Chapter VIII
Function and Homology of Proteins Similar in Sequence: Phylogenetic Profiling

Thomas Meinel
Max Planck Institute for Molecular Genetics, Germany

ABSTRACT

The function of proteins is a main subject of research in systems biology. Inference of function is now, more than ever, required by the upcoming of novel protein sequences in consequence of the discovery of new proteomes. The calculation of sequence similarity is an easily feasible way to compute protein comparisons. The comparison of complete proteomes touches one of the earliest topics in bioinformatics; the biologically meaningful organization of proteins in protein families. Several approaches that interpret function or evolutionary aspects of proteins from sequence similarity are reviewed, which in particular reflects the arsenal of techniques introduced until now. Phylogenetic profiling, a method that compares a set of genes or proteins by their presence or absence across a given set of organisms, is also presented in this chapter. Proteins in a functional context, for example, a pathway or a protein complex, are represented by identical or similar phylogenetic profiles. The detection of functional contexts by phylogenetic profiling is also playing a prospective role as an analytic tool in systems biology. Already established tools for phylogenetic profiling as well as particular biological examples based on the SYSTERS protein family data set are presented.

INTRODUCTION

Protein sequence similarity is a feature that plays a central role in comparative proteomics for the inference of protein function or analysis of protein evolution. To study the complexity of functional cellular units like proteins, basic research can often only be conducted on animal models with the completely available experimental design. These studies are expected to play a significant role in medical research.

Results are considered to be comparable with results from clinical diagnostics. It must be confirmed that results are transferable to human proteins if such experiments are not possible. Therefore, the determination of proteins with identical function in animal models and in human is essential.

Evolution of life can be characterized by the development of species as well as by the divergence of protein sequences, and it is notable that also the development of protein function is an evolutionary process. Several more or less independent biological research fields are introduced to elucidate the backgrounds of a particular evolution event - often only under constraints of the temporary availability of appropriate methods. The development of techniques for a rapid and voluminous sequencing of DNA continues to lead to complete gene sets, genomes, for an increasing list of organisms. Computational techniques calculate the translation of genomic information to proteins including processes like alternative splicing or translation start site variation. The generation of all electronically inferred proteomes is based on such specific algorithms. In parallel to the development of those tools, new evolutionary events were detected and investigated. Some of them are evolutionary events like gene duplication, gene fusion or fission, protein domain rearrangements, horizontal gene transfer, multiple copy number of genes.

In its first part, this book chapter emphasizes the reasons for distinguishing between sequence similarity and homology and function of proteins. Sequence similarity is a parameter that can be computed from a simple biophysical trait of a protein, the sequence, i.e., the primary protein structure. However, it is more complex to determine protein homology, even if it is plausible that proteins with common evolutionary history are similar in sequence. The other way around, in the context of bioinformatics, inference of homology is the interpretation of an observation - namely sequence similarity. The goal here is to determine similar proteins in recent organisms as descendants from a gene with common ancestry and thereby as homologs. The matter of inference of protein function can be discussed in the same way: Similar proteins possess with high probability a common ancestry and therefore similar function. But proteins can adopt function or are specialized during their evolutionary history. Proteins of similar function not necessarily possess similar sequence, therefore.

Consequently, it is necessary to know existing protein sequence comparison methods and underlying methods for the partitioning of proteins into protein families. In fact, a scientist works with more or less closely related members of protein families when using an expression like ‘two homologous genes’. Methodological backgrounds of established data sets are therefore briefly reviewed by this book chapter.

It is observed that proteins in a common functional context are evolutionary conserved in most of the organisms that own such a functional context. A phylogenetic profile is a pattern of presence or absence of a gene or protein across a given set of organisms. Phylogenetic profiling is a method that compares proteins by their phylogenetic profiles. Because proteins are different in organisms, a grouping of proteins is essential for the generation of a phylogenetic profile. Phylogenetic profiling is therefore depending on the method for partitioning of the protein space into protein families. This book chapter in its second part reviews the backgrounds of established phylogenetic profiling tools, restrictions to subsets of organisms on super-kingdom level, general limitations for the detection of functional contexts, and provides particular biological examples of phylogenetic profiles. Phylogenetic profiling as a method to infer unknown protein contexts or to elucidate contexts of proteins unknown in function becomes prospectively relevant in systems biology, and more so with the increasing number of complete eukaryotic proteomes.

Function of proteins is a central issue of this review, as subject of detection for unknown proteins using sequence similarity and as subject of inference of functional contexts in phylogenetic profiling.