Computational Biology

Dariusz Mrozek

Scalable Big Data Analytics for Protein Bioinformatics

Efficient Computational Solutions for Protein Structures

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Springer

Computational Biology

Volume 28

The *Computational Biology* series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

More information about this series at http://www.springer.com/series/5769

Dariusz Mrozek

Scalable Big Data Analytics for Protein Bioinformatics

Efficient Computational Solutions for Protein Structures

Springer

Dariusz Mrozek
Silesian University of Technology
Gliwice, Poland

ISSN 1568-2684
Computational Biology
ISBN 978-3-319-98838-2 ISBN 978-3-319-98839-9 (eBook)
https://doi.org/10.1007/978-3-319-98839-9

Library of Congress Control Number: 2018950968

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

For my always smiling and beloved wife Bożena, and my lively and infinitely active sons Paweł and Henryk, with all my love.

To my parents, thank you for your support, concern and faith in me.

Foreword

High-performance computing most generally refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering, or business. Big Data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

This timely book by Dariusz Mrozek gives you a quick introduction to the area of proteins and their structures, protein structure similarity searching carried out at main representation levels, and various techniques that can be used to accelerate similarity searches using high-performance Cloud computing and Big Data concepts. It presents introductory concepts of formal model of 3D protein structures for functional genomics, comparative bioinformatics, and molecular modeling and the use of multi-threading for the efficient approximate searching on protein secondary structures. In addition, there is a material on finding 3D protein structure similarities accelerated with high-performance computing techniques.

The book is required reading to help in understanding for anyone working with area of data analytics for structural bioinformatics and the use of high-performance computing. It explores area of proteins and their structures in depth and provides practical approaches to many problems that may be encountered. It is especially useful to applications developers, scientists, students, and teachers.

I have enjoyed and learned from this book and feel confident that you will as well.

Knoxville, USA
June 2018

Jack Dongarra
University of Tennessee

Preface

International efforts focused on understanding living organisms at various levels of molecular organization, including genomic, proteomic, metabolomic, and cell signaling levels, lead to huge proliferation of biological data collected in dedicated, and frequently, public repositories. The amount of data deposited in these repositories increases every year, and cumulated volume has grown to sizes that are difficult to handle with traditional analysis tools. This growth of biological data is stimulated by various international projects, such as 1000 Genomes. The project aims at sequencing genomes of at least one thousand anonymous participants from a number of different ethnic groups in order to establish a detailed catalog of human genetic variations. As a result, it generates terabytes of genetic data. Apart from international initiatives and projects, like the 1000 Genomes, the proliferation of biological data is further accelerated by newly developed technologies for DNA sequencing, like next-generation sequencing (NGS) methods. These methods are getting faster and less expensive every year. They produce huge amounts of genetic data that require fast analysis in various phases of molecular profiling, medical diagnostics, and treatment of patients that suffer from serious diseases.

Indeed, for the last three decades we have been witnesses of the continuous exponential growth of biological data in repositories, such as GenBank, Sequence Read Archive (SRA), RefSeq, Protein Data Bank, UniProt/SwissProt. The specificity of the data has inspired the scientific community to develop many algorithms that can be used to analyze the data and draw useful conclusions. A huge volume of the biological data caused that many of the existing algorithms became inefficient due to their computational complexity. Fortunately, the rapid development of computer science in the last decade has brought many technological innovations that can be also used in the field of bioinformatics and life sciences. The algorithms demonstrating a significant utility value, which have recently been perceived as too time-consuming, can now be efficiently used by applying the latest technological achievements, like Hadoop and Spark for analyzing Big Data sets, multi-threading, graphics processing units (GPUs), or cloud computing.

Scope of the Book

The book focuses on proteins and their structures. It presents various scalable solutions for protein structure similarity searching carried out at main representation levels and for prediction of 3D structures of proteins. It specifically focuses on various techniques that can be used to accelerate similarity searches and protein structure modeling processes. *But, why proteins?* somebody can ask. I could answer the question by following Arthur M. Lesk in his book entitled *Introduction to Protein Science. Architecture, Function, and Genomics*. Because proteins are where the action is. Understanding proteins, their structures, functions, mutual interactions, activity in cellular reactions, interactions with drugs, and expression in body cells is a key to efficient medical diagnosis, drug production, and treatment of patients. I have been fascinated with proteins and their structures for fifteen years. I have fallen in love with the beauty of protein structures at first sight inspired by the research conducted by R.I.P. Lech Znamirowski from the Silesian University of Technology, Gliwice, Poland. I decided to continue his research on proteins and development of new efficient tools for their analysis and exploration.

I believe this book will be interesting for scientists, researchers, and software developers working in the field of structural bioinformatics and biomedical databases. I hope that readers of the book will find it interesting and helpful in their everyday work.

Chapter Overview

The content of the book is divided into four parts. The first part provides background information on proteins and their representation levels, including a formal model of a 3D protein structure used in computational processes, and a brief overview of technologies used in the solutions presented in this book.

- **Chapter 1: Formal Model of 3D Protein Structures for Functional Genomics, Comparative Bioinformatics, and Molecular Modeling**
  This chapter shows how proteins can be represented in computational processes performed in scientific fields, such as functional genomics, comparative bioinformatics, and molecular modeling. The chapter provides a general definition of protein spatial structure that is then referenced to four representation levels of protein structure: primary, secondary, tertiary, and quaternary structures.
- **Chapter 2: Technological Roadmap**
  This chapter provides a technological roadmap for solutions presented in this book. It covers a brief introduction to the concept of Cloud computing, cloud service, and deployment models. It also defines the Big Data challenge and

presents the benefits of using multi-threading in scientific computations. It then explains graphics processing units (GPUs) and CUDA architecture. Finally, it focuses on relational databases and the SQL language used for declarative querying.

The second part of the book is focused on Cloud services that are utilized in the development of scalable and reliable cloud applications for 3D protein structure similarity searching and protein structure prediction.

- **Chapter 3: Azure Cloud Services**
  Microsoft Azure Cloud Services support development of scalable and reliable cloud applications that can be used to scientific computing. This chapter provides a brief introduction to Microsoft Azure cloud platform and its services. It focuses on Azure Cloud Services that allow building a cloud-based application with the use of Web roles and Worker roles. Finally, it shows a sample application that can be quickly developed on the basis of these two types of roles and the role of queues in passing messages between components of the built system.
- **Chapter 4: Scaling 3D Protein Structure Similarity Searching with Cloud Services**
  In this chapter, you will see how the Cloud computing architecture and Azure Cloud Services can be utilized to scale out and scale up protein similarity searches by utilizing the system, called *Cloud4PSi*, that was developed for the Microsoft Azure public cloud. The chapter presents the architecture of the system, its components, communication flow, and advantages of using a queue-based model over the direct communication between computing units. It also shows results of various experiments confirming that the similarity searching can be successfully scaled on cloud platforms by using computation units of different sizes and by adding more computation units.
- **Chapter 5: Cloud Services for Efficient *Ab Initio* Predictions of 3D Protein Structures**
  In this chapter, you will see how Cloud Services may help to solve problems of protein structure prediction by scaling the computations in a role-based and queue-based *Cloud4PSP* system, deployed in the Microsoft Azure cloud. The chapter shows the system architecture, the Cloud4PSP processing model, and results of various scalability tests that speak in favor of the presented architecture.

The third part of the book shows the utilization of scalable Big Data computational frameworks, like Hadoop and Spark, in massive 3D protein structure alignments and identification of intrinsically disordered regions in protein structures.

- **Chapter 6: Foundations of the Hadoop Ecosystem**
  At the moment, Hadoop ecosystem covers a broad collection of platforms, frameworks, tools, libraries, and other services for fast, reliable, and scalable data analytics. This chapter briefly describes the Hadoop ecosystem and focuses on two elements of the ecosystem—the Apache Hadoop and the Apache Spark.

It provides details of the MapReduce processing model and differences between MapReduce 1.0 and MapReduce 2.0. The concepts defined in this chapter are important for the understanding of complex systems presented in the following chapters of this part of the book.

- **Chapter 7: Hadoop and the MapReduce Processing Model in Massive Structural Alignments Supporting Protein Function Identification**
  Undoubtedly, for a variety of biological data and a variety of scenarios of how these data can be processed and analyzed, Hadoop and the MapReduce processing model bring the potential to make a step forward toward the development of solutions that will allow to get insights in various biological processes much faster. In this chapter, you will see MapReduce-based computational solution for efficient mining of similarities in 3D protein structures and for structural superposition. The solution benefits from the Map-only processing pattern of the MapReduce, which is presented and formally defined in this chapter. You will also see results of performance tests when scaling up nodes of the Hadoop cluster and increasing the degree of parallelism with the intention of improving efficiency of the computations.
- **Chapter 8: Scaling 3D Protein Structure Similarity Searching on Large Hadoop Clusters Located in a Public Cloud**
  In this chapter, you will see how 3D protein structure similarity searching can be accelerated by distributing computation on large Hadoop/HBase (HDInsight) clusters that can be broadly scaled out and up in the Microsoft Azure public cloud. This chapter shows that the utilization of public clouds to perform scientific computations is very beneficial and can be successfully applied when performing time-consuming computations over biological data.
- **Chapter 9: Scalable Prediction of Intrinsically Disordered Protein Regions with Spark Clusters on Microsoft Azure Cloud**
  Computational identification of disordered regions in protein amino acid sequences became an important branch of 3D protein structure prediction and modeling. In this chapter, you will see the IDPP meta-predictor that applies an ensemble of primary predictors in order to increase the quality of prediction of intrinsically disordered proteins. This chapter presents a highly scalable implementation of the meta-predictor on the Spark cluster (Spark-IDPP) that mitigates the problem of the exponentially growing number of protein amino acid sequences in public repositories.

The fourth part of the book focuses on finding 3D protein structure similarities accelerated with the use of GPUs and on the use of multi-threading and relational databases for efficient approximate searching on protein secondary structures.

- **Chapter 10: Massively Parallel Searching of 3D Protein Structure Similarities on CUDA-Enabled GPU Devices**
  Graphics processing units (GPUs) and general-purpose graphics processing units (GPGPUs) promise to give a high speedup of many time-consuming and computationally demanding processes over their original implementations on CPUs. In this chapter, you will see that a massive parallelization of the 3D structure similarity searching on many-core CUDA-enabled GPU devices leads to the reduction of the execution time of the process and allows to perform it in real time.

- **Chapter 11: Exploration of Protein Secondary Structures in Relational Databases with Multi-threaded PSS-SQL**
  In this chapter, you will see how protein secondary structures can be stored in the relational database and explored with the use of the PSS-SQL query language. The PSS-SQL is an extension to the SQL language. It allows formulation of queries against a relational database in order to find proteins having secondary structures similar to the structural pattern specified by a user. In this chapter, you will see how this process can be accelerated by parallel implementation of the alignment using multiple threads working on multiple-core CPUs.

Summary

In this book, you will see advanced techniques and computational architectures that benefit from the recent achievements in the field of computing and parallelism. Techniques and methods presented in the successive chapters of this book will be based on various types of parallelism, including multi-threading, massive GPU-based parallelism, and distributed many-task computing in Big Data and Cloud computing environments (Fig. 1). Most of the problems are implemented as pleasantly or embarrassingly parallel processes, except the SQL-based search engine presented in Chap. 11, which employs multiple CPU threads in single search process.

Beautiful structures of proteins are definitely worth creating efficient methods for their exploration and analysis, with the aim of mining the knowledge that will improve human life in further perspective. While writing this book, I tried to pass through various representation levels of protein structures and show various techniques for their efficient exploration. In the successive chapters of the book, I described methods that were developed either by myself or as a part of projects that I was involved in. In the bibliography lists at the end of each chapter, I also cited other solutions for the presented problems and gave recommendations for further

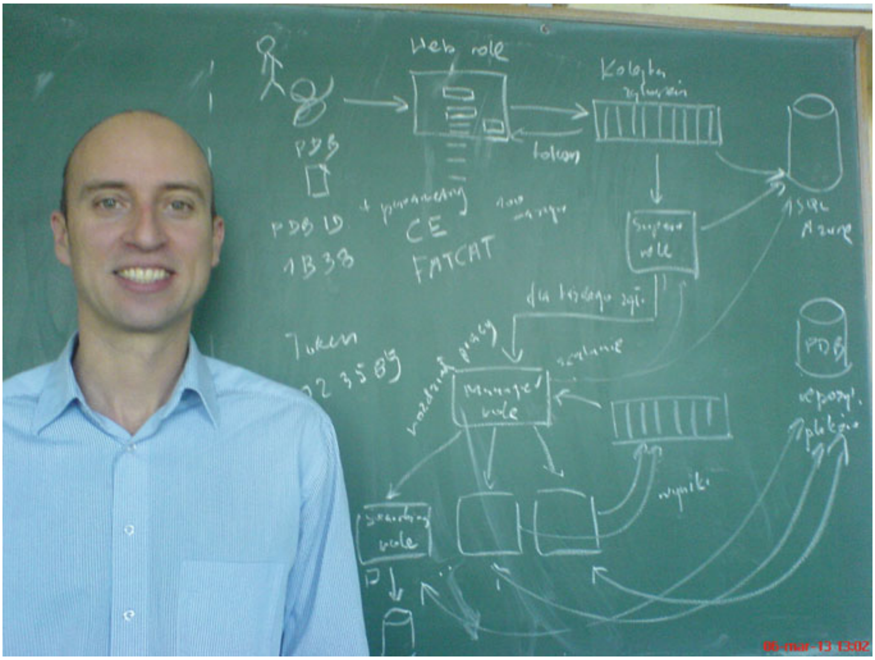


Fig. 1 Preliminary architecture of the cloud-based solution for protein structure similarity searching drawn by me during the meeting (March 6, 2013) with Artur Kłapciński, my associate in this project. Institute of Informatics, Silesian University of Technology, Gliwice, Poland

reading. I hope that the solutions presented in the book will turn out to be interesting and helpful for scientists, researchers, and software developers working in the field of protein bioinformatics.

Gliwice, Poland
June 2018

Dariusz Mrozek

Acknowledgements

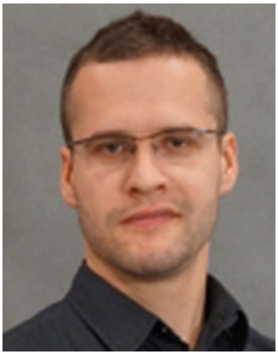
For many years, I have been trying to develop various efficient solutions for proteins and their structures. Through this time, there were many people involved in the research and development works that I carried out. I find it hard to mention all of them. I would like to thank my wife Bożena Małysiak-Mrozek, and also Tomasz Baron, Miłosz Brożek, Paweł Daniłowicz, Paweł Gosk, Artur Kłapciński, Bartek Socha, and Marek Suwała, for their direct cooperation in my research leading to the emergence of the book. A brief information on some of them is shown below. I would like to thank Alina Momot for her valuable advice on mathematical formulas, Henryk Małysiak for his mental support and constructive guidance resulting from the decades of experience in the academic and scientific work, and Stanisław Kozielski, a former Head of Institute of Informatics at the Silesian University of Technology, Gliwice, Poland, for giving me a space where I grew up as a scientist and where I could continue my research.

Bożena Małysiak-Mrozek received the M.Sc. and Ph.D. degrees, in computer science, from the Silesian University of Technology, Gliwice, Poland. She is an Assistant Professor in the Institute of Informatics at the Silesian University of Technology, Gliwice, Poland, and also a Member of the IBM Competence Center. Her scientific interests cover information systems, computational intelligence, bioinformatics, databases, Big Data, cloud computing, and soft computing methods. She participated in the development of all solutions and system for protein structure exploration presented in the book.

Tomasz Baron received the M.Sc. degree in computer science from the Silesian University of Technology, Gliwice, Poland in 2016. He currently works for Comarch S.A. company in Poland as software engineer. His interests cover cloud computing, front-end frameworks, and Internet technologies. He participated in the development of the Spark-based system for prediction of intrinsically disordered regions in protein structures presented in Chap. 9.

Miłosz Brożek received the M.Sc. degree in computer science from the Silesian University of Technology, Gliwice, Poland in 2012. He currently works for JSofteris company in Poland as Java programmer. His interests in IT cover microservices, cloud applications, and Amazon Web Services. He participated in the development of the CASSERT algorithm for protein similarity searching on CUDA-enabled GPU devices presented in Chap. 10.

Paweł Daniłowicz received the M.Sc. degree in computer science from the Silesian University of Technology, Gliwice, Poland in 2014. He currently works for Asseco Poland S.A. company in Poland as senior programmer. His interests in IT cover databases and business intelligence. He participated in the development of the HDInsight-/HBase-/Hadoop-based system for 3D protein structure similarity searching presented in Chap. 8.



Marek Suwała received the M.Sc. degree in computer science from the Silesian University of Technology, Gliwice, Poland in 2013. He currently works for Bank Zachodni WBK in Wrocław, Poland, as system analyst. His interests cover business process modeling and Web Services technologies. He participated in the development of the MapReduce-based application for identification of protein functions on the basis of protein structure similarity presented in Chap. 7.

Additional contributors to the development of the presented scalable and high-performance solutions were: (1) **Paweł Gosk** who participated in the implementation of the scalable system for 3D protein structure prediction working in the Microsoft Azure cloud presented in Chap. 5, (2) **Artur Kłapciński** who was the main programmer while constructing the cloud-based system for 3D protein structure alignment and similarity searching presented in Chap. 4, and (3) Bartek Socha who participated in the development of the multi-threaded version of the PSS-SQL language for efficient exploration of protein secondary structures in relational databases presented in Chap. 11.

Also, I would like to thank Microsoft Research for providing me a free access to computational resources of the Microsoft Azure cloud within the Microsoft Azure for Research Award grant. My special thanks go to Alice Crohas and Kenji Takeda from Microsoft, without whom my adventure with the Azure cloud would not be so long, interesting and full of new challenges.

The emergence of this book was supported by the Statutory Research funds of Institute of Informatics, Silesian University of Technology, Gliwice, Poland (grant No BK/213/RAU2/2018).

On a personal note, I would like to thank my family for all their love, patience, unconditional support, and understanding in the moments of my absence resulting from my desire to write this book.

Contents

Part II Cloud Services for Scalable Computations

Part III Big Data Analytics in Protein Bioinformatics

Part IV Multi-threaded Solutions for Protein Bioinformatics

Acronyms

AFP	Aligned fragment pair
BLOB	Binary large object
CASP	Critical Assessment of protein Structure Prediction
CE	Combinatorial Extension
CPU	Central processing unit
CUDA	Compute Unified Device Architecture
DAG	Directed acyclic graph
DBMS	Database management system
DNA	Deoxyribonucleic acid
ETL	Extract, transform, and load
FATCAT	Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists
GPGPU	General-purpose graphics processing units
GPU	Graphics processing unit
GUI	Graphical user interface
H4P	Hadoop for proteins
HDFS	Hadoop Distributed File System
IaaS	Infrastructure as a Service
MAS	Multi-agent system
MR	MapReduce
NoSQL	Non-SQL, non-relational
OODB	Object-oriented database
PaaS	Platform as a Service
PDB	Protein Data Bank
RDBMS	Relational database management system
RDD	Resilient distributed data set
RMSD	Root-mean-square deviation
SaaS	Software as a Service
SIMD	Single instruction, multiple data
SIMT	Single instruction, multiple thread

SQL	Structured Query Language
SSE	Secondary structure element
SVD	Singular value decomposition
VM	Virtual machine
XML	Extensible Markup Language
YARN	Yet Another Resource Negotiator

Part I
Background

Proteins are complex molecules that play key roles in biochemical reactions in cells of living organisms. They are built up with hundreds of amino acids and thousands of atoms, which makes the analysis of their structures difficult and time-consuming. This part of the book provides background information on proteins and their representation levels, including a formal model of a 3D protein structure used in computational processes related to protein structure alignment, superposition, similarity searching, and modeling. It also consists of a brief overview of technologies used in the solutions presented in this book, solutions that aim at accelerating computations underlying protein structure exploration.

Chapter 1
Formal Model of 3D Protein Structures for Functional Genomics, Comparative Bioinformatics, and Molecular Modeling

> *The great promise of structural bioinformatics is predicted on the belief that the availability of high-resolution structural information about biological systems will allow us to precisely reason about the function of these systems and the effects of modifications or perturbations*
>
> Jenny Gu, Philip E. Bourne, 2009

Abstract Proteins are the main molecules of life. Understanding their structures, functions, mutual interactions, activity in cellular reactions, interactions with drugs, and expression in body cells is a key to efficient medical diagnosis, drug production, and treatment of patients. This chapter shows how proteins can be represented in processes performed in scientific fields, such as functional genomics, comparative bioinformatics, and molecular modeling. The chapter begins with the general definition of protein spatial structure, which can be treated as a base for deriving other forms of representation. The general definition is then referenced to four representation levels of protein structure: primary, secondary, tertiary, and quaternary structures. This is followed by short description of protein geometry. And finally, at the end of the chapter, we will discuss energy features that can be calculated based on the general description of protein structure. The formal model defined in the chapter will be used in the description of the efficient solutions and algorithms presented in the following chapters of the book.

Keywords 3D protein structure · Formal model · Primary structure
Secondary structure · Tertiary structure · Quaternary structure · Energy features
Molecular modeling

D. Mrozek, *Scalable Big Data Analytics for Protein Bioinformatics*,
Computational Biology 28, https://doi.org/10.1007/978-3-319-98839-9_1

1.1 Introduction

From the biological point of view, the functioning of living organisms is tightly related to the presence and activity of proteins. Proteins are macromolecules that play a key role in all biochemical reactions in cells of living organisms. For this reason, they are said to be molecules of life. And indeed, they are involved in many processes, including reaction catalysis (enzymes), energy storage, signal transmission, maintaining cell's cytoskeleton, immune response, stimuli response, cellular respiration, transport of small bio-molecules, regulation of cell's growth and division.

Analyzing their general construction, proteins are macromolecules with the molecular mass above 10 kDa ($1\text{Da} = 1.66 \times 10^{-24}\text{g}$) built up with amino acids (>100 amino acids, aa). Amino acids are linked to each other by peptide bonds forming a kind of linear chains [5]. Proteins can be described with the use of four representation levels: primary structure, secondary structure, tertiary structure, and quaternary structure. The last three levels define the protein conformation or protein spatial structure. The computer analysis of protein structures is usually carried out on one of the representation levels.

The computer analysis of protein spatial structure is very important from the viewpoint of the identification of protein functions, recognition of protein activity and analysis of reactions and interactions that the particular protein is involved in. This implies the exploration of various geometrical features of protein structures. There is no doubt that structures of even small molecules are very complex—proteins are built up of hundreds of amino acids and then thousands of atoms. This makes the computer analysis of protein structures more difficult and also influences a high computational complexity of algorithms for the analysis.

For any investigation related to protein bioinformatics it is essential to assume some representation of proteins as macromolecules. Methods that operate on proteins in scientific fields, such as functional genomics, comparative bioinformatics, and molecular modeling, usually assume a kind of model of protein structure. Formal models, in general, allow to define all concepts that are used in the area under consideration. They guarantee that all concepts that are used while designing and performing a process will be understood exactly as they are defined by an author of the method or procedure. This chapter attempts to capture the common model of protein structure which can be treated as a base model for the creation of dedicated models, derived either by the extension or the restriction, and used for the computations carried out in the selected area. In the following sections, we will discover a general definition of protein spatial structure, and we will reference it to four representation levels of protein structure.

1.2 General Definition of Protein Spatial Structure

We define a 3D structure (S^{3D}) of protein P as a pair shown in Eq. 1.1.

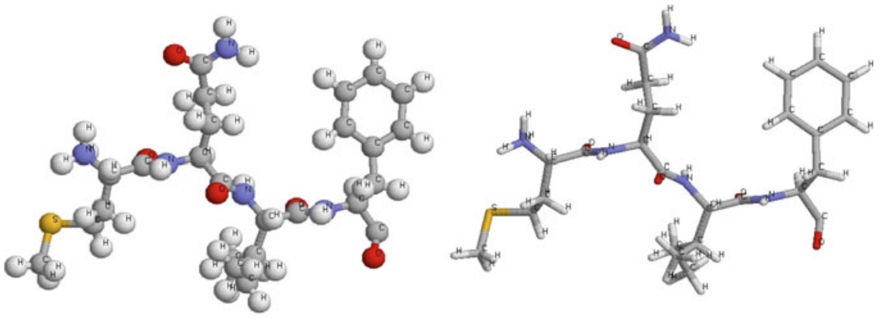


Fig. 1.1 Fragment of sample protein structure: (left) atoms and bonds, (right) bonds only. Colors and letters assigned to atoms distinguish their chemical elements. Visualized using RasMol [52]

$$S^{3D} = \langle A^{3D}, B^{3D} \rangle, \tag{1.1}$$

where A^{3D} is a set of atoms defined as follows:

$$A^{3D} = \left\{ a_n : n \in (1, \ldots, N) \quad \wedge \quad \exists f_E : A^{3D} \longrightarrow E \right\} \tag{1.2}$$

where N is the number of atoms in a structure, f_E is a function which for each atom a_n assigns an element from the set of chemical elements E (e.g., N—nitrogen, O—oxygen, C—carbon, H—hydrogen, S—sulfur).

The B^{3D} is a set of bonds b_{ij} between two atoms $a_i, a_j \in A^{3D}$ defined as follows:

$$B^{3D} = \{ b_{ij} : b_{ij} = (a_i, a_j) = (a_j, a_i) \quad \wedge \quad i, j \in (1, \ldots, N) \}. \tag{1.3}$$

Fragment of a sample protein structure is shown in Fig. 1.1.

Each atom a_n is described in three-dimensional space by Cartesian coordinates x, y, z:

$$a_n = (x_n, y_n, z_n)^T \quad \text{where} \quad x_n, y_n, z_n \in \mathbb{R}. \tag{1.4}$$

Therefore, the length of bond b_{ij} between two atoms a_i and a_j can be calculated using the Euclidean distance:

$$\|b_{ij}\| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}, \tag{1.5}$$

which is equivalent to the norm calculation [8]:

$$\|b_{ij}\| = \|a_i - a_j\| = \sqrt{(a_i - a_j)^T (a_i - a_j)}. \tag{1.6}$$

We can also state that:

$$a_n \in A^{3D} \Longrightarrow \forall_{n \in \{1,..,N\}} \; \exists f_{Va} : A^{3D} \longrightarrow \mathbb{N}_+ \; \wedge \; \exists f_{Ve} : E \longrightarrow \mathbb{N}_+, \qquad (1.7)$$

where f_{Va} is a function determining the valence of an atom and f_{Ve} is a function determining the valence of chemical element. For example, $f_{Ve}(C) = 4$ and $f_{Ve}(O) = 2$.

1.3 A Reference to Representation Levels

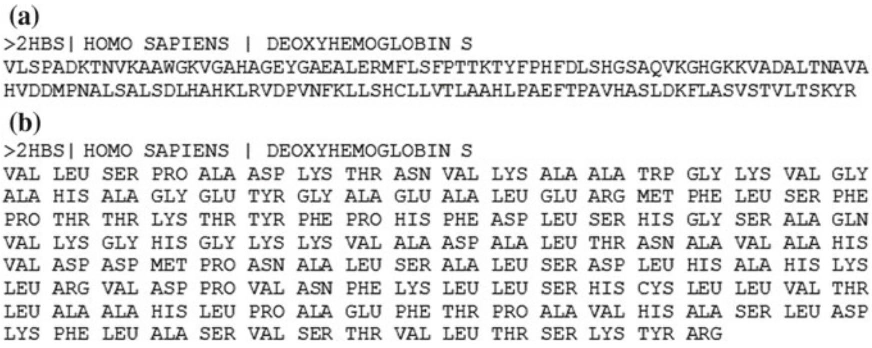
Having defined such a general definition of protein spatial structure, we can study what are the relationships between this structure and four main representation levels of protein structures, i.e., primary, secondary, tertiary and quaternary structures. These relationships will be described in the following sections.

1.3.1 Primary Structure

Proteins are polypeptides built up with many, usually more than one hundred amino acids that are joined to each other by a peptide bond, and thus, forming a linear amino acid chain. The way how one amino acid joins to another, e.g., during the translation from the mRNA, is not accidental. Each amino acid has an N-terminus (also known as amino-terminus) and C-terminus (also known as carboxyl-terminus). When two amino acids join to each other, they form a peptide bond between C-terminus of the first amino acid and N-terminus of the second amino acid. When a single amino acid joins the forming chain during the protein synthesis, it links its N-terminus to the free C-terminus of the last amino acid in the chain. Therefore, the amino acid chain is created from N-terminus to C-terminus. Primary structure of protein is often represented as the amino acid sequence of the protein (also called protein sequence, polypeptide sequence), as it is presented in Fig. 1.2. The sequence is reported from N-terminus to C-terminus. Each letter in the sequence corresponds to one amino acid. Actually, the sequence is usually recorded in one-letter code, and rarely in three-letter code.

Protein sequence is determined by the nucleotide sequence of appropriate gene in the DNA. There are twenty standard amino acids encoded by the genetic code in the living organisms. However, in some organisms two additional amino acids can be encoded, i.e., selenocysteine and pyrrolysine. All amino acids differ in chemical properties and have various atomic constructions. Proteins can have one or many amino acid chains. The order of amino acids in the amino acid chain is unique and determines the function of the protein.

The representation of protein structure as a sequence of amino acids from Fig. 1.2a is very simple and frequently used by many algorithms and tools for protein comparison and similarity searching, such as Needleman–Wunsch [46] and Smith–Waterman [58] algorithms, BLAST [1] and FASTA [49] family of tools. The representation



(a)

```
>2HBS| HOMO SAPIENS | DEOXYHEMOGLOBIN S
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVA
HVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

(b)

```
>2HBS| HOMO SAPIENS | DEOXYHEMOGLOBIN S
VAL LEU SER PRO ALA ASP LYS THR ASN VAL LYS ALA ALA TRP GLY LYS VAL GLY
ALA HIS ALA GLY GLU TYR GLY ALA GLU ALA LEU GLU ARG MET PHE LEU SER PHE
PRO THR THR LYS THR TYR PHE PRO HIS PHE ASP LEU SER HIS GLY SER ALA GLN
VAL LYS GLY HIS GLY LYS LYS VAL ALA ASP ALA LEU THR ASN ALA VAL ALA HIS
VAL ASP ASP MET PRO ASN ALA LEU SER ALA LEU SER ASP LEU HIS ALA HIS LYS
LEU ARG VAL ASP PRO VAL ASN PHE LYS LEU LEU SER HIS CYS LEU LEU VAL THR
LEU ALA ALA HIS LEU PRO ALA GLU PHE THR PRO ALA VAL HIS ALA SER LEU ASP
LYS PHE LEU ALA SER VAL SER THR VAL LEU THR SER LYS TYR ARG
```

Fig. 1.2 Primary structures of *Deoxyhemoglobin S* chain A in *Homo Sapiens* [PDB ID: 2HBS] [19]: **a** in a one-letter code describing amino acid types, **b** in a three-letter code describing amino acid types. First line provides some descriptive information

is also used by methods that predict protein structures from their sequences, like I-TASSER [63], Rosetta@home [29], Quark [64], and many others, e.g., [61] and [69].

Let us now reference the primary structure to the general definition of the spatial structure defined in the previous section. We can state that protein structure S^{3D} consists of M amino acids $P_m^{3D} \subsetneq S^{3D}$ such that:

$$P_m^{3D} = \langle A_m^P, B_m^P \rangle, \tag{1.8}$$

where A_m^P is a subset of the set of atoms A^{3D}, and B_m^P is a subset of the set of bonds B^{3D}:

$$A_m^P \subsetneq A^{3D} \text{ and } B_m^P \subsetneq B^{3D}. \tag{1.9}$$

Sample protein P can be now recorded as a sequence of peptides p_m:

$$P = \{p_m | i = 1, 2, \ldots, M \quad \wedge \quad \exists f_R : P \longrightarrow \Pi\}, \tag{1.10}$$

where M is a length of the sequence (in peptides), and f_R is a function which for each peptide p_m assigns a type of amino acid from the set Π containing twenty (twenty-two) standard amino acids.

Assuming that $p_m = P_m^{3D}$ we can associate the primary structure with the spatial structure S^{3D} (Fig. 1.3):

$$S^{3D} = \left\{P_m^{3D} | m = 1, 2, \ldots, M\right\}. \tag{1.11}$$

Although:

$$\bigcup_{m=1}^{M} P_m^{3D} \subsetneq S^{3D}, \tag{1.12}$$