CHAPTER 3

Prediction of Protein Function:
Two Basic Concepts and One Practical Recipe

Frank Eisenhaber*

Abstract

The analysis of uncharacterized biomolecular sequences obtained as a result of genetic screens, expression profile studies, etc. is a standard task in a life science research environment. The understanding of protein function is typically the main difficulty. This chapter intends to give practical advise to students and researchers that have only introductory knowledge in the field of protein sequence analysis.

Applicable theoretical approaches range from (1) textual analyses, interpretation in terms of patterns of physical properties of amino acid side chains and (2) the extrapolation of empirically established relationships between local sequence motifs with known structural and functional properties to the collection of sequence segment families with sequence distance metrics and protein function derivation with annotation transfer (concept of homologous families). Here, the impact of different techniques for the biological interpretation of targets is discussed from the practitioner's point of view and illustrated with examples from recent research reports. Although sequence similarity searching techniques are the most powerful instruments for the analysis of high-complexity regions, other techniques can supply important additional evaluations including the assessment of applicability of the sequence homology concept for the given target segment.

Introduction

The genome has become the integrating principle for the various fields of biology and the clarification of pathways that lead to the realization of genome information into phenotypes under varying environmental conditions has become the central task for life sciences. As a first step, it is critical to understand the function of genes at least in qualitative terms; i.e., to name the molecular function of encoded proteins and to uncover the topology of interactions of networks involving them. Given that, currently, the molecular function of at least two thirds of all genes in completely sequenced eukaryote genomes remains more or less clouded, this would represent a dramatic progress. At the same time, it should be noted that real theoretical predictability of biological systems above the level of educated guesses (for example, for drug engineering) typically requires quantitative characterization of gene and protein activity and modeling of biological networks, which will be, in most cases, not a matter of the coming handful of years. Possibly, this is even an optimistic assessment.

With the central role of the genome in the functioning of biological systems, it is not surprising that experimental screens for genes relevant for the processes investigated are a standard approach in today's experimental biology; for example, expression profiling with DNA

*Frank Eisenhaber—Research Institute of Molecular Pathology, Dr. Bohr-Gasse 7, A-1030 Vienna, Republic Austria. Email: Frank.Eisenhaber@imp.univie.ac.at

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

microarrays, yeast two hybrid screens, etc. If the biological phenomenon has not been well described in already published research, the screens lead typically to sequence tags of yet uncharacterized genes. Their sequence information has then to be interpreted in functional terms within the given physiological context. Stereotypically, the sequence is submitted to a similarity search in sequence databases. As a rule, the amount of insight produced by such a direct approach is indirectly proportional to the novelty of the gene target. In this tractate, we want to discuss the few fundamental principles that underlie state-of-the-art protein sequence analysis approaches. Then, we propose a general recipe for the practitioner who looks for research hints in his target sequences. We will give interpretation guides for sequence analytic findings and emphasize limitations where appropriate.

The Beginning: Deriving the Protein Sequence and the Definition of Protein Function

Typically, the starting point is a partial nucleic acid sequence representing a piece of mRNA. Whereas the experimental extension of the sequence to a full transcript was mandatory before the era of large-scale sequencing, this step can often be avoided now. In this case, it is necessary to find (1) a longer expressed sequence tag (EST), (2) a cluster of ESTs with a consensus sequence or, luckily, (3) a complete cDNA in the databases that obviously contains the reliably sequenced segment of the partial sequence obtained in the screen. The completeness of the putative transcript sequence can be investigated by mapping relevant ESTs onto the genome sequence. Especially in the case of incomplete transcripts involving only 3' untranslated regions, searching for the closest predicted gene upstream in the genome might yield the desired gene.[1,2] Searches for ESTs that bridge the distance between the detected gene and the mapped site are a possible reliability check and can also discriminate cases of alternative splicing. Further, the possibility of stumbling onto a pseudogene must be ruled out.[3,4]

Whereas all the steps leading to the protein sequences possibly encoded in the given transcript (in this essay, we do not consider untranslated RNAs) are sometimes complicated by sequencing errors (frameshifts, single point exchanges, genome fusion errors) but, in most cases, are just a technical exercise, the insufficient understanding of biological function for proteins known only as conceptual translations has become the major bottleneck in sequence data interpretation.

A few words on protein function: Protein function requires a hierarchical concept for the description of its many aspects that reflects the complexity of living systems.[5] The protein's function at the molecular level is rather a list of potential capabilities determined by its primary and tertiary structure. *Molecular function* description includes qualitative and quantitative aspects of diffusion properties in solution and membrane environments, conformational flexibility, allosteric conformational changes, possible ligand-binding (or catalytic) activities and ability for posttranslational modifications. Depending on cellular context (subcellular localization), different features of the molecular function may become important. A set of many cooperating proteins is responsible for a *cellular function* (metabolic pathway, signal transduction cascade, cytoskeletal complex, etc.). Since gene expression is regulated in a time- and tissue-dependent manner, regulatory sequences in the genomic environment of the gene considered come additionally into play at this level.[2] Finally, the presence and activity of a gene product may be directly associated with a *phenotypic function* at the organism or population level. Typically, only some aspects of molecular or cellular function are in the reach of sequence analytic studies.

Concept No. 1: Function Inheritance from a Common Ancestor Gene

The most widely known, the evolutionary (historic) approach for inferring protein function with nonexperimental means is based on the frequent observation of similarity between biomolecular sequences coding proteins with similar molecular function. Since the early examples were typically metabolic enzymes or transporters (such as hemoglobin) for which the 3D structure was available, the insight materialized soon in the paradigm of both equal/similar