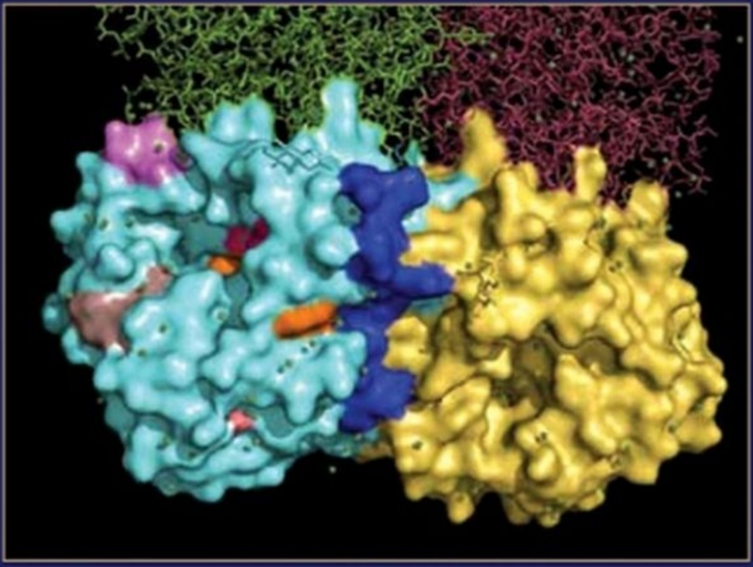
Advances in PROTEIN CHEMISTRY and STRUCTURAL BIOLOGY

VOLUME 83

Protein Structure and Diseases



Edited by
Rossen Donev

ADVANCES IN PROTEIN CHEMISTRY AND STRUCTURAL BIOLOGY

Volume 83

Protein Structure and Diseases

ADVANCES IN PROTEIN CHEMISTRY AND STRUCTURAL BIOLOGY

Protein Structure and Diseases

EDITED BY

ROSSEN DONEV
Institute of Life Science
College of Medicine, Swansea University
Swansea
United Kingdom

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA

First edition 2011

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

ISBN: 978-0-12-381262-9
ISSN: 1876-1623

For information on all Academic Press publications
visit our website at www.elsevierdirect.com

Printed and bound in USA
11 12 13 14 10 9 8 7 6 5 4 3 2 1

Contents

Graphical Representation and Mathematical Characterization of Protein Sequences and Applications to Viral Proteins

Ambarnil Ghosh and Ashesh Nandy

Structural, Thermodynamic, and Mechanistical Studies in Uroporphyrinogen III Synthase: Molecular Basis of Congenital Erythropoietic Porphyria

Arola Fortian, David Castaño, Esperanza Gonzalez, Ana Laín, Juan M. Falcon-Perez, and Oscar Millet

Role of Fibrin Structure in Thrombosis and Vascular Disease

Amy L. Cilia La Corte, Helen Philippou, and Robert A. S. Ariëns

Structural, Dynamic, and Functional Aspects of Helix Association in Membranes: A Computational View

Anton A. Polyansky, Pavel E. Volynsky, and Roman G. Efremov

Proteins Move! Protein Dynamics and Long-Range Allostery in Cell Signaling

Zimei Bu and David J. E. Callaway

Structural Diversity of Class I MHC-like Molecules and its Implications in Binding Specificities

MD. Imtaiyaz Hassan and Faizan Ahmad

This page intentionally left blank

GRAPHICAL REPRESENTATION AND MATHEMATICAL CHARACTERIZATION OF PROTEIN SEQUENCES AND APPLICATIONS TO VIRAL PROTEINS

By AMBARNIL GHOSH* AND ASHESH NANDY†

*Physics Department, Jadavpur University, Jadavpur, Kolkata, India
†Centre for Interdisciplinary Research and Education, Jodhpur Park, Kolkata, India

Abstract

Graphical representation and numerical characterization (GRANCH) of nucleotide and protein sequences is a new field that is showing a lot of promise in analysis of such sequences. While formulation and applications of GRANCH techniques for DNA/RNA sequences started just over a decade ago, analyses of protein sequences by these techniques are of more recent origin. The emphasis is still on developing the underlying technique, but significant results have been achieved in using these methods for protein phylogeny, mass spectral data of proteins and protein serum profiles in parasites, toxicoproteomics, determination of different indices for use in QSAR studies, among others. We briefly mention these in this chapter, with some details on protein phylogeny and viral diseases. In particular, we cover a systematic method developed in GRANCH to determine conserved surface exposed peptide segments in selected viral proteins that can be used for drug and vaccine targeting. The new

DOI: 10.1016/B978-0-12-381262-9.00001-X

GRANCH techniques and applications for DNAs and proteins are covered briefly to provide an overview to this nascent field.

I. Introduction

A. *Protein Basics*

Proteins, the most versatile macromolecules in the living system, primarily constitute complex folded chain of amino acids which are encoded by genes. The information content of the folded complex constitutes a functional unit that plays a crucial role in biological processes. The origin of the word ''Protein'' is from the Greek ''prota'' which means ''of primary importance.'' This name was coined by Jöns Jakob Berzelius in 1838 for large organic compounds with a very close similarity in their empirical formulae and of primary importance in animal nutrition, though the evidences were not so prominent at that time. A landmark in protein chemistry came through Frederick Sanger and his colleagues, at the University of Cambridge in 1954 when, after 10 years of hard work, they succeeded in solving the complete primary structure of insulin (Sanger and Tuppy, 1951; Sanger, 1952; Sanger and Thompson, 1953). The very next milestone in protein chemistry was Max Perutz (Perutz and Weisz, 1947; Perutz, 1960; Perutz et al., 1960) and Sir John Cowdery Kendrew (Kendrew et al., 1958, 1960; Kendrew, 1959) solving the 3D structure of hemoglobin and myoglobin. These findings are the basis of modern age of advanced structural protein chemistry research.

Proteins form the building blocks of the structure and function of biological entities. A typical mammalian cell contains as many as 10,000 different proteins having a diverse array of functions (Karp, 2008). The set of proteins expressed in a cell or cell type is called a proteome. Proteins are generally a few hundred amino acids in chain length but can vary in size from a few tens of amino acids to over 34,000 amino acids, for example, the human titins, also known as the largest in protein world (MW = 3–3.7 MDa; Opitz et al., 2003). While a single protein chain can theoretically fold in an unlimited number of ways (Chou and Fasman, 1974b; Fasman, 1989; Feldman and Hogue, 2002), typically a specific amino acid chain folds to a particular structure through a process that is not yet clearly understood (Dill et al., 2007, 2008; Ghosh et al., 2007), but which is the basis for all protein interactions; recent research shows that the folded structure might have conformational changes depending on the environment too (Makowski et al., 2008). Protein structure is often referred to in terms of four aspects: The primary structure consisting of the amino acid chain, the secondary structure which contains regularly repeating structures like alpha helices and beta sheets stabilized by hydrogen bonds, the tertiary structure which is the final folded structure incorporating the various secondary structures, and a quaternary structure where several proteins are bound together to form one protein complex such as are found in the neuraminidase body of an influenza virion (Russell et al., 2006) or the VP7 of a rotavirus particle (Li et al., 2009b). The tertiary and quaternary structures of a large number of proteins have become available through X-ray crystallography and NMR spectroscopy studies and the data are available in Protein Data Bases (PDB) such as World Wide Protein Data Bank (WWPDB; Berman et al., 2007), RCSB Protein Data Bank (RCSB-PDB; Deshpande et al., 2005; Dutta et al., 2007), Protein Data Bank Europe (PDBe; Velankar et al., 2010), Protein Data Bank Japan (PDBj; Nakamura et al., 2002; Kinjo et al., 2010), and Biological Magnetic Resonance Databank (BMRB; Markley et al., 2008). The difficulty of crystallizing proteins has restricted the number of proteins whose structures are sufficiently well known (Chayen, 2004, 2009; Chayen and Saridakis, 2008). However, taking the protein primary structure as the source material for all subsequent structures, structural genomics and protein structure prediction methods theoretically predict protein secondary and tertiary structures based on known structures (Baker and Sali, 2001).

The importance of proteins in biological function have led to wide ranging studies to understand how proteins fold (Dobson, 2004; Dill et al., 2007; Ghosh et al., 2007), interact with other proteins to regulate enzyme activity (Frieden, 1971), oligomerize to form fibrils (Powers and Powers, 2008), aggregate to protein complexes that lead to conformational changes, and enable signaling networks. These interactions are mediated by the chief characteristic of a protein: the ability to bind other molecules specifically and tightly to it. The specificity arises from unique shapes in the tertiary structure of the protein surface (Roach et al., 2005, 2006) where, for example, a depression acts as a binding site or pocket and by the chemical natures of the side chains of the neighboring amino acids. This also results in total inability to bind in cases where changes in the amino acid composition render conformational changes to the binding site (Moscona, 2005). Such changes arising out of mutations in the amino acid chains are among the main factors responsible for development of drug resistance in bacterial and viral diseases (Moscona, 2004). Enzymatic role of proteins helps catalyze metabolic reactions but only a small region of the protein consisting of a few amino acids are active in the catalysis; a noncatalytic example of protein includes the antibodies that are part of the adaptive immune systems and act as a binder to antigens for destruction (MacCallum et al., 1996). Ligand-binding proteins such as hemoglobin bind specific small molecules to transport them to other locations in the body of a multicellular organism (Baldwin and Chothia, 1979). Structural proteins such as actin and tubulin confer stiffness and rigidity to the cytoskeleton (Doherty and McMahon, 2008); other structural proteins such as myosin and kinesin generate mechanical forces and are responsible for the motility of many single cell organisms (Rayment, 1996).

Thus, there are numerous processes, and there are numerous proteins that take part in them. These processes and the functions of the proteins are studied through *in vivo* and *in vitro* analysis. *In vitro* analysis helps understand how a protein functions, *in vivo* analysis often helps in understanding its functional location and related parameters in the living system; however, the specifics of how a protein targets particular organelles or cellular structures are often unclear (Bejarano and Gonzalez, 1999). Site-directed mutagenesis techniques (Ruvkun and Ausubel, 1981) that alter the protein sequence and hence its structure and cellular location/function that help to identify susceptibility to regulation provide guidelines to rational drug design or development of new proteins with novel properties.

Among the simplest of biological entities, and of particular interest for this chapter, is the virus. A virus particle like the influenza or rotavirus contains about 8–11 protein-coding genes in a multiprotein coat that protects the RNA or DNA of the virus and also enables the proteins and genetic materials to enter and leave cells. A great range of variability in amino acid composition is observed for these viral proteins (Reid et al., 2000; Ghosh et al., 2009), specifically the surface situated ones like NA (neuraminidase; Ghosh et al., 2010), HA (hemagglutinin), VP4 (variable protein), VP7 (Gunn et al., 1985), and gp120 (of the HIV) but the functional impact remains the same. Often, a single change in the side chain of a single amino acid is enough for producing a new mutant (Lopez et al., 2005). Viruses use this highly mutable property for escaping the host defense mechanism and they are also frequently found to generate escape mutants against a naturally occurring immunity or artificially designed drug or vaccine (Air and Laver, 1989).

B. Drugs and Proteins

Proteins are involved by function or malfunction, in diseases of organism. Bacterial, viral, and other pathogens disrupt the normal protein functions and thereby destabilize the infected host organism (Goldsby et al., 2000). While immunological defenses are called into action by the infection, often these are inadequate by themselves and have to be supported by drugs, vaccines, and other therapeutical regimes. Design of drugs and study of their actions have therefore been an important area of research. Drugs can act through formation of drug–DNA complexes (Chaires, 1997, 1998) or protein–drug complexes (Chicault et al., 1981). Major trends of research into drug–DNA relationships have been recently reviewed (Nandy and Basak, 2010). Stated simply, DNA drugs and vaccines are made of plasmids designed to carry a selected gene into cells where it is translated into a protein. In the case of antiviral DNA vaccine, for example, plasmids are created for producing the selected viral protein in the cell and immune systems are expected to act to prevent future infections from the virus (Ulmer et al., 1996a,b; Gurunathan et al., 2000). Advanced techniques such as codon optimization (Deml et al., 2001) are enhancing the protein production from the plasmids and others such as adjuvant incorporation are enhancing the immune response leading to more effective vaccines and therapies, several of which are already available for treatment of specified animals afflicted with the West Nile virus (Kramer et al., 2007), melanoma and fetal loss, while applications for humans for treating HIV, influenza, hepatitis C, and other diseases are under trial (Morrow and Weiner, 2010).

Pharmaceutical proteins effective against a wide range of bacterial infections can be traced to penicillin, and developed into new class of drugs referred to as antibiotics. Conventional production processes for antibiotics are expensive and face many regulatory issues. Vaccines that enhance the body's immune system consist of attenuated viruses but can, in rare cases, harm the host with a full-blown viral occupancy (Ball et al., 1998; Colgrove and Bayer, 2005). Since viruses use the host's cells to replicate, designing safe and effective antiviral drugs is difficult and also makes it difficult to find targets for the drugs that would interfere with the virus without also harming the host organism's cells. But almost all antimicrobials, including antivirals, are subject to drug resistance as the pathogens mutate over time (Gold and Moellering, 1996), becoming less susceptible to the treatment. Small molecules are often used as drugs, but the new technology of recombinant proteins (Geigert, 1989; Dingermann, 2008), commonly produced using bacteria or yeast in a bioreactor, potentially provide greater efficacy and fewer side effects because their action can be more precisely targeted toward the cause of a disease rather than treatment of symptoms, is yet to gain wide acceptance.

Peptide-based drugs operate by stimulating the immune response to the peptide and thereby to the invading pathogen. Peptides play an important role in modulating many physiological processes in our body. Use of peptides as drugs have the benefit that they are small, easily optimized, and can be quickly investigated for therapeutic potential. However, peptide drug screening process (Otvos, 2008), although a well-established approach, is long and arduous resulting in high manufacturing costs, and the fact that they have short half-life, and limited *in vivo* bioavailability hampers their effectiveness; new approaches have been proposed to overcome the difficulty of generating sufficient amount of the required tRNAs (Owens, 2004). The peptides can be naturally derived or chemically synthesized, with the latter method being more prevalent. Novel peptide analogs (Lee et al., 2002) are also being synthesized to create more potent drugs.

In practice, protein and peptide drugs are finding increasing acceptance in therapeutics. A drug's efficiency is related to the degree of its binding with the proteins in blood serum (Meyer and Guttman, 1968; Koch-Weser and Sellers, 1976): The less bound a drug is, the more efficiently it can diffuse through cell membrane. Common drug-binding proteins in plasma are human serum albumin, lipoprotein, glycoprotein, etc. It is the unbound fraction of the drug–protein complex that exhibits therapeutic effect and excessive binding may mitigate against rapid action of the drug. However, the same effect can be used for long-lasting dosage by designing drugs that bind to the protein and act as a reservoir so that the unbound fraction is released slowly.

But degradation of the proteins during storage and drug administration routes remains a challenging problem (Frokjaer and Otzen, 2005). These issues of stability of therapeutic proteins toward aggregation and misfolding in long-term storage as well as means of efficacious delivery that avoid adverse immunogenic side effects are engaging the attention of the pharmaceutical industry (Frokjaer and Otzen, 2005). While invasive routes such as subcutaneous injections are often used, oral delivery faces difficulties in poor permeability across biological membranes due to the hydrophobic nature and large molecular size, susceptibility to enzymatic attack, among others. Formulation strategies for protein therapeutics thus continue to remain a challenging problem.

C. *Bioinformatics in Protein Studies*

The complexities of protein function and structure have necessitated the development of computational techniques to analyze available data and help in formulating novel ways to predict structure, function, and interaction of proteins. Especially, in view of the requirements of new approaches to drug development through recombinant proteins, synthesizing new peptides, and investigating drugs–DNA complexes, use of computational methods is now of vital importance.

The increased availability and accessibility of genomic and protein sequence data have opened up new possibilities for the search for target proteins, and the success of protein and peptide therapeutics is revolutionizing the biotech and pharmaceutical market, spurring the creation of next-generation products with reduced immunogenicity (Schellekens, 2002; Tangri et al., 2005), improved safety, and greater effectiveness. The protein engineering market is expected to cross $100 billion in sales in 2010 from about $36 billion 4 years ago. The top-selling therapeutic protein is reported to be Amgen's Aranesp (Locatelli and Vecchio, 2001), a reengineered variant of the company's first-generation product Epogen (recombinant human erythropoietin). A number of such products have been launched by Genetech and others, and nonparenteral delivery systems, alongside parenteral protein and peptide drug delivery systems have also been approved (Packhaeuser et al., 2004). Progress in bioinformatics and computational biology as well as new techniques in protein engineering (recombinant proteins through site-directed mutagenesis and posttranslational modifications) are aiding the development of reengineered, improved, whole antibody, and antibody fragment-based products, reducing immunogenicity by using fully human recombinant antibodies or human antibodies derived from transgenic mice and allowing biosimilar products to be differentiated on the basis of superior characteristics. Screening experiments for appropriate molecules rely critically on bioinformatics support for design of experiments and for interpreting the generated data, for example, to identify interesting differentially expressed genes and to predict the function and structure of putative target proteins (Lengauer and Zimmer, 2000).

Protein characterization and *in silico* protein design and structure analyses form an integral part of these developments. Phylogenetic analyses based on primary sequences have been used to group related proteins and understand their evolutionary history, algorithms have been developed to predict protein secondary structures, and web accessible systems are available to suggest possible folding patterns (Shen and Chou, 2009). A number of epitope prediction tools have been devised with varying degrees of success to aid in drug design (Yang and Yu, 2009); one area of nascent research is concerned with understanding of allosteric conformations that may help or hinder protein interactions (Teague, 2003). In a broader area, computational biology has already proved itself as one of the powerful tools for handling the large genomic databases. The basic applications involve killer tools like sequence alignment, phylogenetic tree drawing, sequence comparison, etc. *In silico* motif search algorithms on primary protein structure can be applied for finding structural information like signal sequence prediction (Menne et al., 2000), cleavage site prediction (Chou, 2001), glycosylation prediction (Blom et al., 2004), posttranslational modifications prediction, etc. Large datasets are frequently found to be utilized in predictions of protein structural levels from primary structure. Software like Modeller (Eswar et al., 2007, 2008), Discovery Studio, etc., can predict 3D structure of proteins from a database of known crystallized proteins. Many theories have been developed in this prediction research but they are often ineffective in case of a completely new protein for comparison with the preexisting database (comparative protein modeling) or a protein without appropriate template. Another very important application of data mining is the use of computational power in handling the proteomics data. In proteomics, proteins are detected by matching a part of it with the whole existing protein database in mass spectrophotometer software (Perkins et al., 1999). The basis of all these data mining and related computational techniques is mathematics and statistics. Different theories like dot-matrix algorithm (Gibbs and McIntyre, 1970), Needleman Wunsch algorithm (Needleman and Wunsch, 1970), Smith Waterman algorithm (Smith and Waterman, 1981; Smith et al., 1985), Hidden Marakov model (Eddy, 1996), Chou-Fasman algorithm (Chou and Fasman, 1974a,b), etc., are widely used.