Tutorial on *Network Genomics*

for ISMB 2001

by

Christian Forst
Bioscience Division, Mailstop M888
Los Alamos National Laboratory, Los Alamos, NM 87544
Tel.: (505) 665-5268; FAX: (505) 665-3024; E-mail `chris @ lanl.gov`

Tutor

I am staff scientist in Bioscience division at the Los Alamos National Laboratory. Here I oversee a bioinformatics group maintaining a database on pathogens related to sexually transmitted disease (http://www.stdgen.lanl.gov). I supervise annotation and genomic analysis of microbial genomes. My main research interest is "Network Genomics", the analysis of genomes in the context of metabolic and regulatory networks. I am trained as chemist and have a background in dynamical systems, complex dynamics, optimization in combinatorial landscapes, graph-theory, sequence/context analysis, whole genome annotation, network reconstruction, gene-expression analysis and phylogeny.

Tutorial's presentation

The information provided by completely sequenced genomes can yield insights into the multi-level organization of organisms and their evolution. At the lowest level of molecular organization individual enzymes are formed, often by assembly of multiple polypeptides, and at a higher level, sets of enzymes group into metabolic networks. Such context information of metabolic networks combined with genomic context on co-occurrence of genes, fusion of genes or gene-order are powerful approaches in prediction of new functional features and in the analysis of multi-level relations between organisms. I will review recent developments and discuss in some detail novel approaches that take into account network information and generic context information. I will discuss gene-expression analysis in connection with metabolic networks. Special reference to relationships between gene-context/operon structure and networks will be made. For this purpose I present a method that extend the conventional sequence comparison and phylogenetic analysis of individual enzymes to metabolic networks. As an example, I investigated the tryptophan biosynthesis pathway, which connects with the serine salvage pathway and the pathway of serine biosynthesis.

Intended audience

The tutorial is aimed at an audience with bioinformatics/computational biology background. Basic knowledge in molecular biology/genetics as well as in sequence analysis is required.

Length

The length of the tutorial is intended to be half day.

Outline

Introduction

I will introduce genomic and non-genomic context information and will browse through the buzz-words that will be referred to in detail in the tutorial. The presented approaches are mainly based on computational techniques, although experimental evidences, obtained from literature, databases and, ideally, from close collaboration with experimentalists are essential for analysis. Special emphasis will be placed on the difference between *context-based* function prediction and homology-based function prediction. Also the role of networks as high-level context information will be addressed.

The tutorial comprises of following sections; the first two sections will discuss the differences between genomic and non-genomic based context information, section three will cover combined methods. Finally, section four lists web-resources and databases. All presented approaches extensively employ comparative methods.

Gene Context

In this section the primary goal will be to discuss different benefits between homology-based and context-based genomic information. The audience will learn the different grades of genome-based contextual information based on genomes such as phyletic profiles, co-occurrence, conserved gene order/conserved operons and gene fusions. The *Rosetta Stone* approach, based on gene-fusion events will be discussed. I will present techniques how to identify the different grades of genome-based context information and how to use context-information for high-level annotation and analysis. Examples will be used to illustrate the advantages of different context information.

Networks as Non-genomic Context Information

In contrast to genomic context metabolic and gene-regulatory networks are representatives of non-genomic context information. Protein-protein interaction networks will be included as an intermediate between genomic and non-genomic context information.

I will introduce network context by employing protein-protein networks and their identification. A powerful approach to identify protein-protein interaction is by domain fusion analysis. Very briefly, prediction of protein-protein interaction by molecular modeling will be mentioned. The relation between protein-protein interaction and enzyme complexes in metabolic networks will be made.

A brief overview on metabolic networks will be given and metabolic profiling techniques in metabolic reconstruction for identifying protein functions will be presented.

Finally in this section, reference to gene-expression profiles and prediction of gene-networks will be made.

Combining various types of context info

A higher predictive value assemble combined approaches between different grades of context information. Of special interest are combinations between genomic and non-genomic context information for high-level annotation and analysis. Similar to the previous sections, examples are extensively used to illustrate the benefit for bioinformatics research of such combined techniques.

I will make the transition from networks to include genomic based context by referring to the connection between protein-protein interaction and metabolic networks.

With the new technique of comparative network genomics, the quantitative combination between genomic information and network connectivity, relationships between operon conservation and metabolic networks will be discussed. Also, reference to gene-expression pattern of adjacent genes will be made.

Superposition of gene-expression information onto metabolic networks represent the combination of two non-genomic based context information. Main focus for such an approach is the analysis of organismic response to environmental stress.

Web-based information, databases

A list of web-based tools and databases relevant to network genomics will be presented.

References

[1] L. Aravind. Guilt by association: Contextual information in genome analysis. *Genome Res.*, 10:1074–1077, 2000.

[2] J.L. DeRisi. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[3] C.V. Forst and K. Schulten. "evolution of metabolism: A new method for the comparison of metabolic pathways using genomic information. *J. Comp. Biol.*, 6:343–360, 1999.

[4] C.V. Forst and K. Schulten. Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.*, 2001. in press.

[5] M. Huynen, B. Snel, W. Lathe III, and P. Bork. Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. *Genome Res.*, 10:1204–1210, 2000.

[6] E.M. Marcotte, M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402:83–86, 1999.

[7] S. Tsoka and C.A. Ouzounis. Recent developments and future directions in computational genomics. *FEBS Let.*, 480:42–48, 2000.