FEBS Letters 587 (2013) 2731–2737

journal homepage: www.FEBSLetters.org

Review

Chemical and genomic evolution of enzyme-catalyzed reaction networks

Minoru Kanehisa *

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

ARTICLE INFO

Article history:
Received 5 June 2013
Revised 18 June 2013
Accepted 18 June 2013
Available online 28 June 2013

Edited by Athel Cornish-Bowden

Keywords:
Evolution of metabolism
Metabolic pathway
Enzyme cluster
Reaction module
KEGG database

ABSTRACT

There is a tendency that a unit of enzyme genes in an operon-like structure in the prokaryotic genome encodes enzymes that catalyze a series of consecutive reactions in a metabolic pathway. Our recent analysis shows that this and other genomic units correspond to chemical units reflecting chemical logic of organic reactions. From all known metabolic pathways in the KEGG database we identified chemical units, called reaction modules, as the conserved sequences of chemical structure transformation patterns of small molecules. The extracted patterns suggest co-evolution of genomic units and chemical units. While the core of the metabolic network may have evolved with mechanisms involving individual enzymes and reactions, its extension may have been driven by modular units of enzymes and reactions.

1. Introduction

Leonor Michaelis was a professor of Aichi Medical College (currently Nagoya University School of Medicine) in Japan from late 1922 to early 1926. His name associated with enzyme kinetics [1] is well recognized, but the fact that the actual person spent three years in Nagoya is no longer widely known among Japanese scientists. Nevertheless, this review is a tribute to his presence in Japan, especially to his contribution to the early days of Japanese biochemistry [2]. In 1995 we started the Kyoto Encyclopedia of Genes and Genomes (KEGG) database project under the then ongoing Human Genome Program in Japan. The original concept was to create a reference knowledge base of metabolism and other cellular processes from published literature, so that it can be used for biological interpretation of genome sequence data. The KEGG database has expanded significantly over the years to meet the needs for integrating and interpreting large-scale datasets generated by various types of high-throughput experimental technologies [3], but this basic concept is unchanged. At first the KEGG metabolic pathway maps were created using the book "Metabolic Maps" [4] compiled by the Japanese Biochemical Society. This Society was founded in 1925 during Michaelis' stay in Japan, and the biochemistry of enzymes was an active field since then. The original KEGG that owes to this tradition still remains in the metabolic pathway section of the KEGG PATHWAY database. The KEGG pathway map identifiers such as map00010, map00020, and map00030 for glycolysis, citrate cycle, and pentose phosphate pathway correspond to the map numbers 1, 2, and 3 in the Japanese Biochemical Society's Metabolic Maps.

Since its inception the KEGG metabolic pathway map is drawn to represent two types of networks: the chemical network of how small molecules are converted and the genomic network of how genome-encoded enzymes are connected to catalyze consecutive reactions. This dual aspect has been utilized for metabolic reconstruction. A set of enzyme genes encoded in the completely sequenced genome will identify enzyme relation networks when superimposed on the KEGG pathway maps, which in turn characterize chemical structure transformation networks allowing interpretation of biosynthetic and biodegradation potentials of the organism. In addition to this type of genome analysis, the KEGG metabolic pathway maps can be used for chemical analysis of small molecules and reactions [5–8]. This review focuses on our efforts to integrate genomics and chemistry toward better understanding of intrinsically related genome evolution and chemical evolution of enzyme-catalyzed reactions.

2. The KEGG resource

2.1. KEGG metabolic pathway map

The KEGG metabolic pathway maps are graphical diagrams representing knowledge of enzyme-catalyzed reaction networks. Each map is manually drawn to capture the overall architecture of how

* Fax: +81 774 38 3269.
E-mail address: kanehisa@kuicr.kyoto-u.ac.jp

0014-5793/$36.00
http://dx.doi.org/10.1016/j.febslet.2013.06.026

main compounds are converted. The details of individual reactions involving all substrates and products can be examined in the KEGG REACTION entries linked from the map. It is also drawn as a generic map combining and summarizing experimental evidence in different organisms, so that it can be used for interpretation of any genome. This is accomplished by the KEGG Orthology (KO) system described below. Basic graphics objects in the KEGG metabolic pathway maps are boxes for enzymes and circles for chemical compounds (see, for example, http://www.kegg.jp/pathway/map00010). Each circle is identified by the chemical compound identifier (C number). Each box is given two types of identifiers: the reaction identifier (R number) and one or more KO identifiers (K numbers). Although the Enzyme Commission (EC) numbers are usually displayed in the boxes, they are not identifiers and are treated as attributes to KO identifiers. Note that the EC numbers may represent reaction classification of the EC system or gene/protein functional classification in the genome annotation. These two aspects of enzymes are clearly separated by the R number and the K number identifiers in KEGG, enabling the analysis of chemical networks and genomic networks in much better defined ways than using the EC numbers.

2.2. KEGG Orthology

The KEGG Orthology (KO) system is a collection of manually defined ortholog groups (KO entries) for all proteins and functional RNAs that appear in the KEGG pathway maps (both metabolic and non-metabolic) as well as in the KEGG BRITE functional hierarchies (ontologies). Whenever a pathway map is drawn based on experimental observations in specific organisms, an additional manual work is performed for generalizing gene information from those specific organisms to other organisms. This is done by assigning KO entries to the map objects (boxes) and, when necessary, by defining a new KO entry and creating a corresponding set of orthologous genes from available genomes. Each KO entry also represents a sequence similarity group. This allows computational assignment of KO identifiers in newly determined genomes and metagenomes by sequence comparison, which may then be used for KEGG pathway mapping (reconstruction) analysis. Note that the degree of similarity in each group varies significantly because each KO is defined in a context (pathway) dependent manner.

2.3. KEGG reaction class

The KEGG REACTION database contains all biochemical reactions that appear in the KEGG metabolic pathway maps together with the set of experimentally characterized enzymatic reactions in the Enzyme Nomenclature [9], i.e., those with the official EC numbers. Less than one half of the reactions in the KEGG pathway maps correspond to the Enzyme Nomenclature reactions, suggesting the difficulty of using EC numbers for a comprehensive analysis. In order to analyze chemical compound structure transformation patterns, the following processing is performed for all reactions both computationally and manually. First, reactant pairs are defined as one-to-one relationships of substrate-product pairs by considering the reaction type (as classified by the EC system) and the flow of atoms. Second, structure transformation patterns are computed, manually curated, and represented by the so-called RDM patterns of KEGG atom type changes [5–7]. Third, the identity of RDM patterns for the main reactant pairs, i.e., the reactant pairs that appear in the KEGG pathway maps, is used to define KEGG reaction class [8]. The resulting KEGG reaction class (identified by RC number) is like an ortholog group of reactions defined by localized structural changes and accommodating global structural differences of reactants.

2.4. KEGG module

Functional units of enzyme complexes and subpathways are often encoded in positionally correlated gene sets (operon structures) in prokaryotic genomes. When complete genome sequences first became available, a graph analytical method was used to extract enzyme gene clusters on the chromosome that encode consecutive reaction steps in the metabolic pathways [10]. Such functional units are now accumulated in the pathway module section of the KEGG MODULE database. Each KEGG module (identified by M number) is manually defined as a combination of KO identifiers. For example, the reaction sequence involving oxaloacetate + acetyl-CoA, citrate, isocitrate, and 2-oxoglutarate in the citrate cycle (map00020) is the KEGG pathway module M00010 named as "Citrate cycle, first carbon oxidation" and defined by:

$$\mathrm{K01647}\ (\mathrm{K01681}, \mathrm{K01682})\ (\mathrm{K00031}, \mathrm{K00030})$$

where alternative enzymes are given in parentheses. The positional correlation of operon-like structures is not always observed, but when it exists, at least, in certain organism groups, as is the case for many KEGG pathway modules, it well supports the definition of functional units.

2.5. KEGG reaction module

An alternative way to define functional units in the metabolic pathways has been developed recently [8]. It relies only on the chemistry of reactions without using the information about genes and proteins. As mentioned, KEGG pathway nodes (boxes) are given both K numbers (gene/protein orthologs) and R numbers (reactions), where the latter can be converted to RC numbers (reaction class or reaction orthologs). While KEGG pathway modules are conserved subnetwoks of the K number network, different types of conserved subnetworks may exist in the RC number network. This is in fact the case, and conserved reaction sequences termed reaction modules can be extracted from known metabolic pathways [8]. Furthermore, reaction modules (also called RC modules) tend to correspond to KEGG pathway modules (also called KO modules) despite the fact that they are separately defined from different properties. A case in point is the RC module RM001, which exactly matches the KO module M00010, for the reaction sequence from oxaloacetate to 2-oxoglutarate. RM001 is named as "2-Oxocarboxylic acid chain extension by tricarboxylic acid pathway" and defined by:

$$\begin{aligned}
&\mathrm{RC00067}\ (\mathrm{RC00498}+\mathrm{RC00618}, \mathrm{RC00497})(\mathrm{RC00084}+\mathrm{RC00626}, \mathrm{RC00114})\\
&\mathrm{RC01205}\ \mathrm{RC00976}+\mathrm{RC00977RC00417}\\
&\mathrm{RC00470}\ \mathrm{RC01041}+\mathrm{RC01046RC00084}+\mathrm{RC00577}
\end{aligned}$$

where the notation is somewhat more complex because of the existence of three subtypes and multi-step reactions denoted by plus signs.

3. Modular architecture of metabolic network

3.1. Reaction modules used in combination

The analysis of reaction modules has revealed the modular architecture of the metabolic network with two interesting aspects: the existence of chemical units containing chemical logic of organic reactions and the correspondence of chemical and genomic units [8]. The chemical units of reaction modules are used in combination as if they are building blocks of the metabolic network, generating different chemical substances in different pathways. A notable example is illustrated in Fig. 1 for 2-oxocarboxylic acid chain