
Preface

Knowing the function of a protein and understanding how it is carried out are the ultimate goals of molecular biology and biochemistry. From the early stage of bioinformatics in the 1980s, the development of computational tools to aid in elucidating protein function was a major focus of the field. Numerous methods have been developed since then. Computationally, protein function can be predicted through similarity searches because similarity implies homology from an evolutionary standpoint, and also because it indicates that the proteins have the same physical structures where the function takes place. Thus, based on this similarity principle, methods were developed to compare global or local sequences and the structures of proteins. Databases were also developed, which organize function information of proteins and serve as references to be queried against. In this book, well-established sequence- and structure-based tools and databases are introduced, which are very useful for biology labs. In addition, this book introduces software which addresses function beyond its conventional meaning, reflecting the diversity of the current active research field.

This book begins by introducing two sequence-based function prediction methods, PFP and ESG, in Chapter 1. The chapter also describes a web server, NaviGO, which can analyze Gene Ontology annotations. Then, Chapters 2, 3, and 4 discuss tools suitable for the functional analysis of metagenomics data. The tools in these three chapters are based on sequence database searches faster than conventional homology search methods, a necessity when processing the large amounts of sequence data which typify metagenome sequences. Chapter 2 introduces GhostX, which uses a suffix array for fast sequence comparison. Fun4Me in Chapter 3 is a pipeline that combines protein coding gene detection in query sequences and a fast sequence database search utilizing a hashing technique. SUPER-FOCUS in Chapter 4 combines fast search algorithms with preclustered reference sequence databases. In Chapter 5, we have MPFit, a program that detects when query proteins are moonlighting proteins, i.e., a protein with dual functions.

The next chapter (Chapter 6) describes SignalP, a well-established web server that predicts subcellular localization by recognizing a signal peptide in a query sequence. Subcellular localization is one of the three functional categories in the Gene Ontology (Cellular Component), and it can be a clue for other biological functions of a protein since localization and biological function are closely correlated.

The following four chapters deal with protein structures. ProFunc in Chapter 7 is a popular web server that performs multiple different analyses on a query protein structure, including global and local structure matching to known proteins. Chapter 8 describes G-LoSA, which finds ligand binding sites similar to a query binding site within a reference database. eMatchSite, the following chapter (Chapter 9), aligns two ligand binding sites to quantify similarities between them. In Chapter 10, WATsite2.0 is introduced, which predicts bound water molecules in a ligand binding site. Water molecules bound to proteins mediate ligand-protein interactions and are thus important in protein function.

The subsequent five chapters cover resources that address protein function through pathways, networks, and genomes. Chapter 11 discusses recent updates of KEGG, focusing on enzymes and pathways. KEGG is one of the most comprehensive databases of pathways, genomes, and other biomolecules and is a fundamental resource for understanding protein

function at a systems level. Chapter 12 is about the Microbial Genome Database, a valuable resource to perform comparative genomics. The *Saccharomyces* Genome Database (SGD) is described in Chapter 13. *S. cerevisiae* is one of the most extensively studied organisms. SGD has long served as a reliable source for protein function and other resources, including gene expression and phenotypes, in *S. cerevisiae*. Chapter 14 introduces MouseNet, which predicts gene function in mice from a gene expression network. FANTOM5 in Chapter 15 is a database of human and mouse genomes. Transcription start sites and promoter activities of various cells can be browsed and searched. The last chapter (Chapter 16) introduces Spatiocyte, a software for simulating the diffusion and localization of proteins in a cell. Results from the simulation, i.e., a phenotype, can be compared against microscope observations. Proteins exhibit their function through dynamic interactions in a cell environment. Thus, ultimately functions must be considered in a dynamic system, which this software aims to do.

I hope readers enjoy this book as a practical guide for using bioinformatics tools related to protein function prediction. Moreover, I also hope that this compilation itself exhibits a snapshot of the current research field and our understanding of the concept of protein function, while indicating the future direction of the field.

Editing of this book was greatly aided by Mr. Joshua McGraw, Ms. Sarah Rodenbeck, Ms. Lenna X. Peterson, and Mr. Charles Christoffer of my research group. I would like to conclude this preface by recognizing and acknowledging their help as a happy memory of my research activities.

West Lafayette, IN, USA

Daisuke Kihara

Protein Function Prediction

Methods and Protocols

Kihara, D. (Ed.)

2017, X, 239 p. 82 illus., 71 illus. in color., Hardcover

ISBN: 978-1-4939-7013-1

A product of Humana Press