CHAPTER 1

Reliable and Specific Protein Function Prediction by Combining Homology with Genomic(s) Context

Martijn A. Huynen,* Berend Snel and Toni Gabaldón

Abstract

Completely sequenced genomes and other types of genomics data provide us with new information to predict protein function. While classical, homology-based function prediction provides information about a proteins' molecular function (what does the protein do at a molecular scale?), the analysis of the sequence in the context of its genome or in other types of genomics data provides information about its functional context (what are the proteins' interaction partners, and in which biological process does it play a role?) Genomic context data are however inherently noisy. Only by combining different types of genomic(s) context data (vertical comparative genomics) or by combining the same type of genomics data from different species (horizontal comparative genomics) do they become sufficiently reliable to be used for protein function prediction. Homology-based function prediction and context-based function prediction provide complementary information about a protein's function and can be combined to make predictions that are specific enough for experimental testing. Here we discuss the genomic coverage and reliability of combining genomics data for protein function prediction and survey predictions that have actually led to experimental confirmation. Using a number of examples we illustrate how combining the information from various types of genomics data can lead to specific protein function predictions. These include the prediction that the Ribonuclease L inhibitor (RLI) is involved in the maturation of ribosomal RNA.

Introduction

Genome sequencing provides us with an abundance of genes whose functions are not determined experimentally and have to be predicted by bioinformatics. The classic tool to do so, homology detection, is mainly suited to predict the molecular function of a protein. Having complete genome sequences we would also like to know protein function at a higher level, like the pathway or complex a protein belongs to.[1] Bioinformatics supplies us with a growing number of so-called genomic context methods that exploit the genomics data themselves to predict such interactions. These methods exploit the fact that the genes of functionally interacting proteins tend to be associated with each other in genomes or in other types of genomics data. At the level of genome sequences, gene fusion,[2,3] the conservation of gene order,[4,5] the co-occurrence of genes among sequenced genomes,[6,7] or genes having a

*Corresponding Author: Martijn A. Huynen—Nijmegen Center for Molecular Life Sciences, p/a Center for Molecular and Biomolecular Informatics, Toernooiveld 1, 6525 ED Nijmegen, Netherlands. Email: huynen@cmbi.ru.nl

Discovering Biomolecular Mechanisms with Computational Biology,
edited by Frank Eisenhaber.

complementary distribution,[8] the sharing of regulatory elements,[9-11] and methods that use sequence information of the proteins itself.[12-14] have been proposed and implemented (Fig. 1). These methods have in common that they exploit the availability of multiple sequenced genomes to increase an in itself weak signal for functional interaction between proteins. To give one example, that two genes are neighbors in one bacterial species is only a weak signal that they functionally interact, but when that gene order is conserved among many genomes it does become a strong signal, even of physical interaction between the proteins.[15] Genomic-context methods are becoming well established and have been the subject of many reviews already.[8,16-18] Presently the focus is on combining and integrating them with each other and with other types of genomics data. As is the case of methods based on comparing genomes, also using the evolutionary conservation of coexpression,[19-21] or of physical interaction as measured with yeast-2-hybrid[19,22,23] leads to a drastic increase in the reliability of the results. Another way of combining genomics data, detecting the same interaction in different types of data, e.g., two genes tend to be coexpressed in *Saccharomyces cerevisiae* and their proteins interact in a yeast-2-hybrid experiment in the same species, increases the likelihood of that interaction.[24,25] In this book chapter we will focus on the practical applicability of these methods: can we use these tools to make predictions that are not only reliable, but that are also specific enough to design experiments to test them, and therewith complete a research "circle" from (genomics) experiment to theory to experiment?

We survey the predictions that were actually experimentally verified and make a number of new ones. The latter will illustrate that, notwithstanding all the advances in making the data and tools available on the web, making specific predictions still requires manual intervention and creativity to integrate the different types of information and make specific predictions.

Types of Genomic Context

Gene Fusion

The finding of two or more proteins encoded by separate genes of which orthologs in a different species are encoded in a single gene (Fig. 1A), reveals a gene fusion or gene fission event.[26] This is the most direct form of genomic context and, from a functional point of view, the fusion of two proteins can result in an enhancement of the interaction between their respective biochemical activities to facilitate, for example, the channelling of a substrate.[27] Using this approach to predict functional interactions in complete genomes was introduced in 1999 by Marcotte et al[2] and Enright et al.[3] In concordance with the above mentioned substrate channelling effect, most of the observed fusions events involve metabolic enzymes, although the fusions do not always involve subsequent steps in the pathway.[3,28] In *Escherichia coli* three quarters of the total of gene fusions affect metabolic genes.[29]

Conservation of Gene Order in Prokaryotes

The first pairwise genome-wide sequence comparisons revealed that even closely related species lack large scale conservation of gene order,[7,30-32] indicating that in the course of evolution genomes are rapidly rearranged and shuffled. Yet in prokaryotes some clusters of genes appear conserved in evolution (Fig. 1B), including the relative location of the genes within them, over large evolutionary distances. Further inspection of these genes revealed that they tend to encode proteins that functionally interact,[5,33] and that they tend to be part of the same operon.[34] As in the case of gene fusion, since conservation of chromosomal proximity has functional meaning it can be used to predict functional interaction between the components of conserved gene clusters. This was proposed in 1998 by Overbeek et al[33,35] and Dandekar et al,[5] by measuring conservation of genes in runs (sets of genes encoded in the same strand and separated by less than 300 bases) and conservation of neighboring genes respectively. Although there are some hints of chromosomal clustering of functionally interacting genes in eukaryotes, e.g., in polycistronic transcripts in Nematodes[36] these do not appear strong enough to predict functional interactions with any level of confidence.