Chapter 1

Introduction

Why it is important to study proteins and especially protein structures? Proteins play a variety of roles in life processes. Major examples of their biochemical functions include binding, catalysis, operating as molecular switches, and serving as structural components of cells and organisms [185]. The extraordinary functional diversity and versatility of proteins derives from their collective properties as a class of biomolecules: all proteins have the same underlying chemical unity; proteins have the ability to organize themselves in three dimensions; and the system that produces them can create inheritable structural variations, conferring the ability to evolve. Although structural data is not as complete as sequential data, detailed atomic protein structures are now available for over $40,000$ proteins and lead to implications in related fields like protein function analysis, protein evolution, protein structure prediction, protein engineering, or drug design.

Proteins perform their function by their three-dimensional (3D) structure. The catalytic activity of enzymes can be explained in terms of physico-chemical properties based on spatial contacts between amino acids. The amino acid sequences dictate the 3D structures of proteins. Anfinsen showed that all information necessary for a protein to fold to the native state resides in its amino acid sequence [9,10]. Under physiological conditions of solvent and temperature, most proteins fold spontaneously to an active native state. The final 3D structure of a protein is commonly referred to as its fold. Appropriately, the process by which a linear polypeptide chain achieves its distinctive fold is known as protein folding. Protein folding is the point at which nature makes the step from the one-dimensional information stored in the genetic code to the 3D world: DNA sequence encodes for protein sequence that encodes the 3D conformation of single polypeptide chains.

The Protein Data Bank (PDB, `http://www.rcsb.org`) [22] at Research Collaboratory for Structural Bioinformatics (RCSB) is the main collection of publicly available structures of proteins, nucleic acids, and other biological macromolecules determined with X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or low-temperature electron microscopy (cryo-EM). By May 2007 the PDB contained $43,633$ entries that can be classified into:

- $40,083$ protein structures, which may include cofactors, substrates, inhibitors, or other ligands including nucleic acids,

- 1,747 oligonucleotide or nucleic acid structures,
- 1,768 protein/nucleic acid complexes, and
- 35 other biomolecules including carbohydrate structures.

There is high amount of redundancy on sequential as well as on structural level within the PDB (see also Figure 1.1). Using Blast [4] the current version contains single protein chains that can be clustered on 90% level of sequence identity in 15,932 non-redundant clusters, and on a 30% sequence identity level into 8,850 different clusters.

Upon their determination of the first 3D globular protein structure, the oxygen-storage protein myoglobin, in 1958, John Kendrew and his co-workers registered their disappointment [123]:"Perhaps the most remarkable features of the molecule are its complexity and lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is more complicated than has been predicated by any theory of protein structure.".

Despite these initial frustrations, subsequent studies of protein structures based on more and more data of higher quality revealed that protein structure has some regularities and underlying principles. These can be organized in a four-tiered abstraction hierarchy describing protein structure: The primary structure is given by the amino acid sequence of a single protein chain. Within these chains there are regions in which the chains are organized into regular structures, namely helices and sheets. These secondary structure elements (SSEs) were predicted by Pauling and co-workers [180, 181] based on known physical constraints in polypeptide chains, prior to the experimental determination of protein structures. The tertiary structure of a protein is a description of the way the whole chain (including the SSEs) folds itself into its final 3D shape. Compact, globular regions within a single chain are called domains. Quaternary structure is the arrangement of multiple protein chains in a protein complex.

1.1 Why is Structure Comparison Important?

Analyzing and comparing protein structures are central issues of the post-genomic era. Only 15 years ago, sequencing the whole genome of even a simple organism appeared to be a task that would require decades. Major progress in molecular biology has made this process become reality. At the end of May 2007, the Genome Online Database lists 699 completed and 1,814 uncompleted genomes (see `http://www.genomesonline.org`). The full value of this large amount of sequence data will only be realized when function is assigned to every gene sequence. As it is not feasible to study every protein experimentally in all genomes, the function and biological role of a newly sequenced protein is usually inferred from a characterized protein using sequence and/or structure comparison methods. Functional inference based on sequence alone is limited by the so-called 'twilight zone', where sequence similarities can no longer be reliably detected (around 25% sequence identity) [200]. A striking feature of the protein structures deposited in the PDB is that nearly all proteins have structural similarities may arise from general physico-chemical principles that limit the number of different protein folds, or from evolutionary relationships. Therefore,

the theoretical analysis of protein structures that is based on a mathematically unique description became more and more important in order to search for similarities in proteins at different abstraction levels. Structure comparison of

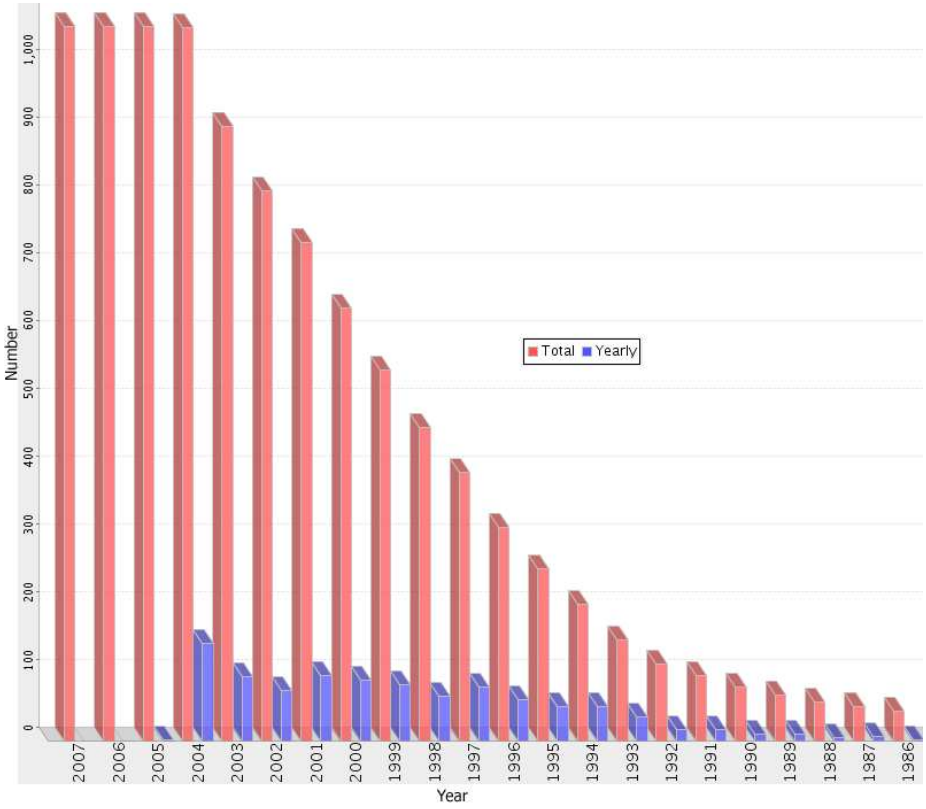

Figure 1.1: **Number of different folds in the PDB [22].** The diagram displays data on the growth in the number of unique folds per year in the PDB based on the SCOP [169] classification since 1986. The total number of folds per year is shown in red bars, the number of new folds per year in blue bars. Statistics are for experimentally determined structures. Data from the PDB website (`http://www.rcsb.org`).

a protein of known function with a protein of unknown function can provide new insight into the function of the unknown. Since structure is much more conserved than sequence during evolution, the discovery of structural alignment algorithms and the development of structural classification schemes like SCOP [169] or CATH [176] have made a significant contribution to the understanding of evolutionary mechanisms as they have enabled much more distant evolutionary relatives to be identified. In protein classification structure alignments help to determine fold classes and can be used subsequently in establishing libraries of templates for proteome annotation or protein structure prediction. Sometimes an entire 'new' structure will resemble that of another protein whose structure is already known. In most cases, however, the overall fold of the protein will be 'new', but the structure will be divisible into a number of domains, at least one, which resembles the tertiary structure previously observed in another

protein. Therefore, despite the increasing amount of solved protein structures the total amount on different folds in the PDB is on unchanging level of about 1100 fold types since three years (see Figure 1.1). This observation corresponds to the fact that the number of different protein folds in nature is limited [46]. They are repeatedly used in different combinations to create the diversity of proteins found in living organisms. Since sequence similarity between related folds is often absent, evolutionary relationship can only be detected by structure alignment methods alone.

Any categorization of a set of objects into clusters of similar objects requires a definition of similarity and dissimilarity. In the case of protein structures, such a measure is provided by structure comparison methods. Although significant process has been made over the past decades, a fast, reliable, and convergent method for protein structure alignment is not yet available. Recent developments have focused both on the search algorithm and on defining the scoring function to be optimized, that is, a quantitative measure of the quality of an alignment. A variety of programs for structure alignment have been introduced, but most of them ignore the fact that similar proteins often do not share the same ordering of their SSEs. However, there are biological meaningful structural motifs like the Rossmann fold whose arrangements of SSEs have been found in different sequential orderings.

The main topic of this thesis is the investigation of non-trivial similarities and functional relationships between protein structures. For the comparison and analysis of protein structures, it is of interest to find maximal common substructures between pairs of structures. This problem is also relevant for the discovery of biological important structural motifs and structure classification. In this thesis we describe suitable representations of protein structures as contact maps or protein graphs on the secondary structure level. Based on these representations we introduce graph-theoretical methods to search for common protein topologies or to perform pairwise structure alignments.

1.2 Outline of this Work

Chapter 2 introduces the basic concepts of protein structures and describes the different representations for protein structures that are used throughout this work: residue contact maps and protein graphs based on secondary structure. Additionally, short introductions into protein folding and protein structure evolution are given as well as a statistical analysis of globular protein structures. Chapter 3 shows how protein topologies can be modeled using graphs and how this description can be adopted to define linear notations that can be used to search efficiently for structural motifs. The most common supersecondary structure motifs are defined using these linear notations. Chapter 4 addresses the general protein structure alignment problem: the state-of-the-art structure alignment methods are introduced. Furthermore, it is described how structural similarity can be measured and how difficult it is to obtain significance for structure alignments. Structure comparisons are also the basis for the most important classification schemes for protein domains. In Chapter 5 a hierarchical method for non-sequential and gapped protein structure comparison is introduced. The basic step of the method is a maximal common subgraph search between protein graphs using a genetic algorithm. We have evaluated the new

alignment method on manually curated alignments and on large database scans. Chapter 6 is dealing with an exact graph-theoretical solution for the heuristic approach introduced in the preceding chapter. General properties of protein graphs are discussed. At the end we present a summary of the current work, discuss advantages and disadvantages of our structure alignment method, and give an outline how the method could be extended to multiple protein structure alignment.

We will use throughout this thesis the naming convention given in Appendix A to denote protein structures from the PDB, protein domains from SCOP and CATH, and protein graphs as defined in Chapter 3. Mathematical and graph-theoretical terms are defined in the text when they are used for the first time.