Reconstructing the Metabolic Network of a Bacterium from Its Genome

Marko Laakso

Helsinki February 25, 2007

Seminar report

UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author
Marko Laakso

Työn nimi — Arbetets titel — Title
Reconstructing the Metabolic Network of a Bacterium from Its Genome

Oppiaine — Läroämne — Subject
Bioinformatics

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Seminar report	February 25, 2007	11 pages

Tiivistelmä — Referat — Abstract

Bacteria form a significant part of the biomass on Earth and they have a huge impact on healthcare, status of the environment and industrial processes. Automated methods are needed in the studies of these widely varying species. This paper gives an overview of the reconstruction procedure that can be used to create metabolic networks of bacteria based on the knowledge about other related species. The original review about this procedure has been given by Francke, Siezen, and Teusink [FST05].

The reconstruction of a metabolic network involves the identification of genes, comparison of these genes against the genes of other related organisms and the reconstruction of the pathways. The pathways are reconstructed based on the functional annotations of the genes. It is assumed that a similar gene of the studied bacterium carries out the same reactions as has been documented for the corresponding gene in other organisms.

The loci of genes of a bacterium can be predicted based on the known promoter and stop sequences. The predicted genes are then aligned against the genes of other organisms. Enzyme activities of the well aligned genes are imported into the model and an initial model is constructed based on the reactions that could be catalyzed by these enzymes.

Further curation of the model is based on the verification of the predicted reactions and the model integrity in terms of a balance between the reactants and reaction products.

ACM Computing Classification System (CCS):
I.6.4 [Model validation and analysis]
J.3 [Life and medical sciences]

Avainsanat — Nyckelord — Keywords
bioinformatics, metabolism, sequence alignment

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1 Introduction

Bacteria can be found practically everywhere in the biosphere. These tiny unicellular organisms make over 90% of the cells of our bodies [BM02] and they are responsible for many vital roles (such as the production of organic nitrogen) of the surrounding ecosystems. Some bacteria are known pathogens of humans, domestic animals and cultivated plants and some are used to drive industrial processes such as waste water purification, food production and synthesis of many chemicals. It is no wonder that the properties of these creatures are of great interest.

Reconstruction of the metabolic networks is a process, which aims to model biochemical reactions that may take place within an organism. This seminar report explains some bioinformatics aspects of the process. The original review of this subject has been published in [FST05]. The review has been used as a primary source of information for this paper.

Metabolism refers to chemical reactions that take place within the living organisms. *Metabolic network* describes a set of biochemical reactions and the relations between them. Each reaction is bind to its reactants and produced compounds. A network is formed as some reactions are producing compounds that are consumed by other reactions. Reactions of the metabolic network are catalysed by enzymes, which are linked to the reactions. The network can be seen as a representation of the dependencies between the reactions as it describes which reactions are affecting the substrate concentrations of the reaction under concern. An example of graph reactions and reactants is shown in Figure 1. Each reaction has been labelled with the associated enzyme numbers. The numbering scheme is explained in Section 3. Reactants have not been labelled into the image as the molecules have quite long and complex names.

The *genome* of a bacterium represents all the genetic material of the particular species. The bacteria are so called *procaryotic* organisms which means that they lack many cellular

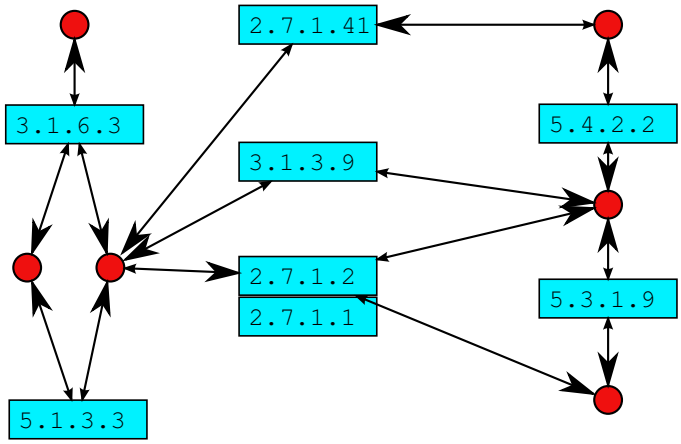

Figure 1: An example network of some compounds (red circles) and reactions (cyan rectangles).

inner compartments like nucleus and endoplasmic reticulum and Golgi apparatus that are in important roles in the regulation of the metabolism of eukaryotic cells. All the circular chromosomes of bacteria are in their cytoplasm. Size of the bacterial genome is usually less than that of eukaryotes but there are exceptions. The shortest bacterial genome that has been found belongs to *Mycoplasma genitalium* and it consists of 0.58 million base pairs (Mbp). On the other hand, some bacteria are known to have genomes of more than 10 Mbp. The known lengths of the eukaryotic genomes range from 2.9 Mbp of some parasites to some plants with over 4000 Mbp [Mal06].

2 Identification of the Genes

Reconstruction of the metabolic network of a new bacterium starts from the analysis of its genome. First, the whole genome is sequenced. Second, possible sequences of genes are identified from the genome. Third, gene candidate sequences are compared against the sequences of known and closely related organisms. Fourth, the original gene candidate is annotated based on the information that is known of the best matching genes of the other organisms.

Genes of a bacterial genome are easier to predict than those of eukaryotes as the mechanisms of the transcription and translation are straighter forward in prokaryotes. Prokaryotic genes are usually initiated by a well known promoter sequence and the coding sequence contains no introns[1]. The coding sequence is terminated using a special stop codon.

A sequence of a gene that has been found from the bacterium can be compared against the sequences of the other related organisms that are known. The sequence alignment and similarity comparison is motivated by the evolution theory that supports a hypothesis about a common origin of the organisms. A gene of interest may have evolved already in the common ancestor of both organisms and thus its homologies may still have the same function in both organisms. The homologous (that is they have a common origin) genes of two organisms are called *orthologous* if they have evolved from the common gene of their latest common ancestor and if they have not undergone a duplication event. The homologous genes of an organism that have evolved from the chromosomal duplication of a gene are called *paralogous*. Paralogous genes are likely to participate in different cellular functions as the organism is given a change to alter the function of each copy without losing the pathway supported by the other copy.

Pairwise sequence alignment is an optimisation procedure that tries to align two character sequences so that the characters of the corresponding indices match the best. Usually, gaps are allowed not only in the beginning and end of the sequences but also in the middle of both sequences. Two sequences of n and m characters can be aligned $\binom{n+m}{m}$ ways preserving the order of the characters. Several so called dynamic programming algorithms exists that can be used to find the optimal solution. *Dynamic programming* refers to it-

[1]Translation of an eukaryotic gene leads to a premature messenger RNA, which is complementary to the encoding DNA. Some parts of the RNA are spliced out before it gets delivered to the protein transcription machinery. The spliced out regions are called *introns*, while the remaining protein coding sequences are referred as *exons*.

erative algorithms that advance by deriving new values from the values that have been calculated already. The complexity of the alignment algorithms is $O(n \times m)$ if they are using dynamic programming.

Similarities between the characters of the sequences can be given in a form of a scoring matrix. The matrix consists of the alphabet on its both axis and each element represents the corresponding similarity of the row and column character. The characters can be the nucleotides of the DNA (adenosine, thymine, cytosine, and guanine) or they may represent the amino acids of the proteins. A reasonable scoring matrix obtains high scores for the identical characters and lower scores for those combinations that are biologically unlikely to substitute each other.

The alignment algorithms can be divided into two categories based on the alignment of interest. *Global alignments* are complete alignments of two strings. For example two coding regions of a gene can be compared using their global alignments. *Local alignments* are alignments of subsequences of two strings and they can be used to find common regions in these strings. A gene can be compared against a whole chromosome to find possible homologies, for instance.

Basic local alignment search tool (BLAST) [Alt90] is a widely used algorithm for the sequence comparisons across species. BLAST takes a heuristic approach to sequence alignment by restricting the actual aligning to those parts of the sequence that have the most probable candidate alignments. The candidate alignments are constructed by creating a table of high scoring alignments of short (for example 11 base pairs) subsequences of the other string and the corresponding locations for the other string. Each candidate region is extended from both ends by maximising the ungapped alignment. Clearly, BLAST is capable in sorting the extended alignments based on their scores and a rank of possible solutions can be obtained to the user [Dur98].

An alternative sequence align method to BLAST is FASTA [PL88]. FASTA starts by

searching for the perfect matches of short sequences, say 4–6 base pairs. The short matches are extended and scored. Alignments that do not exceed a given threshold are dropped and a dynamic programming algorithm is applied to sequence ranges around sequences with well scored alignments. The dynamic programming may find continuous alignments combining multiple seed alignments together but the optimal alignment cannot be guaranteed due the heuristic restriction to well scored regions.

Proteins that are not known from the previously studied organisms can be profiled against other proteins for common residues or motifs. Sequence comparisons employ hidden Markov models (HMM). Hidden Markov models are statistical models that can be used to find most probable interpretations for different parts of sequences or likelihoods that sequences represent the given phenomenon such as certain active parts of a protein.

Hidden Markov models consist of states, transition probabilities between the states and emission probabilities for having a certain value in each state. For example, a HMM can be constructed so that the sequence consists of amino acids coded by an unknown gene. States may represent structural sites of amino acids in known protein structures and the emission probabilities would be the probabilities of observing different amino acids within the given state. The transition probabilities would combine states so that the transitions from the previous position of the same structure would be more probable than a transition to that of a complete different structure. Algorithms such as Viterbi [Vit67] can be used to find the most probable sequence of states for the observed sequence of emitted values (structural sites of the protein). The algorithm keeps list of probabilities of possible state sequences during the iteration of the amino acids and selects the most likely transition based on the transition probability and the probability of emitting the observed value in that state.

3 Merging Genes to Metabolic Reactions

The initial annotations of the predicted (or found) genes provide a set of proteins. Many proteins are known to exhibit catalytic activity within the cell and they are termed *enzymes*. Information about the enzyme activities of the known proteins have been collected to many biological databases such as ENZYME [Bai00], IntEnz [Fle04], Uniprot [Bai05], etc. A single protein may have several roles within the cell albeit they are famous for their substrate specifity.

Enzyme activity of the protein means that the cell is able to carry out certain metabolic reactions in the presence of the protein. In theory, all reactions that are accelerated by a catalyst could happen even if the catalyst is absent or if it is in an inactive state. In practise, the lack of an enzyme would make many biological reactions so low that they would be of no significance for the organism or its behaviour. A missing enzyme in a pathway could lead to accumulation of the substrates if the substrates are produced by an energetically favoured reaction. At the same time, the downstream reactions would be ceased in the lack of their substrates that would be the products of the stalled reaction.

Existence of an enzyme coding gene may suggest that the organism has a pathway with a reaction that is catalysed by the enzyme. On the other hand, a metabolic pathway known from another organism is unlikely used by the bacterium if it lacks an enzyme that would be needed by a reaction of the pathway. An initial reconstruction of the metabolic network can be done by simply binding enzymes to their substrates and end products. The initial model can be compared against the known and expected pathways to see if there are some interesting alterations or missing reactions. KEGG [Kan06] is one of the most famous databases for known biological pathways.

Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) provides a widely used nomenclature for the enzyme activities [Nom92]. The standard defines a hierarchical naming convention that consists of four layers. The

layers are used to form so called EC codes that have four decimal numbers separated with dots. The left most number represents the most generic grouping of the enzyme catalyst reactions, which are: oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. The next number divides each of these groups into more detailed groups and the next numbers are used for ever finer clustering of reaction. The EC numbers provide a direct mapping from proteins to their roles in pathways.

4 Model Curation

The content of biological databases is usually based on predictions and measurements, which are both likely to be inaccurate or misleading. Yet, different databases may use the same concept or identifier in different meanings or the same thing has been stored in multiple copies with different identifiers. It is not only the source information of the databases that may cause errors in the metabolic network but the actual assumption that the role of the homologous genes is preserved. Currently, all these issues cannot be solved automatically and some manual curation of the metabolic model is needed.

The manual curation involves the verification of the sources of information that has been used for each reaction. Practical implication of this step is that the model be constructed with backward references pointing to the original sources of information. The inclusion of these references into the model helps in testing the model. The backward references provide a systematic way of integrating sources of information and they can be used by other applications that are working on top of the metabolic model.

Automated curation of the metabolic model may be based on comparative genomics. In this approach, the model assumptions, such as the functions assigned to the genes, are weighted with the context of the gene in other species [FST05, OO03]. Some of the possible weightening criteria are listed below.

1. Phylogenetic relatedness of the organisms provides a base measure for the plausibility of the annotations.
2. Order of the genes has an important role in the regulation of bacterial genes and thus a conserved order may be related to the function of the genes.
3. Common regulatory motifs in genes may indicate relatedness to the same pathway.
4. Fusion of genes may give arise to hybrid proteins or alterations in their regulations.
5. Essential genes are more likely to be conserved.
6. Chemical properties (such as charge) of the protein may be compared.
7. Expression patterns of co-expressed proteins of the same pathway and protein-protein binding studies with results that are compatible with the model.

STRING [vM05] provides an integrated resource for the connectivity between the proteins. The database can be used to identify links between the proteins and these links can be further applied into the weightening function. For instance, if an active enzyme requires some other proteins the presence of these proteins is bacteria may form a crucial invariant for the model. Several proteins may be needed for a functional complex but the dependencies may be caused by the gene regulatory system or chaperones or other proteins that are involved in the activation of the enzyme.

A solid metabolic network should include all pathways that are vital to the bacteria and the model should be consistent with its known physiology. An elemental balance is required so that the reactions consume and produce equal amounts of elements. A pathway will not work properly unless all substrates are available. A substrate of a reaction should be produced by another reaction or taken in from the environment. Some of the gaps in pathways may be explained by alternative substrates that are different from the expected ones.

Studies of evolution have revealed many cases of convergent evolution where two distinct organisms have developed similar structures or capabilities due the adaptation to similar

selection [CRM99, p. 476–477]. Sometimes, the convergent evolution leads to enzymes that are not homologous but still capable to catalyse same reactions. Genes responsible for these enzymes are called *analogous* and they should be taken into consideration when there are gaps in pathways. Changes are that the missing reaction is actually carried out by a gene that has not been annotated with a matching enzyme activity. Finding out the actual functions for unknown genes is likely to involve some wetlab activities but the initial model of the metabolic network may help in the construction of fruitful hypothesis. The model may for example point out some possible gaps or new branches in pathways.

5 Conclusions

Metabolic networks of new bacteria may be reconstructed using the existing information about the metabolism of the related species. The application of the pre-existing knowledge may give rise to issues of its suitability for the bacterium in concern. Fortunately, some of the issues can be tackled with further annotations and manual curation of the data. Experiments may be needed where the predicted model contains gaps between the metabolic reactions or ambiguities.

The authors of [FST05] suggest that the construction of the metabolic network starts right after the initial annotation of the protein coding genes found from the bacterium that is studied. The reasoning for the early construction is that the metabolic network provides a useful model for further annotations and hypothesis, which aid in the gradual curation of the model.

The model construction is a practically endless procedure that may lead to ever more accurate predictions about the underlying cell chemistry. The construction of the pathway topology can be followed by the flux analysis modelling of the actual reaction rates and the regulation of the pathways. The metabolic network of the enzymes contains less than 30% of all genes of a bacterium [FST05]. The other genes may be responsible for the

gene regulatory system or they may encode structural parts of the cell.

References

Alt90 Altschul, S. e. a., Basic local alignment search tool. *J Mol Biol*, 215,3(1990), pages 403–410.

Bai00 Bairoch, A., The ENZYME database in 2000. *Nucleic Acids Research*, 28,1(2000), pages 304–305.

Bai05 Bairoch, A. e. a., The universal protein resource (UniProt). *Nucleic Acids Res*, 33,Database issue(2005). URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=15608167.

BM02 Bower, J. and Mulvey, M., Gripping tales of bacterial pathogenesis. *Cell*, 111,4(2002), pages 447–448. URL http://www.ingentaconnect.com/content/els/00928674/2002/00000111/00000004/art01119.

CRM99 Campbell, N., Reece, J. and Mitchell, L., *Biology*. Benjamin/Cummings, 1999.

Dur98 Durbin, R. e. a., *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, July 1998. URL http://www.amazon.co.uk/exec/obidos/ASIN/0521629713/citeulike-21.

Fle04 Fleischmann, A. e. a., IntEnz, the integrated relational enzyme database. *Nucl. Acids Res.*, 32, pages 434–437.

FST05 Francke, C., Siezen, R. and Teusink, B., Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 13,11(2005), pages 550–558. URL http://dx.doi.org/10.1016/j.tim.2005.09.001.

Kan06 Kanehisa, M. e. a., From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34,Database issue(2006). URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=16381885.

Mal06 Maloy, S., Microbial genetics, 2006. http://www.sci.sdsu.edu/ smaloy/MicrobialGenetics/. [06.02.2007].

Nom92 Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, *Enzyme Nomenclature*. Academic Press, San Diego, California, US, 1992. ISBN: 0-122-27164-5.

OO03 Osterman, A. and Overbeek, R., Missing genes in metabolic pathways: a comparative genomics approach. *Current Opinion in Chemical Biology*, 7,2(2003), pages 238–251. URL http://www.ingentaconnect.com/content/els/13675931/2003/00000007/00000002/art00027.

PL88 Pearson, W. and Lipman, D., Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*, 85,8(1988), pages 2444–2448. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=3162770.

Vit67 Viterbi, A., Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, pages 260–269.

vM05 von Mering, C. e. a., STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33,Database issue(2005). URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=15608232.