Data Integration Issues in the Reconstruction of the Genome-Scale Metabolic Model of *Zymomonas Mobillis*

José P. Pinto[1], Oscar Dias[2], Anália Lourenço[2], Sónia Carneiro[2], Eugénio C. Ferreira[2], Isabel Rocha[2], and Miguel Rocha[1]

[1] Departament of Informatics / CCTC
[2] IBB - Institute for Biotechnology and Bioengineering, Center of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga - Portugal
{josepedr,mrocha}@di.uminho.pt,
{odias,analia,soniacarneiro,ecferreira,irocha}@deb.uminho.pt

Abstract. Genome-scale model reconstruction represents a major tool in the field of Metabolic Engineering .This paper reports on a study about data integration issues in the process of genome-scale reconstruction of the metabolic model of the bacterium *Zymomonas mobilis,* a promising organism for bioethanol production. Data is retrieved from the Entrez Gene, KEGG, BioCyc and Brenda databases, and the several processes involved in data integration from these sources are described, as well as the data quality issues.

Keywords: Genome-scale model reconstruction, *Zymomonas mobillis*, data integration, data quality.

1 Introduction

Genome-scale reconstructed metabolic models are based on the well-known stoichiometry of biochemical reactions and can be used for simulating *in silico* the phenotypic behaviour of a microorganism under different environmental and genetic conditions, thus representing an important tool in metabolic engineering [1]. However, while the reconstruction of the metabolic network of an organism is likely to become a widespread procedure, starting with the fully sequenced and (partially) annotated genome sequence, it is currently far from being a standardized methodology [2]. This is due in part to the lack of uniform computational tools for model reconstruction, but primarily to the difficulties associated with the extraction of information other than what is available from the annotated genome.

In this paper, we address the reconstruction of the metabolic model of *Zymomonas mobilis* ZM4, among the most promising microorganisms for ethanol fuel production [3]. The genome-scale metabolic reconstruction is imperative for the feasibility of ongoing studies since there is no available genome-scale metabolic model for this organism. The number of reports in current literature studying its *in vivo* physiology remains small and there is a limited use of the metabolic engineering experimental and computational tools in the understanding of its metabolic pathway interconnectivity [4]. Therefore, genome-scale metabolic modeling stands out as one of the most promising approaches to obtain *in silico* predictions of cellular function based on the interaction of the cellular components [5,2].

J.M. Corchado (Eds.): IWPACBB 2008, ASC 49, pp. 92–101, 2009.
springerlink.com

This work is focused on the first steps of metabolic network reconstruction aiming at delivering valuable forms of automation that can assist on the collection and processing of the information required. This case study is invaluable, because of its importance as an ethanologenic source and the scarce availability of data to support related research. However, the workflow was planned in order to account for data issues in an organism-independent way. All the processes and analysis guidelines proposed can be applied to the reconstruction of models of other organisms, adjusting only data retrieval processes from particular repositories.

The main focus of this work lays on data integration planning, in particular on the assessment of data quality. Handling the diversity and quality of the contents along with data formats and structure is determinant to obtain a consistent repository. Most of the times, biologists rely on particular data sources, which they are familiar with. Understanding the reasoning that drives the expert while manually searching for data, allows the identification of the basic set of elements from each source and, far more important, how data sources can be linked together. From then on, source information extraction and preliminary processing can be fully automated and multi-source data integration can be achieved.

The establishment of a fully automated dataflow is inconceivable because data quality poses challenging issues that require expert non-trivial evaluation. Intra-source data quality is often disputable. Misspellings, nulls, duplicates and inconsistencies may undermine data acquisition and further integration. Multi-source integration raises additional quality concerns due to the scarce use of standard nomenclatures that raises terminological issues, namely term novelty, homonymy and synonymy. However, data processing can account for the most common issues, delivering descriptive quality-related statistics and proposing, when possible, candidate solutions to the integration and data quality issues.

2 Information Requirements

The genome-scale reconstruction of a metabolic network encompasses several steps [1], as depicted on Fig. 1. : (1) genome annotation; (2) identification of the biochemical reactions from the annotated genome sequence and available literature; (3) determination of the reaction stoichiometry including cofactor requirements; (4) definition of compartmentation and assignment of reaction localizations; (5) determination of the biomass composition; (6) measurement, calculation, or fitting of energy requirements; and (7) definition of additional constraints.

The process is laborious and requires substantial manual evaluation of the stoichiometry of different reactions in the network: whereas it typically takes 10% of the reconstruction time to collect 90% of all reactions from the annotated genome sequence, the remaining 90% of the time is spent collecting the remaining 10% of data from literature. The present work discusses the shadowed steps of the figure, namely the identification of reactions and collection of stoichiometric data.

For the microorganisms with fully sequenced genomes, the process of reconstructing the metabolic network starts with a careful inspection of the data obtained from the genome annotation. The process can be initiated by consulting a public repository

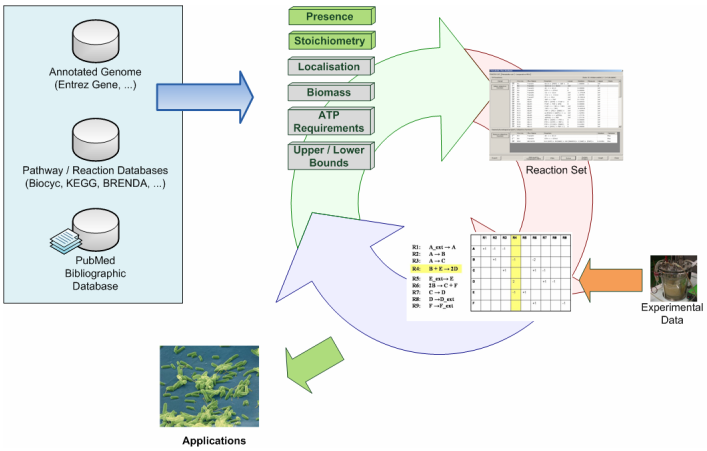

Fig. 1. An illustration of the metabolic network reconstruction process

of genome sequence data, such as GOLD [6], TIGR [7] or NCBI [8]. Important data to be extracted from these sources include gene or open reading frame (ORF) names, assigned cellular functions, sequence similarities, and, for enzyme encoding genes, the Enzyme Commission (EC) number(s) corresponding with the gene products. From the complete set of sequenced genes, only the genes encoding enzymes and membrane transporters are used for the reconstruction.

At the end of this process, the names of the genes assigned during genome annotation, the names of the reactions, reactants and products for each reaction should all be included in the reaction list. Typically, the initial reconstruction only considers genes that code for enzymes with EC numbers assigned. Public pathway databases, such as BRENDA [9] or KEGG [10] provide detailed information about each individual reaction catalyzed by a enzyme with assigned EC number.

Then, the reaction set has to be complemented with reactions catalyzed by enzymes that do not have EC numbers assigned, with transport and exchange reactions, and with reactions known to exist in a given organism, but for which no corresponding genes have been found during annotation. This can only be accomplished by thorough curation of publications and biochemistry textbooks. Curation may be fully manual or comprise the use of Biomedical Text Mining techniques. Either way, due to terminological issues and the challenges posed by unstructured text processing, it is laborious and time-consuming.

Despite the many obstacles faced, this information validates the data deduced from the genome and discarding questionable reactions with poor annotation based on low sequence similarity and those for which no evidence has been found in literature. Also, it supports the selection of reaction(s) specific to the organism being reconstructed from the multiple potential reactions associated with each given EC number in public databases. Furthermore, special cases with more complex than one-gene-to-one-enzyme-to-one-reaction relations need to be considered: (1) many enzymes accept several different substrates; (2) isoenzymes are encoded by different genes, but each of them catalyzes the same reaction(s); (3) for reactions catalyzed by enzyme complexes, several genes are associated with one or more reactions [1]. Information

about reaction stoichiometry can also be found in public databases only for enzymes with assigned EC numbers. For all other reactions, stoichiometric information should be based on the literature data.

3 Data Integration

Our workflow is illustrated in Figure 2 and encompasses the following steps: (1) data loading from original sources into temporary tables; (2) single-source debugging; (3) single-source quality-related processing; (4) detection of conflicts on multi-source integration; (5) semi-automatic conflict resolution; (6) multi-source contents integration; and (7) enforcement of data quality.

3.1 Data Source Description

The parameters related to genome annotation were taken from the NCBI's Entrez Gene [11]. The list of reactions and stoichiometry data was delivered by integrated contents from KEGG, BioCyc and Brenda.

The Kyoto Encyclopedia of Genes and Genomes (**KEGG**) is an information repository that contains several kinds of biological data, with the purpose of linking genomic information with higher order functional information [12]. KEGG is composed by 19 highly integrated databases, each one belonging to one of three categories: Systems, Genomic or Chemical [13]. Due to this widespread hyperlinking, KEGG was considered one of the "central" data sources of this project. KEGG's data is organized in organism-specific and class-specific subdirectories where data is kept in text files. Currently, we extract information from databases of the genomic and chemical categories and also information about pathways.

BRaunschweig ENzyme DAtabase (**BRENDA**) - is the main collection of enzyme functional data available and its data is primarily collected from literature [14]. We use it to complement the enzyme data extracted from generic data sources (for example KEGG or BioCyc). BRENDA is a very extensive database characterized by the fact that it is not limited to specific organisms or aspects of enzymes, covering a wide range of information for all enzymes [15].

In BRENDA, the information is organized by EC number, and within each EC number, it is further organized by organism and by the documents from which it was extracted. One organism can have more than one enzyme with the same EC number, but since in the primary literature EC numbers are rarely associated to specific sequences, discriminating between the enzymes with the same EC number is not possible [16]. BRENDA's contents are delivered in two text files.

BioCyc is a collection of Pathway/Genome Databases (PGDBs), quite popular among biological researchers [17]. It is a very extensive data source containing more than 160 different PGDBs, each one covering a specific organism. BioCyc repositories contain information about most eukaryotic and prokaryotic species whose genome has been sequenced [18]. BioCyc is available in the form of database dumps or as flat files (the latter was chosen).

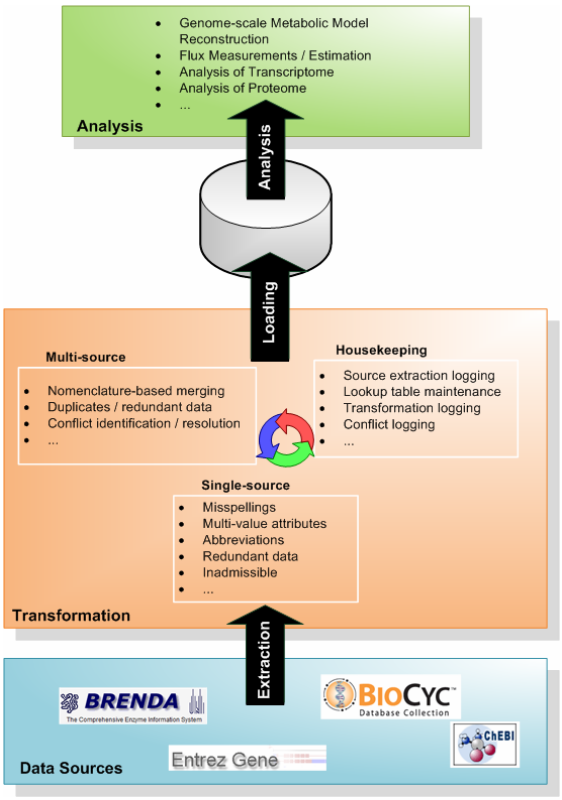

Fig. 2. Workflow for the data integration process

3.2 Data Quality Issues

Data quality issues can be classified into single-source and multi-source issues and, within these, schema and instance-related issues. The quality of a data source depends on schema and integrity constraints. Sources without a schema (e.g. flat files) raise a higher number of errors and inconsistencies. Database systems are expected to enforce a data model and application-specific integrity constraints. Usually, schema-related problems (e.g. uniqueness and referential violations) occur due to the lack of integrity constraints, data model limitations and poor schema design or because integrity constraints were limited to prevent control overhead.

Single-source quality issues get worse when performing data integration. Each source may contain dirty data and the data in the sources may be represented differently, overlap or contradict. At the schema level, data model and schema design differences are to be addressed by the steps of schema translation and schema integration, respectively. Naming conflicts arise when the same name is used for different entities (homonyms) or different names are used for the same entity (synonyms). Structural conflicts occur in many variations and refer to different representations of the same entity, different component structure, different data types, different integrity

constraints, etc. In addition to schema-level conflicts, data problems from single sources can occur with different representations in different sources (e.g., duplicated or contradicting records).

KEGG. The initial analysis of KEGG (Table 1) revealed some issues in the compound data, namely some data such as mass and formula are missing. However, the percentage of records with missing data is relatively small.

Table 1. Characteristics of the KEGG compound data for *Zymomonas mobilis*

Number of compounds	15050
Compounds without formula	2322
Compounds without mass	3457
Compounds without formula & mass	2322

In KEGG, it is considered that in a organism there is only one enzyme for each EC number. This fact makes it difficult to integrate KEGG with data sources that have more detailed enzymatic information (like BioCyc). Also, the study of reaction-enzyme associations (Table 2) revealed that some enzymes in catalogue are not associated with any reaction. One serious problem in KEGG is that pathway data is stored in an image format, hard to parse, and consequently, to combine with other formats of data. This situation appears to be changing however since the KEGG pathway data is being migrated into an XML format (KGML).

Table 2. Characteristics of the KEGG enzymatic data for *Zymomonas mobilis*

Number of enzymes	413
Number of reactions	942
Number of enzymes associated with reactions	380

BRENDA. Two main difficulties were found during BRENDA's data processing: (1) EC numbers are formal record identifiers, but often the identifier field has comments that constrain the integration of the records; (2) the identification of the compounds that affect the enzymes (e.g. inhibitors, cofactors and activators) is only provided by name. After analyzing EC number field values, it was possible to establish a parsing schema that allows adequate record crossing. Compound name resolution is far more delicate and brings a high error probability to the integration. Since the names are the only identifiers available there was no other choice but to use them as the basis for the integration. In order to verify the viability of the integration, a terminology comparison study was undertaken (Tables 3 and 4).

This study showed that there are relatively few name conflicts when the BRENDA compound data is compared with KEGG and BioCyc. Consequently, the integration of the BRENDA data with the information from these two databases should not be a problem. However, KEGG and BioCyc are not databases specific for compounds, so

Table 3. Number of compounds in BRENDA entries about *Zymomonas mobilis* and the number of compounds successfully associated/number of conflicts found during the association of the BRENDA compounds with information from other data sources

Compounds found / Conflicts	N compounds	KEGG	BioCyc	CHEBI
BRENDA cofactors	13	13 / 0	7 / 0	8 / 5
BRENDA inhibitors	52	37 / 3	15 / 0	13 / 24
BRENDA activating compounds	11	5 / 0	1 / 0	2 / 4

there is no guarantee that all or even most of the possible names for a compound are present. For this reason, the BRENDA data was also compared with ChEBI, a compound specific database [19]. The larger number of conflicts obtained with ChEBI leads to the conclusion that the integration of the compound data from BRENDA with the one extracted from other data sources, given the redundancy of the names of the compounds, is far more difficult that it was originally expected. In fact, the only way of insuring a correct integration is manual curation. Another potentially serious problem with data obtained from BRENDA is the fact that in this database there is only a small quantity of information about *Zymomonas mobilis*, specifically only 29 entries.

BIOCYC. The greatest obstacles to the integration of the BioCyc and KEGG genomic data is the scarce use of standard identifiers. The only common gene is the name of the genes and this identifier, similarly to the name of the compounds, is subject to a great degree of redundancy. Furthermore, only some of the genes have names associated in the database. In fact only 663 out of 1998 genes can be integrated (Table 4). Since there is no way to solve the problem only with the information from BioCyc and KEGG, the possibility of using a third database should be considered, preferably with links to both BioCyc and KEGG.

Table 4. Characteristics of the BioCyc genomic data for *Zymomonas mobilis*

Total number of Genes	1998
Genes with no redundant name	663
Genes with redundant names	54

Another problem with the BioCyc data is the fact that the EC number is not associated with the enzymes but rather with reactions. This problem is solved, since when a reaction has an EC number it can be considered that all enzymes that catalyze that reaction have that EC number. This method allows the association of an EC number to most enzymes identified in the BioCyc data (Table 5).

The integration of the KEGG and BioCyc pathway data presents another challenge, because the information is kept in quite different formats. In KEGG, this data is kept as images with hyperlinks in certain regions and in BioCyc the data is stored in text format. Because of their differences, it is not possible to integrate both types of data and it will be necessary to include both or choose only one. There was one more difficulty with the BioCyc database: the references to external database sources in BioCyc

Table 5. Characteristics of protein and enzyme data from BioCyc for *Zymomonas mobilis*

Proteins	2007
Enzymes	610
Reactions	880
Enzyme-reaction associations	837
Reactions with enzymes	636
Enzymes to witch may be associated an encumber	542

are stored in an unusual format: (PID "56544468" NIL |kaipa| 3390578134 NIL NIL). It was found that the first word is a code for the database and the second is the database id. Since there is not data that associates the database code to the corresponding database in BioCyc the association has to be done manually (through consulting the BioCyc site) before integrating the data. Fortunately, there are only a few databases codes in BioCyc.

3.3 Integration Strategy

The data integration strategy is divided in three stages, each corresponding to the integration of one of the data sources into a shared database, the Data Staging Area (DSA). The data sources are loaded to the DSA in a specific order: KEGG first, then BioCyc and finally BRENDA. When information from a data source is added into the DSA the data is compared with the ones already present in the DSA. It is then added if it is not present or used to complement the information already in the DSA in the case that part of the new information is already present. The new data is compared with the data already in the DSA normally by the use of non redundant identifiers (e.g CAS number for compounds) or by the comparison of a set of factors that combined can be used as a nearly non redundant identifier.

The loading of the KEGG information into the DSA is the first step. There is a problem in the KEGG data that affects the integration of BioCyc: in KEGG it is considered that only one enzyme is associated to one EC number. This problem was solved by observing that in KEGG genes, the product of the gene is identified, including its EC number (in the case it is an enzyme). To overcome this difficulty, we cross the information from the KEGG enzyme and gene data, to identify the individual enzymes associated with an EC number. After solving this issue, the remaining of the loading is handled smoothly.

The integration of BioCyc starts with the compound data, since it is the easiest to integrate, followed by the gene and the protein data. This is a two step process: in the first phase, all the proteins are compared with the ones already in the DSA, through the genes that code them to determine which ones are added and which are complemented; the second step is the identification of the new enzymes that can only be executed during the integration of the reaction data. Next, the reaction data is integrated. The reactions are not directly compared with the ones already in the DSA, instead they are compared with the enzymes that catalyze them and their reactants and products. The process ends with the integration of the pathway data.

The integration of BRENDA starts by comparing the EC numbers in its entries with the ones in the DSA; when the values match, the BRENDA data is associated

with that record. BRENDA data includes references to compounds that affect the reaction, such as inhibitors or cofactors. Determining if these compounds are already in the DSA or if they must be added is difficult, since the only identification is the name of the compound. Since there can be multiple enzymes associated with EC numbers, this means that the same BRENDA data will probably be linked with the same enzymes.

4 Conclusions

In this work, we approached a number of issues related to the process of genome-scale reconstruction of the metabolic model of the bacterium *Zymomonas mobilis.* These were mainly related to data quality issues and the implementation of suitable data integration processes. A number of problems were identified and useful guidelines for their solution were proposed. This work is on-going and it will proceed by enlarging the set of handled conflicts and integrating other data sources.

Acknowledgements

The authors acknowledge the support from the Portuguese FCT under the project POCI/BIO/60139/2004 and the PhD grant (ref. SFRH/BD/41763/2007).

References

1. Rocha, I., Forster, J., Nielsen, J.: Design and application of genome-scale reconstructed metabolic models. Methods Mol. Biol. 416, 409–431 (2008)
2. Notebaart, R.A., van Enckevort, F.H., Francke, C., Siezen, R.J., Teusink, B.: Accelerating the reconstruction of genome-scale metabolic networks. BMC Bioinformatics (7), 296 (2006)
3. Seo, J.S., Chong, H., Park, H.S., Yoon, K.O., Jung, C., Kim, J.J., Hong, J.H., Kim, H., Kim, J.H., Kil, J.I., Park, C.J., Oh, H.M., Lee, J.S., Jin, S.J., Um, H.W., Lee, H.J., Oh, S.J., Kim, J.Y., Kang, H.L., Lee, S.Y., Lee, K.J., Kang, H.S.: The genome sequence of the ethanologenic bacterium Zymomonas mobilis ZM4. Nat. Biotechnol. 1(23), 63–68 (2005)
4. Tsantili, I.C., Karim, M.N., Klapa, M.I.: Quantifying the metabolic capabilities of engineered Zymomonas mobilis using linear programming analysis. Microb.Cell Fact. (6), 8 (2007)
5. Borodina, I., Nielsen, J.: From genomes to in silico cells via metabolic networks. Curr. Opin. Biotechnol. 3(16), 350–355 (2005)
6. GOLD (Genomes OnLine Database v 2.0) web site (2008), `http://www.genomesonline.org/`
7. TIGR web site. TIGR web site (2008)
8. NCBI web site. NCBI web site (2008)
9. BRENDA web site. BRENDA web site (2008)
10. KEGG (Kyoto Encyclopedia of Genes and Genomes) web site (2008), `http://www.genome.jp/kegg/`
11. NCBI's Entrez Gene Web site (2008), `http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene`

12. Minorou, K., Susumu, G.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 1(28), 27–30 (2000)
13. Minorou, K., Michihiro, A., Susumu, G., Masahiro, H., Mika, H., Masumi, I., Toshiaki, K., Shuichi, K., Shujito, O., Toshiaki, T., Yoshihiro, Y.: KEGG for linking genomes to life and the environment. Nucleic Acids Research 36 (2007)
14. BRENDA introduction. BRENDA web site (2008)
15. Schomburg, I., Chanf, A., Hofmann, O., Ebeling, C., Ehrentreich, F., Schomburg, D.: BRENDA: a resource for enzyme data and metabolic information. TRENDS in Biochemical Sciences 1(27), 54–56 (2002)
16. Schomburg, I., Chang, A., Schomburg, D.: BRENDA, enzyme data and metabolic information. Nucleic Acids Research 1(30), 47–49 (2001)
17. BioCyc Introduction. BioCyc web page (2008)
18. Karp, P., Ouzounis, C., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., López-Bigas, N.: Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Research 19(33) (2005)
19. CHEBI web site. CHEBI web site (2008)