Application of databases for development of stoichiometric and dynamic models of biochemical networks

Natalja Bulipopa, Ilona Odzina

Latvia University of Agriculture, Liela Str. 2, Jelgava, LV-3001, Latvia
cvetkova.natalja@gmail.com

Abstract: *Metabolic engineering uses tools of bioinformatics, systems biology and synthetic biology to modify organisms for new biotechnological tasks. Stoichiometric and dynamic models are needed for successful design of requested bioprocess. In spite of available model databases both for stoichometric and dynamic models they often are applicable just partly as deposited models may not cover the organism and process of interest. Depending on that intensive use of databases about particular reactions and metabolites may be required. This study analyses and suggests possible application of different publically available databases developing stochiometric and dynamic models. The contents and applications of data bases KEGG, BRENDA, SABIO-RK, IntEnz, MetaCyc, ChEBI, PubChem, BioModelsare analyzed.*

Keywords: metabolic engineering, KEGG, BRENDA, SABIO-RK, BioModels.

Introduction

Metabolic engineering is a growing field of research in microbiology serving as an integrated approach to design new cell factories by providing rational design procedures and valuable mathematical and experimental tools (Patil et al., 2004). Different aspects of bioinformatics, systems biology and synthetic biology are exploited to improve the features of cell factories (Nielsen and Keasling, 2011). In the field of bioinformatics, systems biology and synthetic biology one of the main research challenges are the understanding of biological and biochemical process, integration of existing data in databases thus contributing to the analysis of the use of new analytical methods. The database provides integrated tools, enabling the user to independently perform data analysis.

Both stoichiometric and dynamic models are used in metabolic engineering. The development of stoichiometric models or stoichiometric reconstructions is generally well developed (Thiele and Palsson, 2010) and there are several data bases offering published models. Still important organism specific information has to be implemented in the model by biologists specializing in the particular organism (Pentjuss et al., 2013) because comparison of different models of the same organism often demonstrates high disagreement (Mednis and Aurich, 2012). In case of dynamic modeling there are dynamic models available in dedicated database (Novère et al., 2006). Still dynamic models are relatively small in terms of number of involved reactions (usually some tens of reactions) and often they are not covering the process of interest or it is modeled under different circumstances. Therefore data bases are needed to find very specific kinetic parameters to perform parameter estimation and optimization tasks (Mendes and Kell, 1998; Sulins and Mednis, 2012; Kostromins et al., 2012).

Biological databases are an important tool in assisting scientists to understand biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge facilitates the fight against diseases, assists in the development of medications and in discovering basic relationships amongst species in the history of life (Makheswari and Sudarsanam, 2012).

In case of metabolic engineering the first step is to set the task of the cell factory: substrate, product and limitations of the task (Lee et al., 2010). After that the choice of the organism of interest has to be done. The next challenge is to choose for each analysis stage corresponding initial acquisition of information or database.

The aim of this article is to analyze the workflow of computer modeling in metabolic engineering of biological organism and database usage.

Computer modeling in metabolic engineering of biological organism

Before choosing a database, at the beginning must be defined an objective (task) – to set the task of the cell factory: substrate, product and limitations of the task.

When the objective (task) was defined, must choose the organism and some process of chosen organism which will be analyzed and the structural model of chosen organism will be developed. The structural model contains reactions, their directions and metabolites. For acquisition of this initial information, the following free of charge database are often used: Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2006), BRENDA (Chang et al., 2009) and SABIO-RK (Rojas et al., 2007).

The created structural model allows us to analyze and illustrate the network of biochemical processes – cycles and node degrees (Fig. 1) (Rubina, 2012; Odzina et al., 2010).

This model describes interaction of the reaction and metabolites. In order to create organism's reconstruction the additional information is needed about the reaction's genes, proteins, subsystems, enzyme numbers, metabolites and their full title, neutral formula, charged formula, charge, compartment, ID numbers and encoding (InChi string and SMILES). To obtain information simultaneously can be used several databases: KEGG, BRENDA, IntEnz (Integrated relational Enzyme database) (Alcántara et al., 2013), MetaCyc (Caspi et al., 2012), BioCyc (Caspi et al., 2010), Chemical Entities of Biological Interest (ChEBI) (Matos et al., 2010), and PubChem (Bolton et al., 2008). All this information is necessary for the analysis and inspection of reconstruction by COBRA Toolbox software (Schellenberger et al., 2011).

If after the reconstruction of model development has been ascertained that in the model is interaction between all reactions and metabolites, then can analyze the dynamic parameters of the model. Thus, the dynamic model is created. Model creation is necessary to add information of the kinetic equations and parameters. The following databases: BRENDA (Chang et al., 2009), SABIO-RK (Rojas et al., 2007), and BioModels (Novère et al., 2006) contain most of necessary information.

The last step after the dynamic model development is parameter estimation, based on a dynamic model simulation data. When, model's simulated data are approximated to experimental data, then developed dynamic mode (*In silico*) can be used in laboratory experiments (*In vitro*), in order to reduce the number of failed experiments.

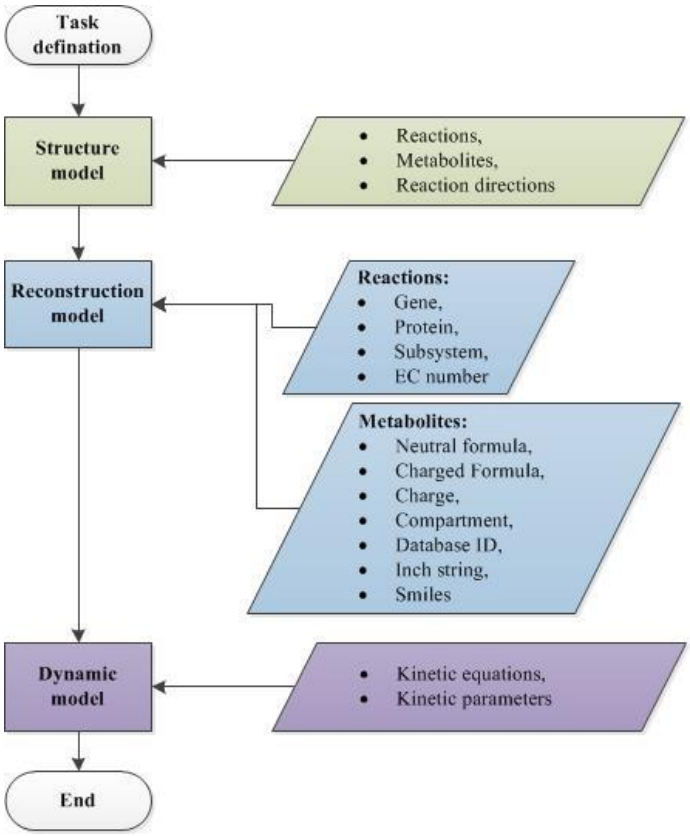

Fig. 1. **Algorithm of metabolic engineering computer modeling of biological organism.**

Biological Databases

Kyoto Encyclopedia of Genes and Genomes (KEGG): is a bioinformatics database containing information on genes, proteins, reactions and pathways. The organisms section is divided into eukaryotes and prokaryotes, encompasses many organisms for which gene and DNA information can be searched by typing in the enzyme of choice. This resource can be extremely useful when building the association between metabolism enzymes, reactions and genes (Ogata et al., 1999). As well as for overall examination of enzymatic reactions and EC numbers the database KEGG can be used. KEGG EC number information and reaction directionality are not organism – specific (Kanehisa et al., 2006).

BRENDA: a comprehensive enzyme database. Database contains functional data for all enzyme classes (~ 4800 entries in six main classes in 2008) that have been classified according to the EC scheme of the IUBMB (International Union of Biochemistry and Molecular Biology) irrespectively of the enzyme's source. The range of data in database is not restricted to specific aspects but includes a wide area of biochemical and molecular properties of enzymes such as a) classification and nomenclature; b) reaction and specificity; c) functional parameters; d) organism-related information; e) enzyme structure; f) isolation and preparation; g) literature

references; h) application and engineering; i) enzyme–disease relationships. All data and information are manually extracted from the primary literature and are connected to the biological source of the enzyme, i.e. the organism, the tissue, the subcellular localization and/or the protein sequence (if available) (Chang et al., 2009). Database also, allows searching for an enzyme by name or EC number.

Organism – specific information was collected mainly from previously published data and the BRENDA database, where each reaction EC number and reaction directionality was checked and validated.

SABIO-RK - SABIO-RK is a web-accessible curated database offering information about biochemical reactions and their kinetic properties. It integrates information about reactions, such as reactants, effectors, and catalyzing enzymes, with information about organisms, tissues and cellular locations where the reactions take place, and with the kinetic properties of these reactions (type of the kinetic mechanism, modes of inhibition or activation and rate equations together with their parameters and measured values). As kinetic constants highly depend on environmental conditions used for their determination these are given together with the description of the kinetics. This also facilitates the comparison of data sets based on experiments assayed under similar experimental conditions. The database is populated by merging data from several sources. The general information about the reactions is mainly obtained from external databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes). In contrast, the kinetic data along with descriptions of the experimental conditions under which they were determined are primarily manually extracted from literature and curated by a team of scientists (Rojas et al., 2007).

IntEnz (Integrated relational Enzyme database) (Alcántara et al., 2013) is a freely available resource focused on enzyme nomenclature. A relational database provides better interoperability with other bioinformatics resources. All the major biological databases at the EBI, such as EMBL Nucleotide Sequence Database, UniProt and MSD, are relational. Enzyme portal (http://www.ebi.ac.uk/enzymeportal) to provide wealth of information on enzymes from multiple in-house resources addressing particular data classes: protein sequence and structure, reactions, pathways and small molecules. The fact that these data reside in separate databases makes information discovery cumbersome (Alcántara et al., 2013).

MetaCyc is a database of nonredundant, experimentally elucidated metabolic pathways (Caspi et al., 2012). MetaCyc contains more than 1928 pathways from more than 2263 different organisms, and is curated from the scientific experimental literature. MetaCyc contains pathways involved in both primary and secondary metabolism, as well as associated compounds, enzymes, and genes. MetaCyc data can be accessed in several ways: a) search for pathways, enzymes, reactions, and metabolites through this Web site; b) install MetaCyc on your computer in conjunction with the Pathway Tools software for faster access and more query options. This local installation also allows you to query MetaCyc programmatically using Java or PERL programs; c) Download MetaCyc data files (Caspi et al., 2012).

BioCyc is a collection of 2038 Pathway/Genome Databases (PGDBs) (Caspi et al., 2010). Each PGDB in the BioCyc collection describes the genome and metabolic pathways of a single organism. The BioCyc Web site contains many tools for navigating, visualizing, and analyzing these databases, and for analyzing omics data, including the following: a) genome browser; b) display of individual metabolic pathways, and of full metabolic maps; c) visual analysis of user-supplied omics datasets by painting onto metabolic maps, regulatory maps, and genome maps; d) store groups of genes and pathways in your account; share, analyze, transform those groups; e) comparative analysis tools (Caspi et al., 2010).

Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical compounds (Matos et al., 2010). The term 'molecular entity' refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable entity. The molecular entities in question are either product of nature or synthetic products used to intervene in the processes of living organisms.

In order to create ChEBI, data from a number of sources were incorporated and subjected to merging procedures to eliminate redundancy.

Four of the main sources from which the data are drawn are: a) IntEnz – the Integrated relational Enzyme database of the EBI. IntEnz is the master copy of the Enzyme Nomenclature, the recommendations of the NC-IUBMB on the Nomenclature and Classification of Enzyme-Catalysed Reactions; b) KEGG COMPOUND – One part of the Kyoto Encyclopedia of Genes and Genomes LIGAND database, COMPOUND is a collection of biochemical compound structures; c) PDBeChem – The service providing web access to the Chemical Component Dictionary of the wwPDB as this is loaded into the PDBe database at the EBI; d) ChEMBL – A database of approximately 500,000 bioactive compounds, their quantitative properties and bioactivities, abstracted from the primary scientific literature. It is part of the ChEMBL resources at the EBI (Matos et al., 2010).

PubChem is an open repository for experimental data identifying the biological activities of small molecules. The primary aim of PubChem is to provide a publicē on - line resource of comprehensive information on the biological activities of small molecules accessible to molecular biologists as well as computation al and medicinal chemists. PubChem contents include more than: 1,000 bioassays, 28 million bioassay test outcomes, 40 million substance contributed descriptions, and 19 million unique compound structures contributed from over

70 depositing organizations. PubChem provides a significant, publicly accessible platform for mining the biological information of small molecules (Bolton et al., 2008).

BioModels database part of the international initiative BioModels.net, provides access to published, peer-reviewed, quantitative models of biochemical and cellular systems. Database aims are as follows: a) to define agreed-upon standards for model curation, b) to define agreed-upon vocabularies for annotating models with connections to biological data resources and c) to provide a free, centralized, publicly accessible database of annotated, computational models in SBML and other structured formats. Database is an annotated resource of quantitative models of biomedical interest. Models are carefully curated to verify their correspondence to their source articles. They are also extensively annotated, with a) terms from controlled vocabularies, such as disease codes and Gene Ontology terms and b) links to other data resources, such as sequence or pathway databases.

The models can currently be retrieved in the SBML format, and import/export facilities are being developed to extend the spectrum of formats supported by the resource (Novère et al., 2006).

Conclusion

There are many and varied biological databases and its containing different information from very specific to more comprehensive. But to be able to choose database, first of all, must be defined the initial task. For each model development step it is necessary to choose the most appropriate database to achieve better results. However the same databases can be used several times during the study, such as, KEGG or BRENDA, because the information in the database are ranked by biological factors, such as Metabolic Pathway Databases, Protein Databases, genomic databases et al.. Search in several databases at the same time can increase the probability that creating a computer model can be made a mistake. Consequently, the database information is appropriate to take instead of a single biological experiment, but from several databases.

When model's simulated data are approximated to experimental data, then developed dynamic model (*In silico*) can be used in laboratory experiments (*In vitro*), in order to reduce the number of failed experiments.

Reference

Alcántara, R., Onwubiko, J., Cao, H., Matos, P., Cham, J.A., Jacobsen, J., Holliday, G.L., Fischer, J.D., Rahman, S.A., Jassal, B., Goujon, M., Rowland, F., Velankar, S., López, R., Overington, J.P., Kleywegt, G.J., Hermjakob, H., O'Donova,n C., Martín, M.J., Thornton, J.M. and Steinbeck C., 2013. The EBI enzyme portal, *Nucleic Acids Res, 41*, Database issue D773–D780. doi: 10.1093/nar/gks1112

Bolton, E.E., Wang, Y., Thiessen, P.A. and Bryant, S.H., 2008. PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12. In *Annual Reports in Computational Chemistry*, *4*, American Chemical Society, Washington, DC.

Caspi, R., Altman, T., Dale, J.M., Dreher, K, Fulcher, C.A., Gilham, F., Kaipa, P., Karthikeyan, A.S., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Paley, S., Popescu, L., Pujar, A., Shearer, A.G., Zhang, P. and Karp, P.D., 2010. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research, 33* (19), Database issue 6083-6089. doi: 10.1093/nar/gki892

Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, O., Paley, S., Pujar, A., Shearer, A.G., Travers, M., Weerasinghe, D., Zhang, P. and Karp, P.D., 2012. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research, 40* , Database issue *D742-D753*. doi: 10.1093/nar/gkr1014

Chang, A., Scheer, M., Grote, A., Schomburg, I. and Schomburg, D., 2009. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Research, 37*, Database issue D588–D592, doi:10.1093/nar/gkn820

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M., 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research, 34*, Database issue D354–D357, doi:10.1093/nar/gkj102/.

Kostromins, A., Mozga, I. and Stalidzans, E., 2012. ConvAn: a convergence analyzing tool for optimization of biochemical networks. *Biosystems, 108*(1-3), pp. 73–77. doi:10.1016/j.biosystems.2011.12.004

Lee, F.C., Pandu Rangaiah, G. and Lee, D.Y., 2010. Modeling and optimization of a multi-product biosynthesis factory for multiple objectives. *Metabolic Engineering*, *12*(3), pp. 251–267. doi:10.1016/j.ymben.2009.12.003

Makheswari, U.M. and Sudarsanam, D., 2012. A Review on Bio Informatics for Diabetic Mellitus. *International Journal of Pharma Sciences and Research (IJPSR), **3*** (6), ISSN : 0975-9492, pp 389 - 395

Matos, P., Alcantara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S. and Steinbeck, C., 2010. Chemical Entities of Biological Interest: an update. *Nucleic Acids Research, 38*, Database issue D249–D254, doi:10.1093/nar/gkp886

Mednis, M. and Aurich, M. K., 2012. Application of string similarity ratio and edit distance in automatic metabolite reconciliation comparing reconstructions and models. *Biosystems and Information Technology*, *1*(1), pp. 14–18. doi:10.11592/bit.121102

Mendes, P. and Kell, D., 1998. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, *14*(10), pp. 869–883. doi:10.1093/bioinformatics/14.10.869

Nielsen, J. and Keasling, J.D., 2011. Synergies between synthetic biology and metabolic engineering. *Nature Biotechnology*, *29*(8), pp. 693–695. doi:10.1038/nbt.1937

Novère, L.N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J.L. and Hucka M., 2006. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, *34*, Database issue D689–D691, doi:10.1093/nar/gkj092

Odzina, I., Rubina, T., Rutkis, R., Kalnenieks, U. and Stalidzans, E., 2010. Structural model of biochemical network of Zymomonas mobilis adaptation for glycerol conversion into bioethanol. In Proceedings of Applied Information and Communication Technologies 2010, pp. 50-54. ISBN 978-9984-48-022-0

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1), pp.29-34.

Patil, K.R., Akesson, M. and Nielsen, J., 2004. Use of genome-scale microbial models for metabolic engineering. *Current Opinion in Biotechnology*, *15*(1), pp. 64–69. doi:10.1016/j.copbio.2003.11.003

Pentjuss, A., Odzina, I., Kostromins, A., Fell, D., Stalidzans, E. and Kalnenieks, U., 2013. Biotechnological potential of respiring *Zymomonas mobilis*: a stoichiometric analysis of its central metabolism. *Journal of Biotechnology*, *165*(1), 1-10 . doi:10.1016/j.jbiotec.2013.02.014

Rojas, I., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann, A. and Wittig, U., 2007. SABIO-RK: a database for biochemical reactions and their kinetics. *BMC Systems Biology*, ***1*** (Suppl 1):S6, doi: 10.1186/1752-0509-1-S1-S6

Rubina, T., 2012. Tools for analysis of biochemical network topology. *Biosystems and Information Technology*, *1*(1), pp. 25–31. doi:10.11592/bit.121101

Schellenberger, J., Que, R., Fleming, R.M., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D.C., Bordbar, A, Lewis, N.E, Rahmanian, S, Kang, J, Hyduke, D.R, Palsson, B.Ø., 2011. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature Protocols*, *6*(9), pp. 1290–1307. doi:10.1038/nprot.2011.308

Sulins, J. and Mednis, M., 2012. Automatic termination of parallel optimization runs of stochastic global optimization methods in consensus or stagnation cases. *Biosystems and Information Technology*, *1*(1).

Thiele, I. and Palsson, B.Ø., 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*(1), pp. 93–121. doi:10.1038/nprot.2009.203