Semana de Engenharia 2010
Guimarães, 11 a 15 de Outubro

Reconstructing genome-scale metabolic models with *Merlin*

Oscar Dias
IBB- Institute for Biotechnology and Bioengineering
Department of Biological Engineering
E-mail: odias@deb.uminho.pt

KEYWORDS
Systems Biology, Genome-Scale Reconstruction, BLAST, SBML, Metabolic Engineering

ABSTRACT

The reconstruction of genome-scale metabolic models is based on the well-known stoichiometry of biochemical reactions. Usually the main objective of a reconstruction is the *in silico* simulation of the phenotypic behaviour of a microorganism, under different environmental and genetic conditions, thus representing an important tool in Metabolic Engineering.

The genome of the yeast *Kluyveromyces lactis* was used as a case study for this method, providing information for the first stage of the reconstruction of this eukaryote. Given and input of 5085 gene sequences, *Merlin* identified more than 4200 distinct organisms and approximately 394.000 genes with sequence similarities to the *K. lactis* genome.

This information, after user appraisal, will be used to assemble a metabolic model with the reactions catalysed by the enzymes encoded in the genome. Such model, in the *SBML* format, can be used as a first raw approach to the study of the *K. lactis* metabolism

INTRUDUCTION

The genome-scale reconstruction of metabolic networks encompasses several steps, such as genome annotation, reactions identification and stoichiometry determination, compartmentation, determination of the biomass composition, energy requirements and additional constraints (Rocha *et al*. 2008).

The first of this type of reconstruction is the genome annotation, which is an essential step since precursory data can be retrieved for the model reconstruction. There are currently more than 4.000 fully sequenced genomes, with more than 700 being drafted right now. Hence tools such as BLAST, one of the most widely used bioinformatics tools, are being used to establish sequence similarities between genomes.

Annotating a Genome

Genome Annotation encompasses both gene finding on the sequenced genome and the assigning of biological functions to the recently found genes (Medigue and Moszer 2007; Salzberg 2007). The gene functional annotation procedure can be defined as the assignment of functional information to a specific gene. Such information is often obtained by similarity to formerly characterized sequences, in several online or local databases (Ouzounis and Karp 2002).

Genome Re-Annotation

After the initial annotation of the genome, there are several circumstances that can lead to a genome-wide re-annotation. Whether new genes or protein functions are discovered, a research group tries to determine the reproducibility of an existing annotation, or just because a specific organism information is known to be outdated, a genome-wide re-annotation will update the data assigned to such genome (Ouzounis and Karp 2002; Tamaki *et al*. 2007). However, most of the tools that perform re-annotations are aimed at genome projects and do not provide outputs that allow the easy (re-)annotation for the development of genome-scale metabolic models.

Genome-scale metabolic models reconstruction

Whereas the reconstruction of the metabolic network of a given organism is becoming a widespread procedure, starting with the fully sequenced and (partially) annotated genome sequence, there are still many improvements needed in the current methodologies and a clear lack of computational tools for many of the steps. Some organisms are more prone to rely in similarity information as such organisms are less characterized (Rocha *et al.* 2008).

For the reconstruction of a genome-scale metabolic model *Merlin* (Dias *et al.* 2010) is proposed. *Merlin* performs similarity searches for any organism that has its genome sequenced, and allows a semi-automated dynamic (re-) annotation of the genome. The semi-automated (re-) annotation is supported by a specific algorithm developed for scoring of the BLAST hits results.

MERLIN: METABOLIC MODELS RECONSTRUCTION USING GENOME-SCALE INFORMATION

Merlin is composed by two modules: the Dynamic Annotation Tool and the Model Reconstruction Tool, each of which will be further described in the next sections. The Dynamic Annotation Tool automatically annotates genes list, properly provided in the *Fasta* format. This tool performs BLAST similarity searches to the gene sequences list with user-defined parameters,

scoring the results and allowing the user to dynamically annotate each gene with a quantifiable confidence level. The Models Reconstruction Tool allows the user to load information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000), integrate it with information from the previous module and later build the metabolic model and store it in the well accepted SBML implementation.

RESULTS AND DISCUSSION

Two of the most well studied organisms were selected for *Merlin*'s validation: *Escherichia coli str. K-12 substr. W3110* and *Saccharomyces cerevisiae*.

Table 1. *E.coli* and *S. cerevisiae* matching.

	EC numbers		protein names	
	E. coli	*S. cerevisiae*	*E. coli*	*S. cerevisiae*
Match	903	842	766	674
Partial Match	22	31	-	-
Distinc	68	73	227	272
Only *Merlin*	44	64	44	64
Only BLAST	395	727	395	727
Total	1432	1732	1432	1732

As demonstrated on Table 1 the studied organisms had similar distributions for gene matching between both databases, either on EC numbers or protein names.

The results for the EC numbers integration are similar for E. coli and S. cerevisiae. For the bacterium there were a total of 1432 genes which encoded enzymes. More than 60% of those genes were assigned with the same protein by KEGG and by *Merlin's* similarity search. For the yeast, 1732 enzyme encoding genes were identified, with almost 50% of the genes being assigned with the same enzyme in both databases.

For the two organisms less than 2% of the genes assigned by KEGG were only partially matched by *Merlin's* similarity search for homologues.

For both organisms less than 5% of the genes were assigned with different EC numbers by similarity and on KEGG. Most of the cases were genes that encoded an incomplete EC number on one database and the complete EC number on the other database Over 3% of the genes were assigned as enzyme coding genes in the local database, but no similarity was found by *Merlin* when BLAST was performed. The most unexpected results were obtained on the genes that were only appraised by *Merlin's* similarity search. For E. coli, *Merlin* identified 395 candidate genes, which may encode enzymes, with scores beyond the confidence level threshold.

Moreover, for the yeast, *Merlin* identified (42%) 727 candidate genes from the 1732 total enzyme coding genes.

Conclusion

With the ever increasing amount of genomic data becoming available, every tool developed to interpret and make sense of such data is greatly appreciated, as appraising such bulk loads of data can be very tedious and time consuming.

Merlin is proposed as a user-friendly tool, which allows to attain comprehensible information and perform a semi-automated dynamic annotation, relying in the most up to date information, available in the GenBank database, and integrate such data with the data already available at the well accepted KEGG database, for the development of a more robust metabolic model. Moreover, such model may be retrieved in the *Systems Biology Markup Language* for in silico processing.

Merlin obtains the most up-to-date information from the online databases, allowing the user to perform regular similarity searches and update the genome annotation.

REFERENCES

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. NUCLEIC ACIDS RESEARCH, 28(1), 27-30.

Medigue, C. and Moszer, I. (2007). Annotation, comparison and databases for hundreds of bacterial genomes. RESEARCH IN MICROBIOLOGY, 158(10), 724-736. doi:10.1016/j.resmic.2007.09.009.

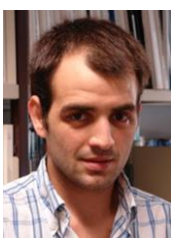
Ouzounis, C. and Karp, P. (2002). The past, present and future of genome-wide re-annotation. Genome Biology, 3(2). doi:10.1186/gb-2002-3-2-comment2001. URL:http://genomebiology.com/2002/3/2/comment/2001.

Dias O., Rocha M., Ferreira E.C., Rocha I. (2010). Merlin: Metabolic Models Reconstruction using Genome-Scale Information. Proceedings of the 11th Computer Applications in Biotechnology International Symposium, Leuven, Belgium, 7-9 July, 2010, 120-125.

Rocha, I., Förster, J., and Nielsen, J. (2008). Design and application of genome-scale reconstructed metabolic models. Methods in molecular biology (Clifton, NJ), 416, 409.

Salzberg, S.L. (2007). Genome re-annotation: a wiki solution? GENOME BIOLOGY, 8(1). doi:10.1186/gb-2007-8-1-102.

Tamaki, S., Arakawa, K., Kono, N., and Tomita, M. (2007). Restauro-G: a rapid genome re-annotation system for comparative genomics. Genomics, Proteomics & Bioinformatics, 5(1), 53-58.



Oscar Dias was born in Braga, Portugal went to the University of Minho, where he studied biological engineering and obtained his degree in 2005. He then studied Informatics and obtained his Masters Degree in 2008. Now he is undergoing a PhD in Chemical and Biological Engineering under the Supervision of Eugénio C. Ferreira (IBB - UM), Isabel Rocha (IBB - UM) and Andreas K. Gombert (USP – Brazil).