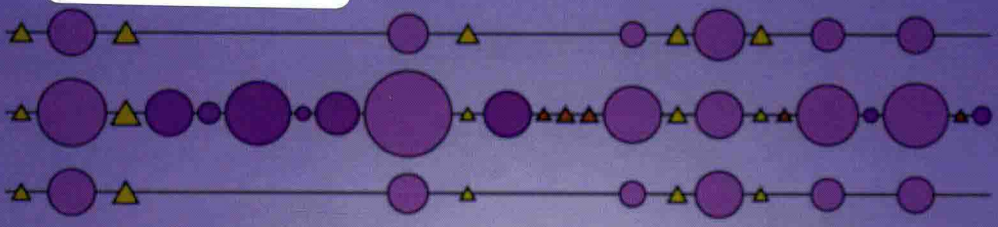
Wiley Series in Protein and Peptide Science

Vladimir N. Uversky, Series Editor



PROTEIN FAMILIES

Relating Protein Sequence, Structure, and Function

Christine Orengo AND Alex Bateman

WILEY

PROTEIN FAMILIES

Relating Protein Sequence, Structure, and Function

Edited by

CHRISTINE ORENGO
ALEX BATEMAN

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Protein families : relating protein sequence, structure, and function / edited by Christine A. Orengo, Alex Bateman.
pages cm. – (Wiley series in protein and peptide science ; 10)
Includes index.
ISBN 978-0-470-62422-7 (hardback)
1. Proteins. 2. Proteomics. 3. Molecular biology–Data processing. 4. Bioinformatics. I. Orengo, Christine A., 1955– editor of compilation. II. Bateman, Alex, 1972– editor of compilation.
QP551.P695925 2014
572′.6–dc23

2013016212

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PROTEIN FAMILIES

WILEY SERIES ON PROTEIN AND PEPTIDE SCIENCE

Vladimir N. Uversky, Series Editor

Metalloproteomics · Eugene Permyakov

Instrumental Analysis of Intrinsically Disordered Proteins: Assessing Structure and Conformation · Vladimir Uversky and Sonia Longhi

Protein Misfolding Diseases: Current and Emerging Principles and Therapies · Marina Ramirez-Alvarado, Jeffery W. Kelly, and Christopher M. Dobson

Calcium Binding Proteins · Eugene Permyakov and Robert H. Kretsinger

Protein Chaperones and Protection from Neurodegenerative Diseases · Stephan Witt

Transmembrane Dynamics of Lipids · Philippe Devaux and Andreas Herrmann

Flexible Viruses: Structural Disorder in Viral Proteins · Vladimir Uversky and Sonia Longhi

Protein and Peptide Folding, Misfolding, and Non-Folding · Reinhard Schweitzer-Stenner

Protein Oxidation and Aging · Tilman Grune, Betul Catalgol, and Tobias Jung

Protein Families: Relating Protein Sequence, Structure, and Function · Edited by Christine Orengo and Alex Bateman

INTRODUCTION

CHRISTINE ORENGO

Institute of Structural and Molecular Biology, University College London, London, United Kingdom

ALEX BATEMAN

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom

The protein machine is a triumph of nature that puts any man-made nanotechnology into the deepest shade. Without the myosin motor proteins that drive the actin filaments along the myosin tails in muscle tissue we cannot move. Without the rotating motor protein complex F0/F1 ATPase we cannot generate chemical energy in the form of ATP that is so essential for all life. Every cell in our bodies is a whirring biochemical machine of immense complexity. We are still ignorant of the exact molecular function of many, or perhaps most, of the protein cogs in this machine. To understand all the molecular components of the cell and how they fit together remains one of the greatest challenges for biology.

Charles Darwin had no idea of the molecular complexity that lay in the heart of every cell. However, his theory of evolution by natural selection has given us a framework that allows us to understand how the complexity of the cell and its protein machinery could have arisen from simpler preexisting proteins. By looking at the amino acid sequence of different proteins we can see that nature's major source of innovation is the duplication and subsequent mutation of proteins. The five human hemoglobin genes that share a common function to transport oxygen around the blood have all arisen from a single ancestral gene during the evolution of animals over the last 800 million years. Each of these hemoglobin genes has small differences in sequence and this causes differences in their affinity for oxygen and other properties. The set of proteins that have arisen from a common ancestor through the process of evolution are known as a protein family.

The concept of a protein family as an evolutionary entity has immense implications for understanding biology. Related proteins arising from a common ancestral protein often share a common function. If we can identify a protein in a newly sequenced organism that belongs to the hemoglobin family, then we can infer that its function is likely to be to transport oxygen. Despite having carried out no experiments on this new protein, we can learn something about its function from its amino acid sequence. By carrying out detailed molecular experiments on proteins from a few model organisms, we might hope to understand all proteins in the millions of species on earth.

Our ability to correctly identify proteins that belong to the same family is essential to understanding biology. Our ability to do this has improved immensely over the past 40 years. These improvements have been due to three different factors: (i) improvements in the algorithms and statistics associated with sequence alignment, (ii) the growth in the number of protein sequences, and (iii) the increase in the availability of protein structures.

1 IMPROVEMENTS IN ALGORITHMS FOR SEQUENCE ALIGNMENT

Our ability to see relationships between proteins has been greatly enhanced not just by the wealth of sequence and structures available to us. The sophisticated algorithms and statistics that have been developed allow us to determine which similarities between protein sequence and structures are of true homology and which reflect only chance similarities. While sequence comparison software such as BLAST and Fasta made comparison of sequences accessible, techniques such as profiles, hidden Markov models, and fold recognition gave experts the ability to find relationships between proteins whose common ancestor may have existed more than a billion years ago. Although algorithmic developments that have been extensively covered elsewhere are not the primary focus of this book, we applaud the computational scientists and mathematicians who have given us the tools to unlock the mysteries of the cell's protein machine.

2 THE GROWTH OF PROTEIN SEQUENCES

International genome projects have brought a wealth of diverse protein sequences and this means that in the last 10 years or so there have been significant increases in the number of protein and nucleic acid sequences available. Protein sequence databases now hold more than 20 million sequences. This also gives rise to a large increase in the number of known protein families. For example, automatic classification of protein families suggests that we now have representatives from more than a million families. Protein family classifications such as PhyloFacts or PANTHER (described by Sjolander in Chapter 6), which focus on specific sequence repositories and involve some limited curation, now contain around 93,000 and 71,000 families, respectively.

However, many proteins (nearly 80% in eukaryotes) are multidomain and the million or more protein families currently identified are built up from different combinations of domains. In this sense, domains are the primary building blocks of life and not surprisingly there are far fewer domain families than protein families. Furthermore, there has been a much slower increase in the numbers of domain families—especially over the last 5 years. The most comprehensive domain family resource, Pfam (reviewed by Bateman in Chapter 3) currently identifies nearly 14,000 families. Moreover, many new Pfam families tend to be quite small and species specific, suggesting that we may be close to knowing a significant proportion of the major domain families in nature. With the growth of next generation sequencing, it is likely that we will soon see improved sampling of unusual taxonomic groups and in the next 20 years we are likely to have access to a true sampling of protein space.

Alongside the activities of the international genome sequencing initiatives, worldwide structure genomics consortia have attempted to increase the structural coverage of domain and protein families. Since the structure of a protein is usually much more highly conserved during evolution than the sequence, this data is valuable for detecting remote homologies and has been exploited by resources such as SCOP and CATH to trace far back in evolution and capture universal families common to all kingdoms of life. There appear to be only a few hundred of these, depending on the criteria used to identify them, and some have been extensively duplicated and are highly populated.

By exploiting structural data we see that there are currently less than 3000 domain superfamilies covering nearly 60% of the domain sequences from completed genomes. The term "*superfamily*" denotes a broad grouping of relatives (i.e., including all paralogs and orthologs) even from very divergent species, and remote relatives can have rather different structures and functions within some superfamilies (see, e.g., the HUP superfamily described in Chapter 8). Structural data can also be used to merge domain "families" identified using purely sequence data—for example, Pfam often recognizes "clans" (comprising remotely related Pfam families) in this manner.

The relatively small number of domain superfamilies relative to protein families and the fact that we have nearly classified a complete set of these domain "building blocks" mean that we can begin to understand the assembly of diverse proteins during evolution from different domain combinations and start to derive rules for predicting the likely functional contributions of the domains or how their roles may change in different contexts. This will hopefully allow us to move toward a domain grammar of function that exploits our understanding of the evolutionary changes occurring in different domain families to build a picture of how the complete protein, containing these domains, may function.

The data from some of the structural genomics initiatives adds further support to the hypothesis that we already know a large proportion of all major domain families. For example, the NIH-funded PSI structural genomics initiatives in the States deliberately sought to identify new domain families for which there was no structural data. In their second phase (PSI2: 2005–2010) they primarily focused

on new, structurally uncharacterized families in Pfam and related classifications. Powerful HMM–HMM strategies were employed to discard any that were, in fact, distantly related to known families (e.g., in SCOP or CATH) and those remaining were targeted for structure determination. However, despite their lack of sequence similarity to known families, it became increasingly clear as the structures were solved that most of the families were simply divergent relatives of existing families in SCOP or CATH. Only about 20% of them represented completely novel families with novel structures, and many of these novel families were very small, species or subkingdom specific, with less than 100 relatives.

As reported in Chapter 5, some resources (SUPERFAMILY, Gene3D) derive sequence patterns (or HMMs) for domain superfamilies in SCOP and CATH and use these to predict domain relatives in sequences from completed genomes. Their data suggests that the population of superfamilies is very uneven. The trends follow scale-free behavior whereby most superfamilies are rather small, that is, comprising less than 500 relatives while a few ($\sim$200) are very large (having >5,000 relatives). This tiny percentage of superfamilies (<5% of all superfamilies) accounts for nearly two thirds of all structural domains classified.

Many are universal and highly promiscuous, combining with multiple other families to give different multidomain combinations. They support a wide range of functions, either by performing a generic role in different protein contexts or by evolving new functions of the domain itself, that is, through residue mutations and structural divergence. For example, changes in the nature and location of catalytic residues in the active site have been observed. Structural variations can alter the active site geometry to enable binding of different substrates and/or reshape surface features promoting changes in domain or protein interaction partners.

As the sequence and structure data grows—and especially as structural genomics initiatives target new families—the mechanisms by which domains change during evolution will become clearer as also the extent to which they fuse with different partners to give new proteins. However, the coverage of current classifications and the insights already derived from them motivated us to compile this book now, both to convey some of the current knowledge and to present some fascinating examples of the role families play in creating the rich diversity of life we see around us and study as biologists.

3 MOTIVATION FOR THE BOOK

The idea that we may now have accumulated knowledge on all the major protein domain families is borne out by the fact that a large proportion (between 70% and 90%) of domain sequences from most completed genomes can be classified in curated domain families in Pfam. In addition, the technologies for recognizing distant relatives of existing families and confidently assigning new families have matured over the last decade with powerful strategies such as profile–profile comparisons identifying incredibly distant and divergent relatives, some of which may have undergone significant structural changes as well.

Protein and domain family classifications are becoming increasingly and routinely used to annotate newly sequenced proteins, for example, from meta-genome studies or completely sequenced genomes. So a review of protein families—how to identify them and what the analyses of these families tells us about the evolution of the proteins and their impact on the phenotypic repertoire of the organisms they are found in—seemed both timely and valuable for biologists wishing to use these resources to infer functions for their proteins of interest.

There are now many protein, domain, and motif classification resources, some very comprehensive (e.g., Pfam or SCOP) and others only focusing on specific families (e.g., related to a disease or a particular functional activity) or biological processes (e.g., kinases). In order to give a flavor of the technologies used for finding families and the insights they bring, we decided to divide the book into three sections. The first covers strategies for identifying and characterizing the families. Since we felt that it would be unrealistic to capture in a single book the different technologies and data exploited and presented by all family classifications, we invited contributions from authors of the larger scale, more comprehensive resources who could provide overviews of the challenges and strategies related to their own types of classification. We decided to organize the book into three sections. The first section titled "Concepts Underlying Protein Family Classification" of this book reviews the major strategies for identifying homologous proteins and classifying them into families. In the second section titled "In-Depth Reviews of Protein Families" of this book, there is a collection of reviews on some fascinating superfamilies for which we have substantial amounts of data (sequences, structures, and functions) allowing us to trace the emergence of functionally diverse relatives and providing structural insights into the mechanisms modifying their functions. Chapters in the third section titled "Review of Protein Families in Important Biological Systems" review groups of families associated with a particular biological theme (e.g., the protein families involved in the cytoskeleton, reviewed by Baines and coauthors).

We would like to thank all of the authors who contributed to this book. We have been delighted that so many experts from the world over were able to devote their time to create this collection of knowledge. We believe that this work will be useful for student and group leaders alike and hope that you enjoy reading the book as much as we have.

CONTRIBUTORS

Saraswathi Abhiman, National Institutes of Health, National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA

Vivek Anantharaman, National Institutes of Health, National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA

L. Aravind, National Institutes of Health, National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA

Patricia C. Babbitt, Department of Biopharmaceutical Sciences, UCSF Mission Bay, San Francisco, CA, USA

Anthony J. Baines, School of Biosciences, University of Kent, Canterbury, UK

Alan E. Barber II, Department of Biopharmaceutical Sciences, UCSF Mission Bay, San Francisco, CA, USA

Alex Bateman, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK

Rostislav Castillo, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

Varodom Charoensawan, Department of Biochemistry, Mahidol University, Bangkok, Thailand, Integrative Computational BioScience (ICBS) Center, Mahidol University, Bangkok, Thailand

Jonathan S. Chen, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

Erik L. Clarke, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

Alison Cuff, Institute of Structural and Molecular Biology, University College London, London, UK

Benoit H. Dessailly, National Institute of Biomedical Innovation, Osaka, Japan

Nicholas Furnham, European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

Julian Gough, Department of Computer Science, University of Bristol, Bristol, UK

Daniel H. Haft, J Craig Venter Institute, Rockville, MD, USA

Andreas Heger, Department of Physiology, Anatomy and Genetics, MRC CGAT/Functional Genomics Unit, University of Oxford, Oxford, OX, UK

Michael A. Hicks, Department of Biopharmaceutical Sciences, UCSF Mission Bay, San Francisco, CA, USA

Gemma L. Holliday, European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

Liisa Holm, Department of Biological and Environmental Sciences, Institute of Biotechnology, University of Helsinki, Helsinki, Finland

Lakshminarayan M. Iyer, National Institutes of Health National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA

Eugene V. Koonin, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Ujjwal Kumar, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

Juliette T.J. Lecomte, Department of Biophysics, Johns Hopkins University, Baltimore, MD, USA

Arthur M. Lesk, Department of Biochemistry and Molecular Biology, Huck Institute for Genomics, Proteomics and Bioinformatics, The Pennsylvania State University, University Park, PA, USA

Kira S. Makarova, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Ankur Malhotra, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

Russell de la Mare, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

Alexey Murzin, MRC Laboratory of Molecular Biology, Cambridge, UK

Christine Orengo, Institute of Structural and Molecular Biology, University College London, London, UK

Neil D. Rawlings, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, United Kingdom

Vamsee S. Reddy, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

Milton H. Saier, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

Maksim A. Shlykov, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

Kimmen Sjölander, Plant & Microbial Biology, Bioengineering, Berkeley, CA, USA

Eric I. Sun, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

Sarah Teichmann, MRC Laboratory of Molecular Biology, Cambridge, UK

Janet M. Thornton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

Steven T. Wakabayashi, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

Corin Yeats, Institute of Structural and Molecular Biology, University College London, London, UK

CONTENTS

SECTION I

CONCEPTS UNDERLYING PROTEIN FAMILY CLASSIFICATION