PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2010, 37(1): 63~68
www.pibb.ac.cn

MetaGen: a Promising Tool for Modeling Metabolic Networks From KEGG*

ZHOU Ting-Ting[1, 3]**, YUNG Kin-Fung[2]**, CHAN Chun-Chung Keith[2],
WANG Zheng-Hua[1], ZHU Yun-Ping[3]***, HE Fu-Chu[3]***
([1] *National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, Changsha* 410073, China;
[2] *Department of Computing, The Hong Kong Polytechnic University, Hong Kong*, China;
[3] *State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing* 102206, China)

Abstract For the computational researches on the large-scale metabolisms, *MetaGen*, a user-friendly utility to batch process and model the organism-specific multi-level metabolisms in KEGG into enzyme and pathway graphs, was developed. An example was given by expanding a little on the bow-tie structure of metabolic networks to show *MetaGen* is a useful and promising tool. *MetaGen* takes advantages of KEGG web service to ensure data reliability, uses relational database to accelerate the modeling process and facilitate data management, and applies advanced software modeling techniques as well as a pluggable software architecture for the easy expansion of functionalities in the future. *MetaGen* saves researchers from the elaboration on preparing metabolic networks for computation, which paves the way to deepen the researches on the large-scale metabolic networks. *MetaGen* is fully open-sourced and available at http://bnct.sourceforge.net/.

Key words metabolic network, network modeling, KEGG, web service, relational database
DOI: 10.3724/SP.J.1206.2009.00464

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [1] is a comprehensive knowledge repository and is popularly regarded as one of the main data resources for research on metabolic networks. KEGG manages the manually curated pathway maps in KEGG/PATHWAY database, organizes them using the functional hierarchies in KEGG/BRITE database[2], and provides a graphical interface for their navigation through KEGG/Atlas [3]. Generally speaking, pathway maps in KEGG are represented as static or semi-static graphs, which is fixed and typically not accessible by computer programs. In this case, one need tools to model the KEGG pathway maps into the computable metabolic networks before feed them to the computer programs to draw the biologically meaningful inferences. To address this problem, several powerful tools have been developed in recent years.

One excellent work is from Klukas and Schreiber[4]. They proposed a set of methods and developed KGML-ED, the powerful pathway visualization and editing system, for the users to navigate and combine KEGG pathways, edit and export them as KGML files or in other graph exchange formats. Similar contributions are also addressed to PaVESy [5], VisANT [6], KEGG spider [7], MEGU [8], MetaViz [9] and the most recent, KEGGgraph[10]. Among other tools, PaVESy introduces a relational database for the storage of biological objects, which allows users to add features, organize and arrange the database flexibly. However, these tools normally models metabolic networks into bipartite graphs that nodes represent both metabolites and genes

*This work was supported by grants from The National Natural Science Foundation of China (60773021, 60603054, 30621063, 30800200), National Basic Research Program of China(2006CB910803, 2006CB910706), Hi-Tech Research and Development Program of China (2006AA02A312), National S&T Major Project (2008ZX10002-016, 2009ZX09301-002) and State Key Laboratory of Proteomics (SKLP-Y200811).

**These authors contributed equally to this work.

***Corresponding author.

HE Fu-Chu. Tel: 86-10-68171208, E-mail: hefc@nic.bmi.ac.cn
ZHU Yun-Ping. Tel: 86-10-80705225, E-mail: zhuyp@hupo.org.cn

Received: July 31, 2009 Accepted: September 30, 2009

(or enzymes). If one needs no more than enzyme graphs where the nodes only represents enzymes, like the research of Pinter *et al.*[11], he has to figure out if there are such filters to exclude the non-enzymatic elements and how to configure them in the right way.

Previously we proposed a recursive approach[12] to model the organism-specific multi-level metabolisms into enzyme graphs. Following the example of PaVESy, this approach sets up a relational database (MySQL) for data analysis and management. In this paper we extend the approach to model metabolisms to pathway graphs, and re-architect the system for a pluggable infrastructure in order to have more choices for the database schema design. Furthermore, based on the architecture we developed *MetaGen*, an easy-to-use utility to handle the whole work in batch processing. With the help of the local relational database, *MetaGen* not only enables the flexible data management and analysis for end-users, but also accelerates the modeling process to some extent. By taking advantages of KEGG web service[13], *MetaGen* ensures the graph models are both up-to-date and reliable.

The remainder is organized as following. Section 1 shows the details of the modeling method. Section 2 focused on the technologies *MetaGen* employed as well as its requirement and availability. Section 3 illustrates the possible applications by a detailed example on the bow-tie structure. Section 4 summarizes the entire work.

1 Method

KEGG pathway maps are organized in the functional hierarchy defined by the KO system. The hierarchy includes three levels. Referred as “the bio-process level” in our work, the first level consists of the biological processes, including the entire “metabolism”. The second level, described as “the sub-process level” , corresponds to a group of pathways with tightly related functions, and the underlying pathway group, such as carbohydrate metabolism and energy metabolism, is named as “the sub-level process”. The third level is called “the pathway level”, which includes pathways such as Glycolysis/Gluconeogenesis, TCA cycle and so forth. Different KO numbers are assigned to the biological process, sub-level processes and pathways respectively.

MetaGen is designed to model metabolic objects——the biological process(the entire metabolism), sub-level processes or metabolic pathways——into enzyme graphs and pathway graphs by batch processing. Enzyme graphs are modeled upon the following principles: vertices denote individual enzymes and arcs denote the relationships between enzymes; if products of one enzyme and substrates of another are overlapped through at least one metabolite, there is an arc directed from the former enzyme to the latter. The bidirectional arc is replaced by two individual arcs with opposite direction. Compared to the enzyme graph model, the pathway graphs on the sub-process level and the bio-process level are modeled into undirected graphs by linking overlapping pathways who share at least one metabolite, pathways as vertices and links as edges.

As shown in Figure 1, when *MetaGen* starts for the first time, it queries KEGG to retrieve the needed data. Then *MetaGen* analyzes the raw data, restructures and stores the formatted data into the local database as soon as it models the metabolic objects into graphs. Only on the first-time modeling does *MetaGen* visit KEGG and retrieve data on a large scale. After that the local data will always be queried first. In the local database, data is set with a time-to-live (TTL) limit. To reduce the frequent visits of KEGG for data, different types of data can have different TTL limits. When the TTL limit is passed, the data in question will become invalidated and then be renewed by re-querying KEGG when the related graph is re-requested.

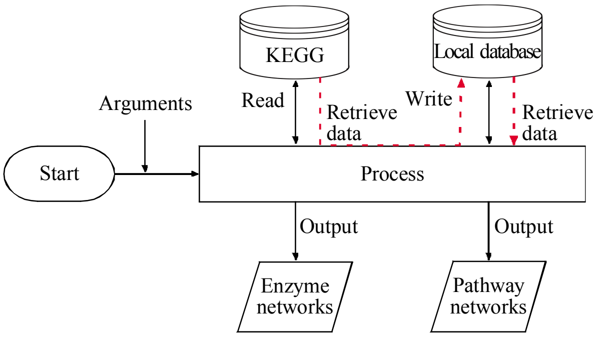

Fig. 1 The overall work schema of *MetaGen*

To model metabolic pathways into enzyme graphs, *MetaGen* use KEGG web service to retrieve enzymes and enzyme relations directly and further assembles them into the enzyme graph model. To model the sub-level processes or the entire metabolism into enzyme graphs, it takes *MetaGen* three steps: first *MetaGen* looks up the KO hierarchy to know all the pathways which belongs to the sub-level process or the entire metabolism; then *MetaGen* models all the pathways into enzyme graphs one after another; and finally *MetaGen* unites all these enzyme graphs into the larger one for the sub-level process or the entire metabolism.

Similar methodology is employed in the modeling of sub-level processes or the entire metabolism into pathway graphs. Firstly *MetaGen* looks up the KO hierarchy to know all the involved pathways; and then *MetaGen* will scan pathways one after another to retrieve all the linking pathway pairs. Having filtered out pathways and pathway links which are not in the metabolic object under investigation, *MetaGen* assembles the linking pathway pairs into pathway graphs as request.

2 Technologies, requirements and availability

MetaGen retrieves data using KEGG web service, manages data using the local relational database server, and creates graph models using JGraphT (http://jgrapht.sourceforge.net/). Currently the modeled graph is written in the format of Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). Other graph formats can be easily added, in the form of plugins to the graph writer architecture inside *MetaGen*.

MetaGen is developed with Eclipse and Java Development Kit (JDK) 6.0, and requires JDK 5.0 or above to run the system. The whole project is built by Maven 2.0, and the system runs on top of Spring Framework 2.5, in which all the internal components are managed by Spring using dependency injection and annotation mechanism. The functionalities provided by Spring Framework enable us to design a lightweight and loosely coupled system which can be easily extended by adding new components. *MetaGen* is database independent; it doesn’t use any vendor-specific SQL syntax, which implies that you can use whatever database you already have. The system ships pre-configured with the MySQL database. For other databases, you will only need to update the database connection property file. *MetaGen* is platform-independent; it have been fully tested on Windows, Linux, Unix and Mac operating systems.

MetaGen is designed to work in several ways. Different argument settings will produce different modeling graph type and scope. *MetaGen* can also accept a command file as the input which is a collection of command line arguments in separate lines, in the case where a set of graphs, especially the enzyme graphs for metabolic pathways, is needed to model. *MetaGen* is freely available for the academic users at http://bnct.sourceforge.net/.

3 Application

MetaGen can handle all the organisms in KEGG by batch processing provided that the organism’s KO system has been set up and metabolic data are relatively complete. *MetaGen* helps users to set up their own metabolic graph database quickly, where the data is up-to-date and non-redundant. This paves the way for most studies on large-scale metabolic networks, such as exploration of metabolic network structure[14–15], prediction of missing enzymes[16], analysis of phylogeny based on the pathway alignment [17] or the network comparison[18–20] and so on. Due to limitations on space, we use one example, the primary analysis on the bow-tie structure of metaboilc networks, to stretch out the prespective on the many possible applications of *MetaGen*.

Bow-tie structure first appeared in the study on the graph structure of web [21], where it is declared to consist of 6 parts: LSCC (Largest Strongly Connect Component), IN, OUT, TUBES, TENDRILS and OTHERS (Figure 2). The core component, LSCC, is the most important part inside this structure. In 2003, Ma *et al.* [22] revealed the directed metabolite networks (Metabolite networks are formed with metabolites as vertices and relations between metabolites as edges.) contain the similar bow-tie structure which had four

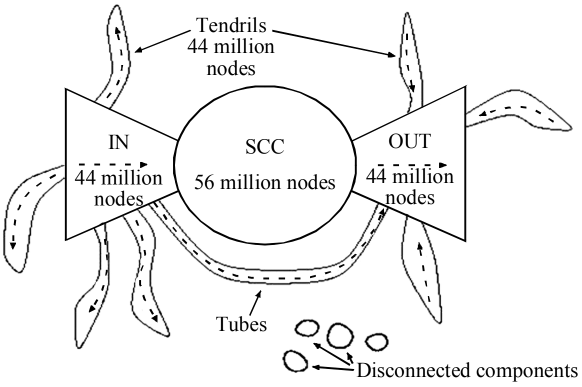

Fig. 2 Bow-tie-Graph structure in the web[21]

One can pass from any node of IN through SCC to any node of OUT. Hanging off IN and OUT are TENDRILS containing nodes that are reachable from portions of IN, or that can reach portions of OUT, without passage through SCC. It is possible for a TENDRIL hanging off from IN to be hooked into a TENDRIL leading into OUT, forming a TUBE——a passage from a portion of IN to a portion of OUT without touching SCC.

components: giant strong component (GSC), substrate subset (S), product subset (P) and isolated subset (S), and shows that the most important metabolic pathways are scoped in GSC. Years later, Zhao *et al*[15] extended this research on GSC and proposed a spread bow-tie model to imply why the metabolism is robust. However, as far as we know, no research answered if there are similar bow-tie structure in metabolic networks represented as the other graphs, and if so, what it implies.

For this purpose we modeled the entire metabolism of *Homo sapiens* (hsa), the most complicated species in KEGG, into the enzyme graph, hsa01100-enzyme, by use of *MetaGen*. The generated enzyme graph consists of 736 vertices and 3 572 arcs. Pajek helps to display and analyze its bow-tie structure (Figure 3). Of all the enzymes, 410 are contained in LSCC, the large component, 77 eznymes in IN, 97 in OUT, 8 in TUBES, 24 in TENDRILS, and 120 in OTHERS. Remarkably, there are 251 enzymes existing in more than one pathways, which are all in LSCC. Enzymes acting at cross points of pathways may take up important roles in the function and evolution [23], which reveals the importance of LSCC from other aspect than the superiority of quantity and topology.

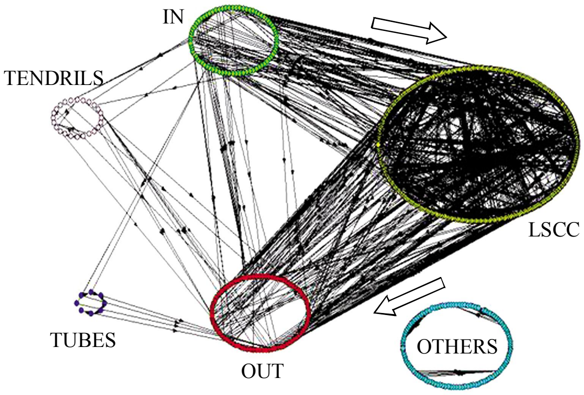

Fig. 3 Bow-tie structure of hsa01100-enzyme, the entire enzyme graph of hsa

Figure 4 shows the pathway graph modeled from the entire metabolism of hsa, named hsa01100-pathway, which consists of 108 vertices (pathways) and 169 edges (pathway links). It is displayed by Cytoscape [24] in the yFile circular layout. Although to our knowledge there were no formal definition for the bow-tie structure in undirected graphs, one can still see its basic graph characteristics: pathways are clustered in mainly three components. In this case, we call it pseudo-bow-tie structure. The central biggest circle, corresponding to LSCC (GSC), contains 63 individual pathways and covers all the 11 sub-level processes, which primarily shows its dominance in the metabolic functions. The outer ring, regarded as union of IN, OUT, TENDRILS and TUBES in the bow-tie structure of directed networks, contains 24 pathways. The part scattered at the lower left corner and the upper right corner corresponds to OTHERS, where 21 pathways are like orphans. Is this structure possibly related to network evolution, or is there any correspondence between this topology and the functions of pathways? More efforts are needed to answer these questions.

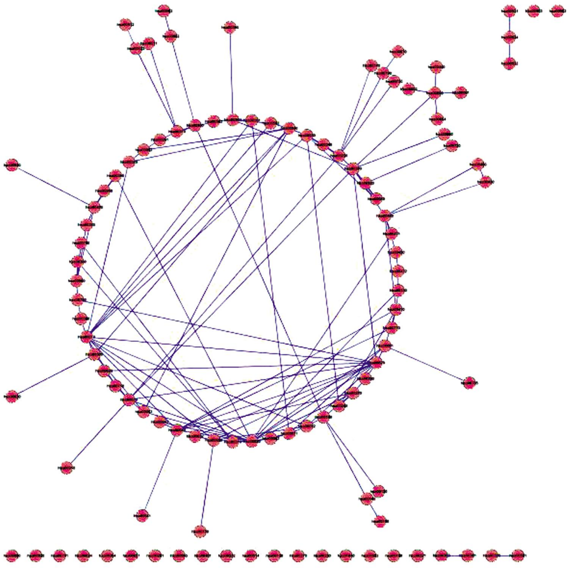

Fig. 4 Pseudo-bow-tie structure of hsa01100-pathway, the entire pathway graph of hsa

4 Conclusion

In this paper we made attempts to develop an easy-to-use utility to provide multiple options for the modeling of metabolic networks from KEGG for the computational purpose. By taking advantages of *MetaGen*, KEGG web service is used to keep data up-to-date and non-redundant; the local SQL database is used to accelerate the modeling process and makes data management and analysis more flexible and predictable; advanced software modeling techniques and a pluggable software architecture is employed for the easy expansion of functionalities in the future. *MetaGen* saves researchers from the elaboration on preparing metabolic networks for computation, which paves the way to deepen the researches on the large-scale metabolic networks. *MetaGen* is LGPL open-sourced and available at http://bnct.sourceforge.net/.

Acknowledgements We thank KEGG technical support for the use of their web service. We also thank Prof. Amos Bairoch for the helpful discussion.

References

[1] Kanehisa M, Araki M, Goto S, *et al*. KEGG for linking genomes to life and the environment. Nucleic Acids Res, 2008, **36** (Database issue): D480-484

[2] Aoki-Kinoshita K F, Kanehisa M. Gene annotation and pathway mapping in KEGG. Methods Mol Biol, 2007, **396**: 71-91

[3] Okuda S, Yamada T, Hamajima M, *et al*. KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res, 2008, **36**(Web Server issue): W423-426

[4] Klukas C, Schreiber F. Dynamic exploration and editing of KEGG pathway diagrams. Bioinformatics, 2007, **23**(3): 344-350

[5] Ludemann A, Weicht D, Selbig J, *et al*. PaVESy: pathway visualization and editing system. Bioinformatics, 2004, **20** (16): 2841-2844

[6] Hu Z, Ng D M, Yamada T, *et al*. VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. Nucl Acids Res, 2007, **35**(suppl_2): W625-632

[7] Antonov A V, Dietmann S, Mewes H W. KEGG spider: interpretation of genomics data in the context of the global gene metabolic network. Genome Biol, 2008, **9**(12): R179

[8] Kono N, Arakawa K, Tomita M. MEGU: pathway mapping web-service based on KEGG and SVG. In Silico Biol, 2006, **6**(6): 621-625

[9] Bourqui R, Cottret L, Lacroix V, *et al*. Metabolic network visualization eliminating node redundance and preserving metabolic pathways. BMC Syst Biol, 2007, **1**: 29

[10] Zhang J D, Wiemann S. KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor. Bioinformatics, 2009, **25**(11): 1470-1471

[11] Pinter R Y, Rokhlenko O, Yeger-Lotem E, *et al*. Alignment of metabolic pathways. Bioinformatics, 2005, **21**(16): 3401-3408

[12] Zhou T, Yung K F, Wang Z, *et al*. A new approach for reconstructing metabolic networks from KEGG[J/OL]. Computer Engineering and Science, 2009 , **32** (7) [2010-01-12]. http://master.dl.sourceforge.net/project/bnct/manuscripts/Approach_draft.pdf

[13] Kawashima S, Katayama T, Sato Y, *et al*. KEGG API: A web service using SOAP/WSDL to access the KEGG system. Genome Informatics, 2003, **14**: 673-674

[14] Zhao J, Ding G H, Tao L, *et al*. Modular co-evolution of metabolic networks. BMC Bioinformatics, 2007, **8**(1): 311

[15] Zhao J, Tao L, Yu H, *et al*. Bow-tie topological features of metabolic networks and the functional significance. Chin Sci Bull, 2007, **52**(8): 1036-1045

[16] Yamanishi Y, Mihara H, Osaki M, *et al*. Prediction of missing enzyme genes in a bacterial metabolic network. Reconstruction of the lysine-degradation pathway of *Pseudomonas aeruginosa*. FEBS J, 2007, **274**(9): 2262-2273

[17] Heymans M, Singh A. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics, 2003, **19**(90001): 138-146

[18] Zhou T, Chan K, Wang Z. TopEVM: using co-occurrence and topology patterns of enzymes in metabolic networks to construct phylogenetic trees. LNCS (LNBI), 2008, **5265**: 225-236

[19] Zhou T, Chan K, Pan Y, *et al*. An approach for determining evolutionary distance in network-based phylogenetic analysis. LNCS (LNBI), 2008, **4983**: 38-45

[20] Mazurie A, Bonchev D, Schwikowski B, *et al*. Phylogenetic distances are encoded in networks of interacting pathways. Bioinformatics, 2008, **24**(22): 2579-2585

[21] Broder A, Kumar R, Maghoul F, *et al*. Graph structure in the web. Computer Networks, 2000, **33**(1-6): 309-320

[22] Ma H W, Zeng A P. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics, 2003, **19**(11): 1423-1430

[23] Greenberg A J, Stockwell S R, Clark A G. Evolutionary constraint and adaptation in the metabolic network of *Drosophila*. Mol Biol Evol, 2008, **25**(12): 2537-2546

[24] Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, 2003, **13**(11): 2498-2504

MetaGen: 从 KEGG 建模代谢网络的新工具 *

周婷婷1, 3)**　容健锋2)**　陈振冲2)　王正华1)　朱云平3)***　贺福初3)***

(1)国防科学技术大学并行与分布式处理国家重点实验室，长沙 410073；2)香港理工大学计算学系，中国香港；
3)蛋白质组学国家重点实验室，北京蛋白质组研究中心，军事医学科学院放射与辐射医学研究所，北京 102206)

摘要　为便于大规模代谢网络的计算，发展了一款方便实用的工具：*MetaGen*，对 Kyoto Encyclopedia of Genes and Genomes (KEGG)中物种特异的各层次代谢系统进行建模，生成的代谢网络以酶图和通路图的方式表示．利用该工具，对人类代谢系统的 bow-tie 结构进行了初步研究，并以此为例展示了该工具广阔的应用前景．*MetaGen* 利用 KEGG web 服务保证建模数据的可靠性，依靠本地关系数据库加速网络建模过程并提供更多的数据管理和利用方式， 并结合高级 JAVA 技术提高代码的可扩展性．*MetaGen* 完全开源，可直接从 http://bnct.sourceforge.net/ 下载．

关键词　代谢网络，网络建模，KEGG，web 服务，关系数据库

学科分类号　Q61，Q591，Q811.4　　**DOI:** 10.3724/SP.J.1206.2009.00464

* 国家自然科学基金资助项目(60773021,60603054, 30800200, 30621063), 国家重点基础研究发展计划(973)(2006CB910803, 2006CB910706), 国家高技术研究发展计划(863)(2006AA02A312), 国家科技重大专项(2008ZX10002-016, 2009ZX09301-002)和蛋白质组学国家重点实验室课题(SKLP-Y200811)资助项目.

** 共同第一作者.

** 通讯联系人.

贺福初. Tel: 010-68171208, E-mail: hefc@nic.bmi.ac.cn

朱云平. Tel: 010-80705225, E-mail: zhuyp@hupo.org.cn

收稿日期：2009-07-31，接受日期：2009-09-30