Reconstruction of a Genome-scale Metabolic Network for Streptococcus pneumoniae R6

João Saraiva, Francisco Pinto, Isabel Rocha
IBB - Institute of Biotechnology and Bioengineering

Abstract

The gram-positive, lancet-shaped bacteria Streptococcus pneumoniae thrives in almost any environment. Under certain conditions this pathogen can cause several infections such as meningitis, otitis media, endocarditis or pneumonia.

Genome-scale metabolic networks (GSMs) are commonly used to study phenotype-genotype relationships using biochemical, physiological and genomic information. These relationships might shed some light on identification of targets for metabolic engineering or, in the case of S. pneumoniae, determine if the bacteria´s increased invasiveness and virulence is dependent on specific genomic regions or determined by environmental conditions.

In order to obtain a robust and reliable metabolic model, a proper, up-to-date genome annotation must be performed.

In our work we aimed to re-annotate the genome of Streptococcus pneumoniae strain R6 and which would be used to reconstruct a metabolic network at a genomic level. For these tasks merlin was used, a software tool capable of performing automatic annotation of the genome using the amino acid sequences as well as reconstruction of the metabolic network. For validation purposes, another in-house tool (Optflux) capable of performing simulations and optimization tasks was used.

Re-annotation of the genome was performed in accordance to an in-house generated pipeline which established rules for gene identification acceptance (and attribution of confidence levels) or rejection. Out of the 2043 genes present in S. pneumoniae´s genome, an initial 822 were identified as metabolic, representing an increase of almost 9 and 15% when compared to those of KEGG and Uniprot. An extended comparison revealed that a large number of genes (359 and 271 when compared to Uniprot and KEGG, respectively) were only present in our re-annotation. Although a significant amount of genes (113) were identified as only being present in KEGG and not in our study, this can be explained by the dismissal of genes associated to DNA and RNA processes from the statistical analysis.

The metabolic network is comprised of 795 genes, 776 that only encode enzymes and 19 that only encode transporters. The biomass equation was adapted from close-related organisms such as B. subtilis and L. lactis cross-referenced with the biomass equation determined by ModelSEED for S. pneumoniae R6.

Despite the considerable amount of essential genes in our model (83), only 38 were in accordance to literature regarding gene essentiality although it identified others (45) which have not been studied to date. The mismatch between results might be related to strain metabolic specificities, regulatory phenomena or even the dismissal of genes that affect DNA and RNA processes and capsule synthesis which should be addressed in future work.

In order to validate the accuracy of the model, simulations were performed using experimental data retrieved from literature. The results obtained were very similar to the ones described in in vitro studies elevating the confidence level of the reconstructed model.