KEGG PEPTIDE: A Database for Peptide Structures

Michihiro Araki[1]
maraki@hgc.jp

Toshiaki Katayama[1]
ktym@hgc.jp

Yuriko Matsuura[1]
yuriko@scl.kyoto-u.ac.jp

Minoru Kanehisa[1, 2]
kanehisa@scl.kyoto-u.ac.jp

[1] Human Genome Center, Institute of Medical Science, University of Tokyo,
4-6-1 Shirokane-dai Minato-ku, Tokyo 108-8639, Japan
[2] Bioinformatics Center, Institute for Chemical Research, Kyoto University,
Uji, Kyoto 611-0011, Japan

Keywords: database, peptide, chemical genomics

1 Introduction

Genomic information provides a wealth of knowledge to extract the design principles of the biological system. Most of the post-genomic researches focus on describing how the components of the biological system assemble into complex networks to execute autonomous functions on the basis of genome, transcriptome and proteome analyses. Chemical compounds such as metabolites, hormones and drugs are also parts of the major components in the biological system, such that it is of great importance to investigate the correlation between genomic and chemical information in a systematic manner. A lot of information on the chemical genomics has been so far accumulated in extensive literature and databases with respect to biosynthetic pathways, metabolic pathways and drug targets [1-4]. In order to extract significant knowledge lying under the chemical genomic information from the data scattered across different sources, it is necessary to perform an integrated analysis of genomic information (genes and proteins) and chemical information (chemical substances and their reactions).

Peptide information provides a suitable opportunity for the integrated analysis since peptide is one of the chemical compounds highly correlated with genomic information. The diverse functions of peptides extend from antimicrobial activities to the control of biological functions, thereby antimicrobial peptide databases, such as APD [5] and ANTIMIC [6], are constructed to collect peptide information from functional aspects. On the other hand, most peptides are gene encoded to be produced through maturing process (ribosomal peptide), while others are enzymically synthesized as secondary metabolites (nonribosomal peptide). The chemical genomic information in the biosynthetic process besides the functional information includes a lot of knowledge to understand the design principles in the production of peptides coded on genomic information. We thus expand the KEGG LIGAND database to include PEPTIDE for the integrated analysis of peptide and genomic information from the chemical genomic viewpoints.

2 Method and Results

2.1 Database description

The integration of genomics and chemistry has been emphasized in KEGG and the LIGAND database including COMPOUND, REACTION, ENZYME and GLYCAN database has been made publicly available for many years [7]. PEPTIDE database is also built in as one of the LIGAND database collections. Entries of peptides are collected from COMPOUND database, literature, PDB, Swiss-Prot and other databases.

COMPOUND database contains various kinds of peptides like peptide hormones, neuropeptides, antimicrobial peptides and synthetic peptides. PDB and Swiss-Prot include large numbers of ribosomal peptides with various kinds of functions and species, while literature search is the best way to find nonribosomal antibiotic peptides [8].

Each entry contains entry number, peptide name, amino acid composition, reference, class including information on biosynthetic pathway (ribosomal, nonribosomal) and biological activity (hormone, antimicrobial, antitumor), and links to pathway, reaction and enzyme if possible. The drawing tool for 2D peptide structures is also developed based on GLYCAN structures [9], in which KCF (KEGG Chemical Function) format is extended to represent peptide structures as graph objects consisted of amino acids (nodes) and peptide bonds (edges). Many peptides with biological functions include non-natural amino acids, modified functional groups in side chains, and inter- or intra-molecular disulfide bond formation to form complex chemical structures. The peptide drawing tool together with KCF representation can deal with such modifications, which gives a great advantage in subsequent chemical genomic researches.

2.2 Data analysis

Ribosomal peptides are processed into matured forms from precursors mostly in a sequence dependent manner. The analyses of ribosomal peptides thus directly offer knowledge on how genomic information is converted into peptide sequences. However, there are a lot of exceptions we cannot easily predict peptide sequences and structures from genomic sequences due to lack of the integration of the chemical genomic analyses. Integrated comparison among ribosomal peptides in terms of sequences, secondary structures and amino acids indices will be discussed to present another rule in the maturing process of peptides, which will provide useful information on the precise prediction of peptide sequences and the discipline of peptide engineering.

Nonribosomal peptides are one representative group of secondary metabolites constructed on multimodular enzymatic assembly lines found in microbes. Modular enzymes are often encoded on genome as gene clusters, and the building blocks, amino acids, are organized into highly structured compounds. Both nonribosomal peptides and the corresponding genomic information are collected, but no sufficient information is currently available for the chemical genomic analyses. Another example of such unit compounds is polyketide consisted of organic acids, which will complement the deficiency in peptide information. The construction of database including polyketide information in addition to peptide information is on the way to understand how to decipher the chemical code imprinted on genomic information.

References

[1] Yadav, G., Gokhale, R. S. and Mohanty, D., SEARCHPKS: a program for detection and analysis of polyketide synthase domains, *Nucleic Acids Research.*, 31:3654-3658, 2003.

[2] 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.

[3] http://chembank.med.harvard.edu/

[4] Chen, X., Li, Z. L. and Chen, Y. Z., TTD: therapeutic target database, *Nucleic Acids Research.*, 30:412-415, 2002.

[5] Wang, Z. and Wang, G. APD: the antimicrobial peptide database, *Nucleic Acids Research.*, 32:D590-D592, 2004.

[6] Brahmachary, M., Krishnan, S. P., Koh, J. L. Y., Khan, A. M., Seah, S. H., Tan, T. W., Brusic, V. and Bajic, V. B. ANTIMIC: a database of antimicrobial sequences, *Nucleic Acids Research.*, 32:D586-D589, 2004.

[7] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. The KEGG resources for deciphering the genome, *Nucleic Acids Research.*, 32:D277-280, 2004.

[8] Walsh, C. *Antibiotics*, ASM Press, 2004.

[9] Hashimoto, K., Hamajima, M., Goto, S., Masumoto, S., Kawashima, M. and Kanehisa, M. GLYCAN: the database of carbohydrate structures, *Genome Informatics*, 14:649-650, 2003.