as a combination of K numbers based on the knowledge of pathways and using the information of gene conservation patterns among organism groups and positional correlations of genes (operon-like structures). The definition is a simple Boolean expression allowing automatic evaluation of whether the gene set is complete, i.e., the module is present. For example, M00010 in citrate cycle and M00433 in lysine biosynthesis are defined as:

M00010: K01647 (K01681,K01682) (K00031,K00030)

M00433: (K01655,K10977) (K17450 K01705,K16792 +K16793) (K05824,K10978)

where a comma sign represents OR and a space or a plus sign (for a molecular complex) represents AND when evaluating this expression. The automatic evaluation is implemented in the genome annotation in KEGG revealing incomplete modules that can be complete when a few more genes are correctly annotated.

Reaction class

The nodes (boxes) of the KEGG metabolic pathway maps are linked not only to K numbers of KO entries but also to R numbers of KEGG REACTION entries in order to represent both the genetic and chemical networks. The KEGG REACTION database contains all known biochemical reactions taken from the KEGG metabolic pathway maps, as well as experimentally characterized enzymatic reactions with the official EC numbers in the Enzyme Nomenclature [12]. The reaction is represented by a reaction formula, which is an equation involving multiple substrates and products. On the KEGG pathway maps this is simplified to indicate only the conversion of main substrates and products along the pathway. The simplified reaction of main substrates and products is represented by a reaction map formula, which is pathway-dependent, and stored in the KEGG REACTION database. Figure 4 shows an example. The reactions of citrate synthase (R00351) and homocitrate synthase

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2013

(R00271) are simplified in the reaction map formulas excluding water and CoA, and sometimes acetyl-CoA as well.

Figure 4 shows additional processing of reaction data in KEGG [8]. The reaction formula is decomposed into a set of reactant pairs, one-to-one relationships of substrate-product pairs, by considering the reaction type and the flow of atoms. Each reactant pair is characterized by the chemical structure transformation pattern, called RDM pattern of KEGG atom type changes. Among the main reactant pairs that appear in the KEGG pathway maps, distinct RDM patterns are used to define reaction class. The resulting KEGG reaction class (identified by RC number) is like an ortholog group of reactions defined by functionally important local structural changes and accommodating global structural differences of reactants. The main reactant pairs of RP00177 (oxaloacetate to citrate) and RP04506 (2-oxoglutarate to homocitrate) in Figure 4 belong to the same reaction class RC00067, accommodating the chain length difference.

Reaction module

Using the reaction class information, a systematic survey was performed both computationally and manually to extract conserved reaction sequence patterns, represented by RC number sequences, from all known metabolic pathways in KEGG [8]. As already mentioned, reaction modules (also called RC modules) tend to correspond to KEGG modules (also called KO modules) despite the fact that they are separately defined from different properties. For example, the reaction module RM001 is defined by three subtypes:

(RC00067 (RC00498+RC00618,RC00497)
(RC00084+RC00626,RC00114)),

(RC01205 RC00976+RC00977 RC00417),

(RC00470 RC01041+RC01046 RC00084+RC00577),

and the first subtype corresponds to the KEGG modules M00010 and M00433, the second to M00535, and the third to M00432. This definition is similar to the simple Boolean expression for the KEGG module, a comma sign representing OR and a space or a plus sign (for a multi-step reaction) representing AND. However, it is more complex because of the existence of subtypes caused by the fact that reaction class is too finely classified. Reaction modules are defined by considering the similarity grouping of RC entries [8], for example, RC00067, RC01205, and RC00470 are the same in this grouping.

Reaction modules for improving annotation

In a sense genome annotation is a prediction, extending limited experimental evidence in certain organisms to many other organisms using sequence similarity, protein

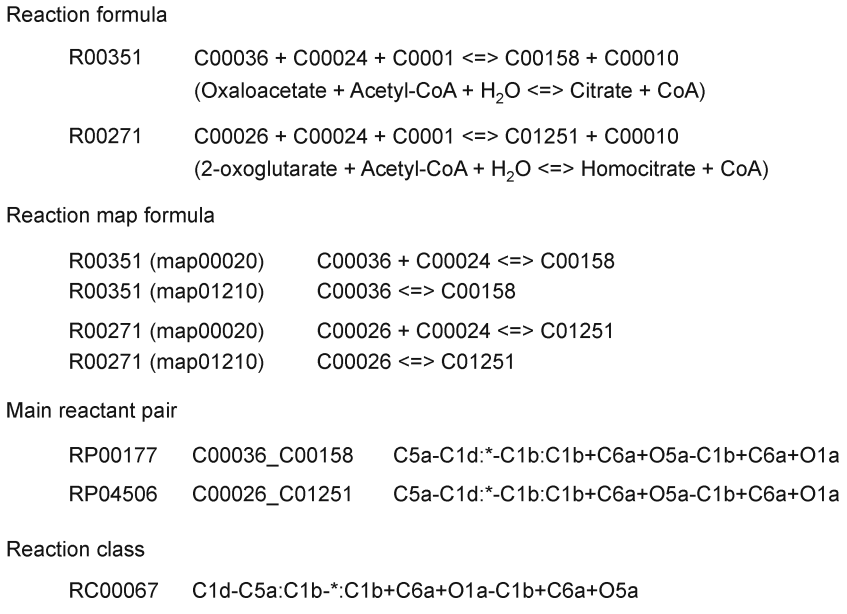

Figure 4. An example of the reaction data processing in KEGG. The KEGG pathway map does not contain all substrates and products defined in the reaction formula. Instead, it uses a simplified representation as shown in the reaction map formula for main compounds. Reactant pairs are defined from the reaction formula as one-to-one relationships of substrate-product pairs. Those on the pathway map are called main reactant pairs. Reaction class corresponds to a set of main reactant pairs that have the same chemical structure transformation patterns defined as RDM patterns.

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2013

domains, and other information. The accuracy of prediction will increase when individually predicted genes are found to come together to form a functional unit. This is exactly the purpose of developing KEGG modules, which can be used to check if functional units are complete. Reaction modules have also turned out to be a useful resource for improving KO grouping and module definition.

Figure 5 shows a global picture of 2-oxocarboxylic acid metabolism (map01210). First, the chain extension module RM001 appears not only in citrate cycle (map00020), lysine biosynthesis (map00300), and valine, leucine and isoleucine biosynthesis (map00290), but also in glucosinolate biosynthesis (map00966) and coenzyme B biosynthesis (map00680). Second, the chain extension module is used in combination with the chain modification modules, as well as the reductive amination step RC00006 or RC00036, which itself involves conversions between 2-oxocarboxylic acids and between amino acids. The carboxyl to amino conversion modules RM032 and RM002, which differ in the use of protective N-acetyl group, are found in basic amino acid biosynthesis. The branched-chain addition module RM0033 is found in branched-chain amino acid biosynthesis. The module RM030 is found in the biosynthesis of glucosinolates, a class of plant secondary metabolites. Furthermore, the chain extension from 2-oxoadipate to 2-oxosuberate is followed by coenzyme B biosynthesis in methonogenic archaea.

The reaction modules and KEGG module have been used for improving genome annotation. The reaction module RM001 corresponds to KEGG modules M00010 (observed in 72% of 2800 genomes), M00433 (4.6%), M00608 (1.7%), M00535 (4.6%), and M00432 (69%), which share paralogous gene sets [8,9]. M00606 defined as (K10977 K16792 + K16793 K10978) is for chain extensions from 2-oxoglutarate to 2-oxoadipate and further to 2-oxosuberate in methanogenic archaea, which are apparently catalyzed by the same set of enzymes [13–15]. According to our annotation, which can be examined from the Ortholog table link of M00606, the synthase K10977 (aksA) is limited to methanogenic archaea, but the dehydratase complex, K16792 (aksD) and K16793 (aksE), is also found in Chloroflexi (green non-sulfur bacteria) and Deinococcus-Thermus. As summarized in Table 1, these organism groups use yeast enzyme orthologs, homocitrate synthase K01655 (LYS21) and homoisocitrate dehydrogenase K05824 (LYS12), to make the module M00433 complete in lysine biosynthesis. Note that yeast homoaconitase K01705 (LYS4) in lysine biosynthesis apparently catalyzes only the second step of the dehydratase reaction and a paralog of aconitase from citrate cycle catalyzes the first step [16]. This observation resulted in the regrouping of K01681 (ACO), from which a tighter group K17450 (ACO2) is defined. The dehydrogenase K10978 (aksF) in M00606 is

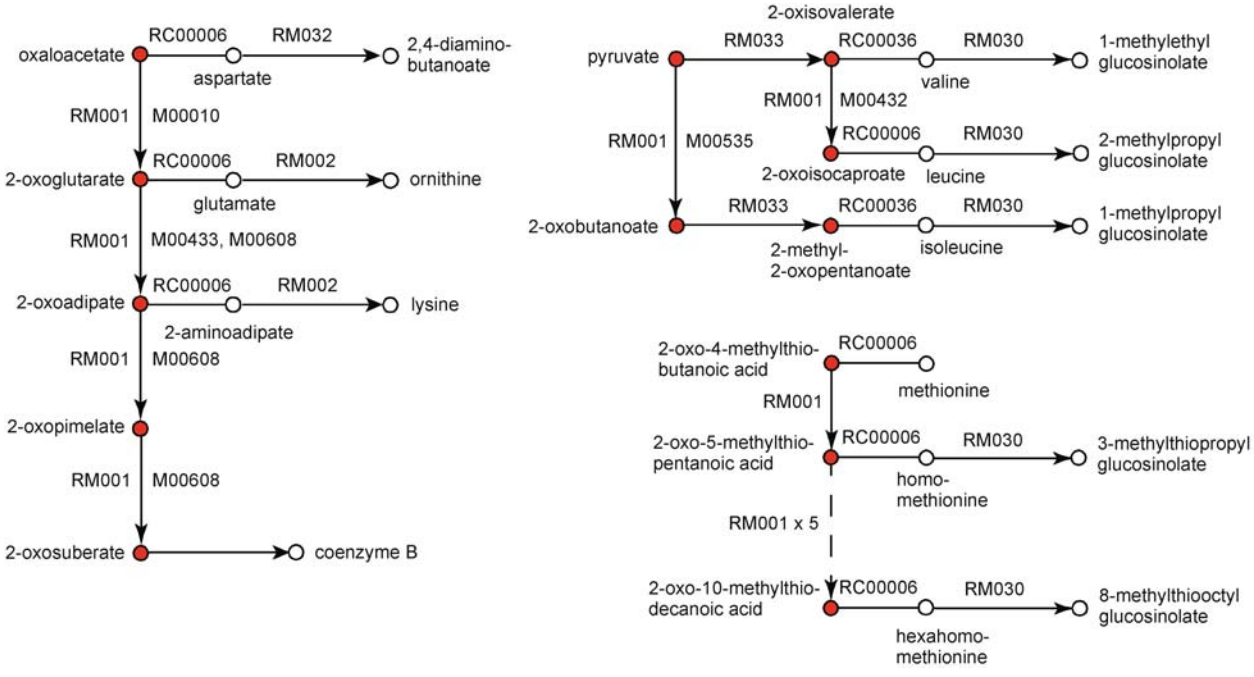

Figure 5. Modular architecture of 2-oxocarboxylic acid metabolism. This is a simplified version of the KEGG pathway map (http://www.kegg.jp/pathway/map01210), where 2-oxocarboxylic acids are denoted by red circles. The chain elongation module RM001 is shown in the vertical direction, and chain modification modules and other reactions are shown in the horizontal direction. The correspondence of RM001 to KEGG modules (M00010, etc.) is also shown.

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2013

Table 1. Combination of paralogous genes in 2-oxocarboxylic acid chain extension

Module	Pathway	Organism group	Synthase	Dehydratase	Dehydrogenase
M00433	Lysine	Fungi	K01655 (LYS21)	K17450 (ACO2)K01705 (LYS4)	K05824 (LYS12)
M00433	Lysine	Green non-sulfur bacteria Deinococcus-Thermus	K01655 (LYS21)	K16792 (aksD) + K16793 (aksE)	K05824 (LYS12)
M00608	Lysine, Coenzyme B	Methanogenic archaea	K10977 (aksA)	K16792 (aksD) + K16793 (aksE)	K10978 (aksF)
M00432	Leucine	Pyrococcus	K01649 (leuA)	K01703 (leuC) + K01704 (leuD)	K00052 (leuB) K10978 (aksF)

mostly limited to methanogenic archaea, but similar genes are found in Pyrococcus possibly playing a role in leucine biosynthesis. These examples indicate extensive mixing of paralogous genes to form functional units. Genome annotation of these paralogs can be improved by comparative evaluation of related KEGG modules.

Signature modules for inferring metabolic capacity

Signature modules were originally conceived as a class of KEGG modules that link genomic features to phenotypic features. However, the use of a single KEGG module (M number) was too restrictive, and a new version of signature modules is under development. A signature module is now defined by a combination of M numbers, although it may also be a combination of M numbers and K numbers, a single M number, or a single K number. Table 2 shows some examples of signature modules that indicate metabolic capacities. Four types of carbon fixation is distinguished by combining different carbon fixation pathways [17] and the molecular complexes of oxygenic and anoxygenic photosystems. For nitrate assimilation [18] and sulfate assimilation [19] reduction pathways are combined with transporters. Methanogenesis [20] and acetogenesis [17] are represented by pathway features.

The metabolic capacities, especially those reflecting the environment of Earth, are limited to specific organism groups. The organisms in the KEGG GENOME database are now given metadata annotations, such as metabolic capacity, pathogenicity, and where they are isolated, based on literature information. In a similar way that gene functions are expanded from experimentally verified organisms to other organisms by KO entries, phenotypic features can be expanded from well characterized organisms to other organisms by signature modules.

Table 2. Examples of signature modules

Metabolic capacity	Signature module	Definition
Carbon fixation in plants and cyanobacteria	(M00161,M00163)+ M00165	M00161 Photosystem II M00163 Photosystem I M00165 Reductive pentose phosphate cycle (Calvin cycle)
Carbon fixation in alphaproteobacteria	M00597 M00165	M00597 Anoxygenic photosystem II M00165 Reductive pentose phosphate cycle (Calvin cycle)
Carbon fixation in green nonsulfur bacteria	M00597 + M00376	M00597 Anoxygenic photosystem II M00376 3-hydroxypropionate bi-cycle
Carbon fixation in green sulfur bacteria	M00598 + M00161	M00598 Anoxygenic photosystem I M00173 Reductive citrate cycle (Arnon-Buchanan cycle)
Nitrate assimilation	(K02575,M00438)+ M00531	K02575 MFS transporter, NNP family, nitrate/nitrite transporter M00438 ABC transporter, nitrate/nitrite transport system M00531 Assimilatory nitrate reduction, nitrate = > ammonia
Sulfate assimilation	(K14708,M00185)+ M00176	K14708 SLC family 26, sodium-independent sulfate transporter M00185 ABC transporter, sulfate transport system M00176 Assimilatory sulfate reduction, sulfate = > H2S
Methanogenesis	M00567,M00357,M00356,M00563	M00567 Methanogenesis, CO2 = > methane M00357 Methanogenesis, acetate = > methane M00356 Methanogenesis, methanol = > methane M00563 Methanogenesis, methylamine = > methane
Acetogenesis	M00377 + M00579	M00377 Reductive acetyl-CoA pathway (Wood-Ljungdahl pathway) M00579 Phosphate acetyltransferase-acetate kinase pathway

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2013

DISCUSSION

Biological science has been a data-driven science. This is bound to change with more data generated, more information extracted, and more knowledge accumulated. Biological science will eventually become, at least partially, first-principle based (see Introduction). Metabolism is the most likely subject for this approach to be effectively applied because of the amount of knowledge already accumulated and chemical logic involved. The reaction modules that have been identified contain interesting features, possibly Nature's design principles of a series of organic reactions, including how to achieve an activated transition state (e.g., phosphorylation), how to introduce a protective group (e.g., N-acetylation), how to increase specificity and efficiency (e.g., using carrier protein and switching a carbon source from acetyl-CoA to malonyl-CoA) [8,9]. These design principles, together with architectural design principles of how to reuse the same module and how to combine with other modules, may have enabled increased complexity of the metabolic network in the evolution of both the chemical network and the genetic network.

We have recently introduced a new category of KEGG pathway maps, called overview maps (map numbers 01200s) to present such design principles of the metabolic network rather than traditional views of individual pathways. The modular architecture of the metabolic network is apparent in 2-oxocarboxylic acid metabolism, fatty acid metabolism, and degradation of aromatic compounds, but the core portion of carbon metabolism (map01200) seems to contain a different design principle. It is an extensive use of the same parts with minor modifications [9]. The new overview maps reflect our efforts to understand "principle of life" from accumulated knowledge.

ACKNOWLEDGMENTS

This work was partially supported by the Japan Science and Technology Agency. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

CONFLICT OF INTEREST

The author Minoru Kanehisa declares that no conflict of interest exists.

REFERENCES

1. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res., 40, D109–D114.
2. Gene Ontology Consortium. (2013) Gene Ontology annotations and resources. Nucleic Acids Res., 41, D530–D535.
3. Papin, J. A., Reed, J. L. and Palsson, B. O. (2004) Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. Trends Biochem. Sci., 29, 641–647.
4. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabási, A. L. (2002) Hierarchical organization of modularity in metabolic networks. Science, 297, 1551–1555.
5. Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. and Dandekar, T. (2002) Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae*. Bioinformatics, 18, 351–361.
6. Yamada, T., Kanehisa, M. and Goto, S. (2006) Extraction of phylogenetic network modules from the metabolic network. BMC Bioinformatics, 7, 130.
7. Ogata, H., Fujibuchi, W., Goto, S. and Kanehisa, M. (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic Acids Res., 28, 4021–4028.
8. Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S. and Kanehisa, M. (2013) Modular architecture of metabolic pathways revealed by conserved sequences of reactions. J. Chem. Inf. Model., 53, 613–622.
9. Kanehisa, M. (2013) Chemical and genomic evolution of enzyme-catalyzed reaction networks. FEBS Lett., doi: 10.1016/j.febslet.2013.06.026.
10. Maeder, D. L., Weiss, R. B., Dunn, D. M., Cherry, J. L., González, J. M., DiRuggiero, J. and Robb, F. T. (1999) Divergence of the hyperthermophilic archaea *Pyrococcus furiosus* and *P. horikoshii* inferred from complete genomic sequences. Genetics, 152, 1299–1305.
11. Pruitt, K. D., Tatusova, T., Brown, G. R. and Maglott, D. R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res., 40, D130–D135.
12. McDonald, A. G., Boyce, S. and Tipton, K. F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. Nucleic Acids Res., 37, D593–D597.
13. Howell, D. M., Harich, K., Xu, H. and White, R. H. (1998) A-keto acid chain elongation reactions involved in the biosynthesis of coenzyme B (7-mercaptoheptanoyl threonine phosphate) in methanogenic Archaea. Biochemistry, 37, 10108–10117.
14. Drevland, R. M., Jia, Y., Palmer, D. R. and Graham, D. E. (2008) Methanogen homoaconitase catalyzes both hydrolyase reactions in coenzyme B biosynthesis. J. Biol. Chem., 283, 28888–28896.
15. Howell, D. M., Graupner, M., Xu, H. and White, R. H. (2000) Identification of enzymes homologous to isocitrate dehydrogenase that are involved in coenzyme B and leucine biosynthesis in methanoarchaea. J. Bacteriol., 182, 5013–5016.
16. Fazius, F., Shelest, E., Gebhardt, P. and Brock, M. (2012) The fungal α-aminoadipate pathway for lysine biosynthesis requires two enzymes of the aconitase family for the isomerization of homocitrate to homoisocitrate. Mol. Microbiol., 86, 1508–1530.
17. Berg, I. A., Kockelkorn, D., Ramos-Vera, W. H., Say, R. F., Zarzycki, J., Hügler, M., Alber, B. E. and Fuchs, G. (2010) Autotrophic carbon fixation in archaea. Nat. Rev. Microbiol., 8, 447–460.
18. Ohashi, Y., Shi,W., Takatani, N., Aichi, M., Maeda, S., Watanabe, S., Yoshikawa, H. and Omata, T. (2011) Regulation of nitrate assimilation in cyanobacteria. J. Exp. Bot., 62, 1411–1424.
19. van der Ploeg, J. R., Eichhorn, E. and Leisinger, T. (2001) Sulfonate-sulfur metabolism and its regulation in *Escherichia coli*. Arch. Microbiol., 176, 1–8.
20. Liu, Y. and Whitman, W. B. (2008) Metabolic, phylogenetic, and ecological diversity of the methanogenic archaea. Ann. N. Y. Acad. Sci., 1125, 171–189.

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2013