CHAPTER 1

Protein Structure Classification

Patrice Koehl

Department of Computer Science and Genome Center, University of California, Davis, California

INTRODUCTION

The molecular basis of life rests on the activity of large biological macromolecules, including nucleic acids (DNA and RNA), carbohydrates, lipids, and proteins. Although each plays an essential role in life, there is something special about proteins, as they are the lead performers of cellular functions. As a response, structural molecular biology has emerged as a new line of experimental research focused on revealing the structure of these bio-molecules. This branch of biology has recently experienced a major uplift through the development of high-throughput structural studies aimed at developing a comprehensive view of the protein structure universe. Although these studies are generating a wealth of information that are stored into protein structure databases, the key to their success lies in our ability to organize and analyze the information contained in those databases, and to integrate that information with other efforts aimed at solving the mysteries behind cell functions. In this survey, the first step behind any such organization scheme, namely the classification of protein structures, is described. The properties of protein structures, with special attention to their geometry, are reviewed. Computer methods for the automatic comparison and classification of these structures are then reviewed along with the existing classifications of protein structures and their applications in biology, with a special focus on computational biology. The chapter concludes the review with a discussion of the future of these classifications.

Reviews in Computational Chemistry, Volume 22
edited by Kenny B. Lipkowitz, Thomas R. Cundari, and Valerie J. Gillet

Classification and Biology

Classification is a broad term that simply means putting things into classes. Any organizational scheme is a classification: Objects can be sorted with respect to size, color, origin, and so on. Classification is one of the most basic activities in any discipline of science, because it is easier to think about a few groups that have something in common than it is to think about each individual of a whole population. Scientific classification in biology started with Aristotle, in the fourth century B.C. He divided all livings things into two groups: animal and plant. Animals were divided into two groups: those with blood and those without (at least no red blood), whereas plants were divided into three groups based on their shapes. Aristotle was the first in a long line of biologists who classified organisms in an arbitrary, although logical way, to convey scientific information. Among these biologists is the Swedish naturalist Carolus Linnaeus from the eighteenth century who set formal rules for a two-name system called the binomial system of nomenclature, which is still used today. With the publication of '*On the Origin of Species*' by Darwin, the purpose of classification changed. Darwin argued that classification should reflect the history of life. In other words, species should be related based on a shared history. *Systematic classifications* were introduced accordingly, the aims of which are to reveal the *phylogeny*, i.e., the hierarchical structure by which every life-form is related to every other life-form. The recent advances in genetics and biochemistry, the wealth of information coming from genome sequencing projects, and the tools of bio-informatics are playing an essential role in the development of these new classification schemes, by feeding to the classifiers and taxonomists more and more data on the evolutionary relationships between species. Note that the genetic information used for classification is not limited to the sequence of the genes, but it also takes into account the products of these genes, and their contributions to the mechanisms of life. Because function is related to shape, protein structure classification will thus play a significant role in our understanding of the organization of life. Paraphrasing Jacques Monod[1], in the protein lies the secret of life.

The Biomolecular Revolution

All living organisms can be described as arrangements of cells, the smallest self-sustainable units capable of carrying functions important for life. Cells can be divided into organelles, which are themselves assemblies of biomolecules. These bio-molecules are usually polymers composed of smaller subunits whose atomic structures are known from standard chemistry. There are many remarkable aspects to this hierarchy, one of them being that it is ubiquitous to all life forms, from unicellular organisms to complex multicellular species. Unraveling the secrets behind this hierarchy has become one of the major

challenges for scientists in the twentieth and now twenty-first centuries. Although early research from the physics and chemistry communities has provided significant insight into the nature of atoms and their arrangements in small chemical systems, the focus is now on understanding the structure and function of bio-molecules. These usually large molecules serve as storage for the genetic information (the nucleic acids) and as key actors of cellular functions (the proteins). Biochemistry, one field in which these bio-molecules are studied, is currently experiencing a major revolution. In hope of deciphering the rules that define cellular functions, large-scale experimental projects are now being performed as collaborative efforts involving many laboratories in many countries to provide maps of the genetic information of different organisms (the *genome projects*), to derive as much structural information as possible on the products of the corresponding genes (the *structural genomics projects*), and to relate these genes to the function of their products, which is usually deduced from their structure (the *functional genomics projects*). The success of these projects is completely changing the landscape of research in biology. As of October 2004, more than 220 whole genomes have been fully sequenced and published, which corresponds to a database of over a million gene sequences,[2] and more than a thousand other genomes are currently being sequenced. The need to store these data efficiently and to analyze their contents has led to the emergence of a collaborative effort between researchers in computer science and biology. This new discipline is referred to as bio-informatics. In parallel, the repository of bio-molecular structures[3,4] contains more than 27,600 entries of proteins and nucleic acids. The same need to organize and analyze the structural information contained in this database is leading to the emergence of another partnership between computer science and biology, namely the discipline of bio-geometry. The combined efforts of researchers in bio-informatics and bio-geometry are expected to provide a comprehensive picture of the protein sequence and structure spaces, and their connection to cellular functions. Note that the emergence of these two disciplines is often viewed as a consequence of a paradigm shift in molecular biology,[5] because the classic approach of hypothesis-driven research in biochemistry is being replaced with a data-driven discovery approach. In reality the two approaches coexist, and both benefit from these computer-based disciplines.

Outline

Given the introduction to classification in biology and an update on the progress of research in structural biology, we can now examine protein structure classification, the topic of this chapter. The next section describes proteins and surveys their different levels of organization, from their primary sequence to their quaternary structure in cells. The following section surveys automatic methods for comparing protein structures and their application to classification. Then the existing protein structure classifications are described, focusing on the Structural Classification of Proteins (SCOP)[6]; the Class, Architecture,

Topology, and Homologous (CATH) superfamilies classification[7]; and the domain classification based on the Distance ALIgnment (DALI) algorithm.[8] Finally, the tutorial concludes with a discussion of the future of protein structure classifications.

BASIC PRINCIPLES OF PROTEIN STRUCTURE

Although all bio-molecules play an important role in life, there is something special about proteins, which are the products of the information contained in the genes. A finding that has crystallized over the last few decades is that geometric reasoning plays a major role in our attempt to understand the activities of these molecules. In this section, the basic principles that govern the shapes of protein structures are briefly reviewed. More information on protein structures can be found in protein biochemistry textbooks, such as those of Schulz and Schirmer,[9] Cantor and Schimmel,[10] Branden and Tooze,[11] and Creighton.[12] The reader is also referred to the excellent review by Taylor et al.[13]

Visualization

The need for visualizing bio-molecules is based on our early understanding that their shape determines their function. Early crystallographers who studied proteins could not rely (as it is common nowadays) on computers and computer graphics programs for representation and analysis. They had developed a large array of finely crafted physical models that allowed them to represent these molecules. Those models, usually made out of painted wood, plastic, rubber, or metal, were designed to highlight different properties of the molecule under study. In space-filling models, such as those of Corey–Pauling–Koltun (CPK),[14,15] atoms are represented as spheres, whose radii are the atoms' van der Waals radii. They provide a volumetric representation of the bio-molecules and are useful to detect cavities and pockets that are potential active sites. In skeletal models, chemical bonds are represented by rods, whose junctions define the position of the atoms. Those models were used for example by Kendrew et al. in their studies of myoglobin.[16] Such models are useful to chemists because they help highlight the chemical reactivity of the bio-molecule under study and, consequently, its potential activity. With the introduction of computer graphics to structural biology, the principles of these models have been translated into software such that molecules as well as some of their properties can now be visualized on a computer display. Figure 1 shows examples of computer visualizations of myoglobin, including space-filling and skeletal representations. Many computer programs are now available that allow one to visualize bio-molecules. Cited here are MOLSCRIPT[17] and VMD,[18] which have generated most of the figures of this chapter.

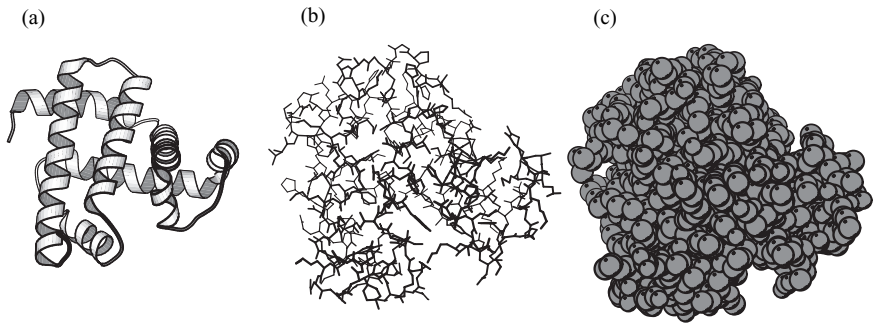


Figure 1 Visualizing protein structures. Myoglobin is a small protein very common in muscle cells, where it serves as oxygen storage. The structure of sperm whale myoglobin using three different types of visualization is depicted without the heme group. The coordinates are taken from the PDB file 1mbd. (a) **Cartoon.** This representation, also referred to as "ribbon" diagram, provides a high-level view of the local organization of the protein in secondary structures, shown as idealized helices. (b) **Skeletal model.** This representation uses lines to represent bonds; atoms are located at their endpoints where the lines meet. (c) **Space-filling diagram.** Atoms are represented as balls centered at the atoms, with radii equal to the van der Waals radii of the atoms. This representation shows the tight packing of the protein structure. Each of the representations is complementary to the others. Figure drawn using MOLSCRIPT.[17]

Protein Building Blocks

Proteins are heteropolymer chains of amino acids often referred to as *residues*. There are 20 naturally occurring amino acids that make up proteins. With the exception of proline, amino acids have a common structure, which is shown in Figure 2a. Naturally occurring amino acids that are incorporated into proteins are, for the most part, the levorotary (L) isomer. Substituents on the alpha carbon, called *side chains*, range in size from a single hydrogen atom to large aromatic rings. Those substituents can be charged, or they may include only nonpolar saturated hydrocarbons[19] (see Table 1 and Figure 2b). Nonpolar amino acids do not have a concentration of electric charges and are usually not soluble in water. Polar amino acids carry local concentration of charges and are either globally neutral, negatively charged (acidic), or positively charged (basic). Acidic and basic amino acids are classically referred to as electron acceptors and electron donors, respectively, which can associate to form salt bridges in proteins. Amino acids in solution are mainly dipolar ions: The amino group NH_2 accepts a proton to become NH_3^+, and the carboxyl group COOH donates a proton and becomes COO^-.

Protein Structure Hierarchy

Condensation between the $-NH_3^+$ and the $-COO^-$ groups of two amino acids generates a peptide bond and results in the formation of a dipeptide.

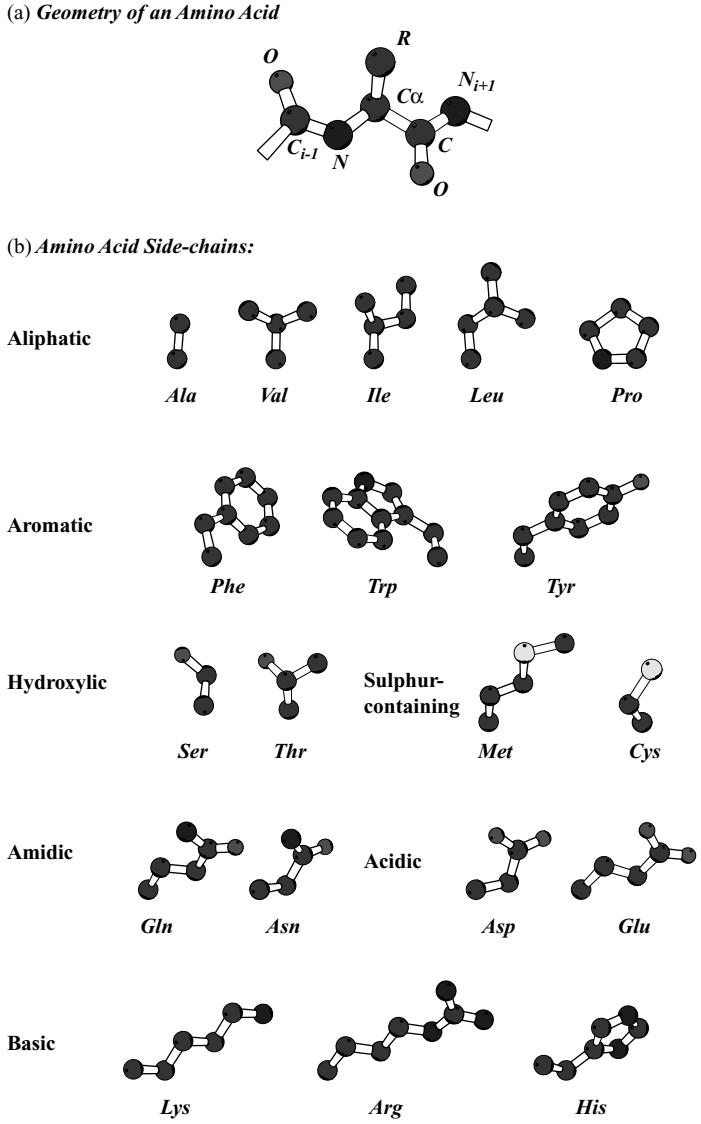


Figure 2 The twenty natural amino acids that make up proteins. (a) Each amino acid has a main-chain (N, C_α, C, and O) on which is attached a side-chain schematically represented as R. Amino acids in proteins are attached through planar peptide bonds, connecting atom C of the current residue to atom N of the following residue. For the sake of simplicity, the hydrogens are omitted. (b) Classification of the amino acid side-chains R is according to their chemical properties. Glycine (Gly) is omitted, as its side-chain is a single H atom.

Table 1 Classification of the 20 Amino Acids Based on Their Interaction With Water[19]

Classification	Amino Acid
Nonpolar	glycine (G)[a], alanine (A), valine (V), leucine (L), isoleucine (I), proline (P), methionine (M), phenylalanine (F), tryptophan (W)
Polar	serine (S), threonine (T), asparagine (N), glutamine (Q), cysteine (C), tyrosine (Y)
Acidic (polar)	aspartic acid (D), glutamic acid (E)
Basic (polar)	lysine (K), arginine (R), histidine (H)

[a] The one-letter code of each amino acid is given in parentheses.

Protein chains correspond to an extension of this chemistry, which results in long chains of many amino acids bonded together. The order in which amino acids appear defines the *sequence* or *primary structure* of the protein. In its native environment, the polypeptide chain adopts a unique three-dimensional shape, which is referred to as the *tertiary* or *native structure* of the protein.[20] The amino acid backbones are connected in sequence forming the protein *main-chain*, which frequently adopts canonical local shapes or *secondary structures*, mostly α-helices and β-strands (see Figure 3). α-helices form a right-handed helix with 3.6 amino acids per turn, whereas the β-strands form an approximately planar layout of the backbone. Helices often pack together to form a hydrophobic core, whereas β-strands pair together to form parallel or antiparallel β-sheets. In addition to these two types of secondary structures, a wide variety of other commonly occurring substructures, which are referred to as *super-secondary structures*. More information about these substructures can be found in the work of Efimov.[21–24]

Three Types of Proteins

Protein structures come in a large range of sizes and shapes. They can be divided into three major groups: *fibrous* proteins, *membrane* proteins, and *globular* proteins.

Fibrous proteins are elongated molecules in which the secondary structure is the dominant structure. Because they are insoluble in water, they play a structural or supportive role in the body and are involved in movement (such as in muscle and ciliary proteins). Fibrous proteins often (but not always) have regular repeating structures. Keratin, for example, which is found in hair and nails, is a helix of helices and has a seven-residue repeating structure. Silk, on the other hand, is composed only of β-sheets, with alternating layers of glycines and alanines and serines. In collagen, the major protein component of connective tissue, every third residue is a glycine and many others are prolines.

Membrane proteins are restricted to the phospholipid bilayer membrane that surrounds the cell and many of its organelles. These proteins cover a large

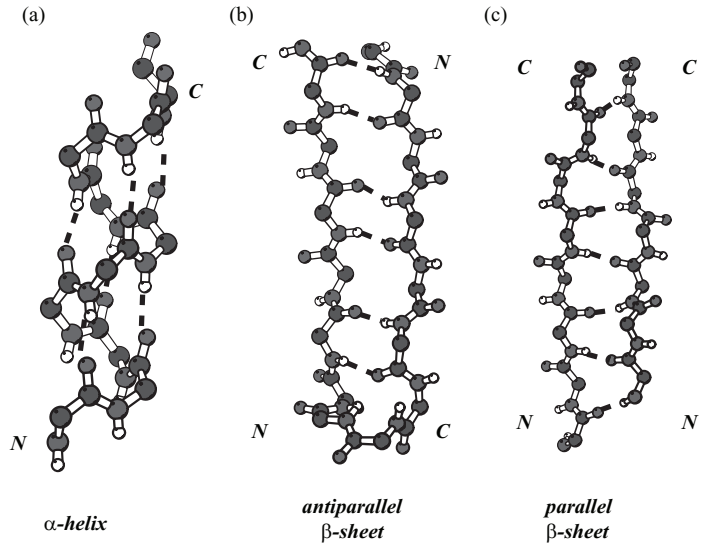
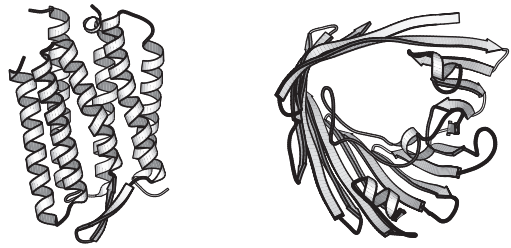


Figure 3 The three most common secondary structure elements (SSE) found in proteins. (a) The regular α-helix is a right-handed helix, in which all residues adopt similar conformations. The α-helix is characterized by hydrogen bonds between the oxygen O of residue *i*, and the polar backbone hydrogen HN (bound to N) of residue $i + 4$. Note that all C = O and N-HN bonds are parallel to the main axis of the helix. (b) An antiparallel β-sheet. Two strands (stretches of extended backbone segments) are running in an antiparallel geometry. The atoms HN and O of residue *i* in the first strand hydrogen bond with the atoms O and HN of residue *j* in the opposite strand, respectively, whereas residues $i + 1$ and $j + 1$ face outward. (c) A parallel β-sheet. The two strands are parallel, and the atoms HN and O of residue *i* in the first strand hydrogen bond with the O of residue *j* and the HN of residue $j + 2$, respectively. The same alternating pattern of residues involved in hydrogen bonds with the opposite strand, and facing outward is observed in parallel and antiparallel β-sheets. A strand can therefore be involved in two different sheets. For simplicity, side-chains and nonpolar hydrogens are ignored. The protein backbone is shown with balls and sticks, and hydrogen bonds are shown as discontinuous lines. Figure drawn using MOLSCRIPT.[17]

range of sizes and shapes, from globular proteins anchored in the membrane by means of a tail to proteins that are fully embedded in the membrane. Their function is usually to ensure transport of ions and small molecules like nutrients through the membrane. The structures of fully embedded membrane proteins can be placed into two major categories: the all helical structures, such as bacteriorhodopsin, and the all beta structures, such as porins (see Figure 4). As of October 2004, there are 158 structures of membrane proteins in the Protein Data Bank (PDB), out of which 86 are unique.

Globular proteins have a nonrepetitive sequence. They range in size from 100 to several hundred residues and adopt a unique compact structure. In globular proteins, nonpolar amino acid side chains have a tendency to cluster together to form the interior, hydrophobic core of the proteins, whereas the



(a) Bacteriorhodopsin (b) Porin

Figure 4 Two examples of membrane proteins. (a) Bacteriorhodopsin is mainly an α-protein containing seven helices. It is a membrane protein serving as an ion pump and is found in bacteria that can survive in high salt concentrations. (b) Porin is a β-barrel. Porins work as channels in cell membranes, which let small metabolites such as ions and amino acids in and out of the cell. Figure drawn using MOLSCRIPT.[17]

hydrophilic polar amino acid side chains remain accessible to the solvent on the exterior of the ''glob.'' In the tertiary structure, β-strands are usually paired in parallel or antiparallel arrangements to form β-sheets. On average, a protein main-chain consists of about 25% of residues in α-helix formation and 25% of residues in β-strands, with the rest of the residues adopting less-regular structural arrangements.[25]

Geometry of Globular Proteins

From the seminal work of Anfinsen,[26] we know that the sequence fully determines the three-dimensional structure of a protein, which itself defines its function. Although the key to the decoding of information contained in genes was found more than 50 years ago (the genetic code), we have not yet rigorously defined the rules relating a protein sequence to its structure.[27,28] Ongoing work in the area of predicting protein structure based on sequence is the topic of Chapter 3 by Shea et al.[29] Our knowledge of protein structure comes from years of experimental studies, primarily using X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. The first protein structures to be solved were those of myoglobin and hemoglobin.[16,30] There are now over 27,700 protein structures in the PDB database[3,4] (see http://www.rcsb.org). It is to be noted that this number overestimates the actual number of different structures available because the PDB is redundant; i.e., it contains several copies of the same proteins, with minor mutations in the sequence and no changes in the structure.

Because only two types of secondary structures (α and β) exist, proteins can be divided into three main structural classes.[31] These are mainly α proteins,[32] mainly β proteins,[33–35] and mixed α–β proteins.[36] A fourth class includes proteins with little or no secondary structures at all that are stabilized by metal ions and/or disulphide bridges. A significant effort has been made by

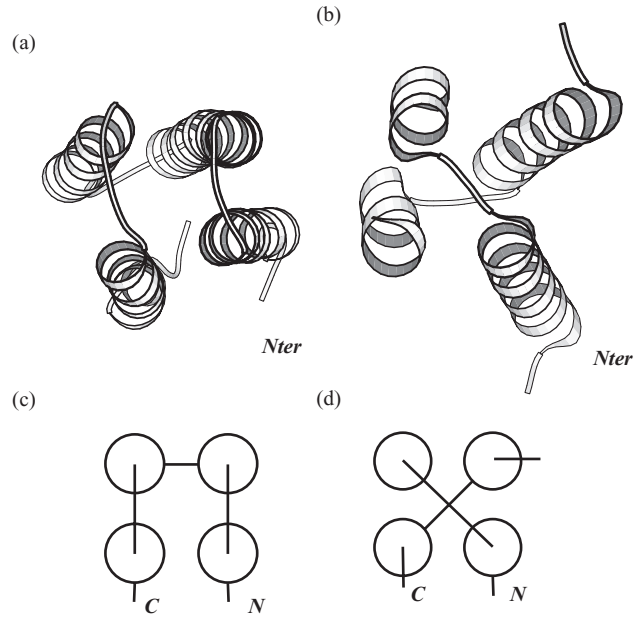


Figure 5 Two different topologies of four-helix bundles. A bundle is an array of α-helices, each oriented roughly along the same (bundle) axis. (a) and (c) show a four helical, up-and-down bundle with a left-handed twist, observed in hemerythrin from a sipunculid worm. (b) and (d) show a four helix bundle with a right handed twist, observed in a fragment of the dimerization domain of a liver transcription factor. (a) and (b) are cartoon representations of the proteins obtained with MOLSCRIPT,[17] whereas (c) and (d) show the schematic topologies produced by TOPS (http://www.tops.leed.ac.uk/).

scientists to a folding class to proteins automatically; these efforts will be reviewed in the next section. There has also been significant work on predicting a protein folding class based on its sequence, the details of which can be found in Refs. 37–44.

The α class, the smallest of the three major classes, is dominated by small proteins, many of which form a simple bundle of α helices packed together to form a hydrophobic core. A common motif in the mainly α class is the four helix bundle structure, which is depicted in Figure 5. The most extensively studied α structure is the globin fold, which has been found in a large group of related proteins, including myoglobin and hemoglobin. This structure includes eight helices that wrap around the core to form a pocket where a heme group is bound.[16]

The β class contains the parallel and antiparallel β structures. The β strands are usually arranged in two β sheets that pack against each other and form a distorted barrel structure. Three major types of β barrels exist, the up-and-down barrels, the Greek key barrels,[45] and the jelly roll barrels (see Figure 6). Most known antiparallel β structures, including the

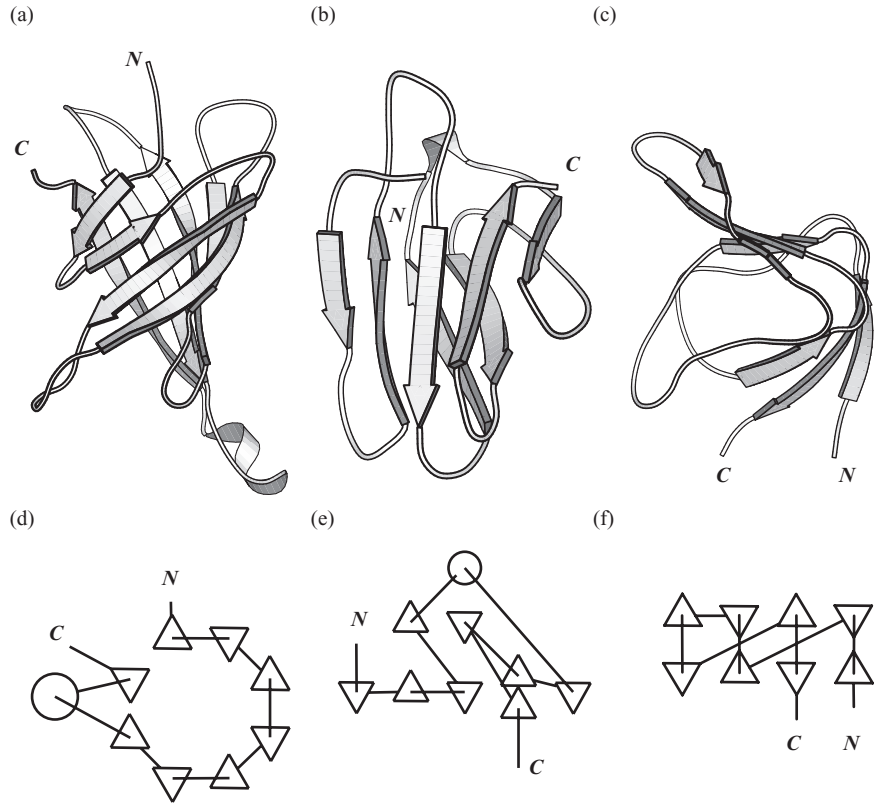


Figure 6 Three common sandwich topologies of beta proteins: a meander (a and d) observed in a glycoprotein from chicken, a Greek key (b and e) observed in and α-amylase (PDB code 1bli), and a jelly roll (c and f) observed in a gene activator protein from *E. Coli* (PDB code 1g6n). A meander (or up-and-down) is a simple topology in which any two consecutive strands are adjacent and antiparallel. A Greek key motif is a topology of a small number of β-sheet strands in which some inter-strand connection exist between β-sheets. The jelly roll topology is a variant of the Greek key topology with both ends crossed by two inter-strand connections. a, b, and c are cartoon representations of the proteins obtained with MOLSCRIPT,[17] while d, e and f show the schematic topologies produced by TOPS (http://www.tops.leed.ac.uk/).

immunoglobulins, have barrels that include at least one Greek key motif. The two other motifs are observed in proteins of diverse function, where functional diversity is obtained by differences in the loop regions connecting the β strands. β structures are often characterized by the number of β-sheets in the structure and the number and direction of the strands in the sheet. It leads to a rigid classification scheme,[46] which is sensitive to the definition of hydrogen bonds and β-strands.

The α–β protein class is the largest of the three classes. It is subdivided into proteins having an alternating arrangement of α helices and β strands

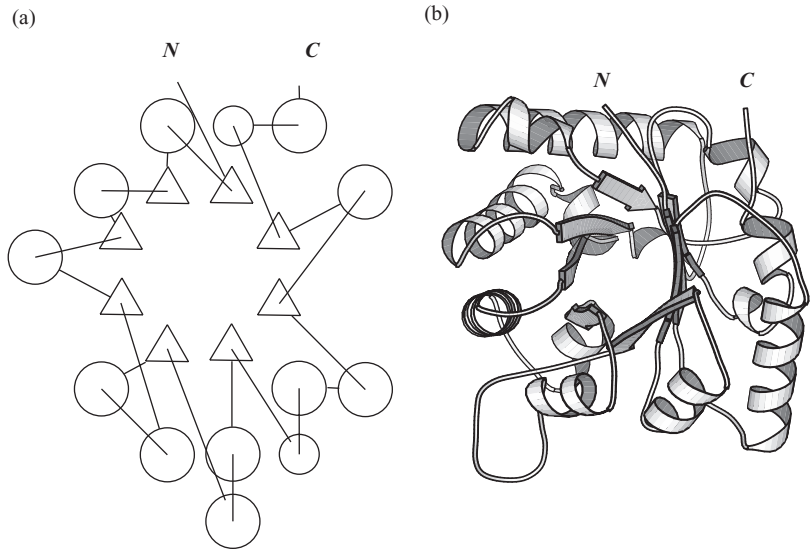


Figure 7 Topology (a) **and cartoon representation** (b) **of the TIM barrel**. The protein chain alternates between β and α secondary structure type, giving rise to a barrel β-sheet in the center surrounded by a large ring of α-helix on the outside. This structure, first seen in the triose phosphate isomerase of chicken, has been observed in many unrelated proteins since then. The topology is drawn using TOPS (http://www.tops.leed.ac.uk/), and the cartoon is generated using MOLSCRIPT.[17]

along the sequence and those with more segregated secondary structures. The former subclass is divided into two groups: one with a central core of (often eight) parallel β strands arranged as a barrel surrounded by α helices, and a second group consisting of an open, twisted parallel or mixed β sheet, with α helices on both sides (see Figure 7). A particularly striking example of a β–α barrel is seen in the eight-fold β–α barrel $(\beta\alpha)_8$ that was found originally in the triose phosphate isomerase of chicken,[47] and is often referred to as the TIM-barrel (for a complete analysis, see Refs. 48–55). Many proteins adopting a TIM barrel structure have completely different amino acid sequences and different biological functions. The open α/β-sheet structures vary considerably in size, number of β strands, and their strand order.

Protein Domains

Large proteins do not contain a single large hydrophobic core, probably because of limitations in their folding kinetics and stability. Large proteins are organized into ''units'' with sizes around 200–300 residues, which are referred to as *domains*.[56–58] Single compact units of more than 500 amino acids are rare. For a detailed analysis of domains in proteins, see Ref. 59. There are five different working definitions of protein domains: (1) regions that display

a significant level of sequence similarity; (2) the minimal part of a protein that is capable of performing a function; (3) a region of a protein with an experimentally assigned function; (4) a region of a structure that recurs in different contexts in different proteins; and (5) a compact, spatially distinct unit of protein structure. As more structures of proteins are solved, contradictions in these definitions appear. Some domains are compact, whereas others are clearly not globular. Some are too small to form a stable domain and thus lack a hydrophobic core. Currently, we are in the awkward situation in which the concept of a structural domain is well accepted, yet its definition is ambiguous;[60] this will be discussed in detail in the next section.

Resources on Protein Structures

Many resources related to protein structure and function exist; the Web addresses of these services are compiled in Table 2. Almost all experimental

Table 2 Resources on Protein Structures

Scheme	Description	Web Address
PDB	Repository of protein structures	http://www.rcsb.org/
PDB at a Glance	Interface to PDB	http://cmm.info.nih.gov/modeling/pdb_at_a_glance.html
Molecules to Go	Interactive interface to the PDB	http://molbio.info.nih.gov/cgi-bin/pdb/
MSD	EBI interface to the PDB, with integration to EBI resources	http://www.ebi.ac.uk/msd/
PDBSum	Summaries and structural analyses of PDB files	http://www.ebi.ac.uk/thornton-srv/databases/pdbsum
Biotech Validation Suite	Suite of programs that generates a quality control on protein structures	http://biotech.ebi.ac.uk:8400/
NRL_3D	Sequence-structure databases	http://laguerre.psc.edu/general/software/packages/nrl_3d/
Entrez	NCBI databases	http://www.ncbi.nlm.nih/gov/Database/index.html
SRS	Sequence Retrieval Services (includes structural information)	http://srs.embl-heidelberg.de:8000/srs5/
DSSP	Database of secondary structures of proteins (available through SRS)	http://srs.embl-heidelberg.de:8000/srs5/
TOPS	Generates a cartoon of the topology of a protein	http://www.tops.leeds.ac.uk/
PISCES	Protein sequence culling server: generates subsets of PDB based on users' criteria	http://dunbrack.fccc.edu/PISCES.php/
ASTRAL	Databases and tools for analyzing protein structure; derived from SCOP	http://astral.berkeley.edu/

protein structures publicly available today are stored in the PDB,[3] a database maintained through the RCSB consortium,[4] and available on the Web at http://www.rcsb.org/. Many services have been developed to supplement the PDB to ease access to the information it contains. For example, the services "PDB at a glance" and "Molecules to Go" were designed as easy-to-use interfaces with simple search engines. The MSD relational database is derived from the PDB and has the aim of providing a knowledge discovery and data mining environment for biological structure data. PDBSum[61,62] and the Biotech Validation Suite are services that allow a user to check the quality of a protein structure. NRL, Entrez, and SRS are integrated services that regroup the PDB with other databases containing information about proteins. For example, SRS includes DSSP,[63] a database of secondary structures of proteins. PISCES[64] and ASTRAL[65–67] can generate subsets of the PDB database, based on the user's criteria.

PROTEIN STRUCTURE COMPARISON

Any attempts to study a large collection of objects usually start with classifying them according to a given measure of similarity. Protein structure similarity is most often detected and quantified by a protein structure alignment program, which is applied to the different domains of the proteins considered. In this section, existing techniques for automatically detecting domains in protein structures are reviewed along with techniques for finding the optimal alignment between two structural domains. The section concludes with a brief description of new techniques for comparing protein structural domains that do not rely on a structural alignment, but instead rely on a direct comparison of the topology of the domains.

Automatic Identification of Protein Structural Domain

Decomposition of multidomain protein structures into individual domains has traditionally been done manually. Because the rate of protein structure determination has increased drastically in the past few years, this manual process has now become a bottleneck in maintaining and updating protein structure classifications; there is a need for automation. Automatic decomposition of proteins into structural domains can be traced back to the work of Rossman and Liljas in 1974,[68] who used $C\alpha$-$C\alpha$ distance maps. They suggested that a domain has, internally, many short residue–residue distances, but few short distances with the rest of the protein. Their analysis of the distance plots, however, required human intervention. Crippen[20] generalized this concept using hierarchical cluster analysis to locate protein fragment–fragment contacts. His procedure generates a tree of protein fragments, from a small, locally compact region of the complete protein. Several methods have

been proposed subsequently that follows this concept of identifying domains based on a difference between intradomain and interdomain properties. Some of these methods follow up on the work of Rossman and Liljas and compare intradomain and interdomain distances,[69–72] whereas others evaluate contact surface area between domains,[73,74] the "compactness" of the domains,[56,75,76] or their dynamics.[77] Recursive algorithms have been developed to find the cutting points that delineate domains in a protein chain. These algorithms either scan the chain to find single cuts such that the two resulting fragments verify a given protein domain definition based on one of the properties enumerated above or look directly for multiple cuts (see, for example, Ref. 71). The problem of delineating protein domains has also been formulated as an eigenvalue problem on the $C\alpha$-$C\alpha$ distance matrix,[77] as well as a network flow problem.[78,79]

These methods take the approach in which a predefined domain definition is imposed on the structural data. In the language of systems analysis, such methods are referred to as "top-down" approaches, and the inherent problem in their applications is the difficulty in recognizing when the data fit or do not fit the model. An alternative approach is to reverse the direction and let the model emerge from the data, in what is often referred to as a "bottom-up" approach. Taylor[80] recently developed a "bottom-up" approach to identify domains in protein, using an Ising model, in which the structural elements of the model change state according to a function of the state of the neighbors. His procedure works as follows. Each residue in the protein chain is assigned a numeric label, usually the sequential residue number. If a residue i with label s_i is surrounded by neighbors with, on average, a higher label, its label increases; otherwise it decreases. This procedure is iterated until the system reaches equilibrium. Special care is taken to ensure (1) that the protein chain does not pass between domains too frequently; (2) that secondary structures, in particular, β-sheets are not broken; and (3) that small domains are either ignored or avoided. Swindells developed an alternative "bottom-up" approach, in which he first identifies core regions in the protein,[81] which are then extended to define the different domains in the proteins.[82]

Most existing methods for identifying protein domains include a refinement scheme to assess the quality of the domains that have been identified. Domain quality is computed according to accessible surface area, hydrophobic moment profile, size, compactness, number of protein segments involved,[79] and presence of intact β sheets.[80]

The diversity of definitions for protein structural domains is a serious issue for the generation of protein structure classifications. Many programs have been developed to delineate domains automatically in multidomain proteins. Table 3 lists the programs that are currently accessible on the Web, either as a Web service or for download. Although the results of these programs agree in most cases, discrepancies still prevent consistent assignments of protein domains.[60] The absence of quality control in the results of protein domain assignment programs has led researchers to use a combination of

Table 3 Websites for Publicly Available Services or Programs for Protein Domain Assignment

Program	Web Access
DIAL	http://www.ncbs.res.in/~faculty/mini/ddbase/dial.html
DomainParser	http://compbio.ornl.gov/structure/domainparser
DOMAK	http://www.compbio.dundee.ac.uk/Software/Domak/domak.html
PDP	http://123d.ncifcrf.gov/pdp.html

automatic and manual methods. For example, CATH[7] defines domains in multidomain proteins based on a consensus of three automatic programs, namely PUU,[77] DOMAK,[83] and Detective.[82] When all three programs agree on an assignment, the corresponding domains are included in CATH. In cases of disagreement, the domains are assigned manually, either from visual inspection or from information available in the literature or on the Web. Several structural domain databases are available on the Web to assist manual assignments of domains (see Table 4).

The Rigid-Body Transformation Problem

Definition

Before one attempts to classify protein structures, it is important to evaluate structure similarities. Many ways exist in which protein structures can be compared, that will be reviewed below. Most of these approaches proceed in two steps: (1) find the transformation that provides the optimal superposition between the two structures, and (2) define the similarity score as the distance between the two structures after superposition. This section describes how to obtain the optimal transformation for step (1).

We begin with the (relatively) easy problem of comparing two protein structures with the same number of atoms and a known correspondence table between these atoms (for a review, see Ref. 84). This problem is often solved when comparing two possible models for the structure of a protein. Because it is such a common problem, and because there still exists some confusion about

Table 4 Databases of Protein Structural Domains

Database	Web Access	Method
3Dee	http://www.compbio.dundee.ac.uk/3Dee	DOMAK
Authors	http://www.bmm.icnet.uk/~domains/test/dom-rr.html	Domains identified in the literature
DALI	http://www.ebi.ac.uk/dali/domain/3.1beta	DALI Domain Definition
DDBASE	http://www.ncbs.res.in/~faculty/mini/ddbase/ddbase.html	DIAL

how it can be solved,[85] a full mathematical description of the problem, as well as a proof for one of its closed form solutions is given.

The problem of comparing two different models of a protein can be formalized as: given two sets of points $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots b_m)$ in three dimensional space and assuming that they have the same cardinality, i.e., n = m, and that the element a_i corresponds to the element b_i, find the optimal rigid body transformation G_{opt} between the two sets that minimizes a given distance metric D over all possible rigid body transformation G, as in Eq. [1]:

$$\min_G \{D(A - G(B))\} \quad [1]$$

When comparing two proteins, the sets of points can include the C_α only, all backbone atoms, or all atoms of the proteins. Different metrics have been used in the literature to determine the geometric similarity between sets of points. For protein superposition, the most common metric is the coordinate root mean square deviation (cRMS), which is defined as follows:

$$D(A, B) = cRMS(A, B) = \| A - B \| = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (a_i - b_i)^2} \quad [2]$$

A rigid body transformation is a transformation that does not produce changes in the size, shape, or topology of an object. Mathematically, it can be defined as a mapping $G: \Re^3 \rightarrow \Re^3$ that satisfies the properties:

$$\|G(x) - G(y)\| = \|x - y\| \text{ for all points } x \text{ and } y \quad [3]$$

and

$$G(x \wedge y) = G(x) \wedge G(y) \text{ for all vectors } x \text{ and } y \quad [4]$$

where $\wedge$ is the cross-product.

Equation [3] states that distances are conserved, whereas Eq. [4] says that internal reflection is not allowed. Rotations and translations are two examples of rigid body transformation, and in fact, a general rigid body transformation can be expressed as a combination of a rotation R and a translation T. The transformation problem can then be restated as finding the optimal rotation R and optimal translation T such that $\| A - RB - T \|$ is a minimum.

A Closed-Form Solution Based on Singular Value Decomposition

Many algorithms exist in the literature that solve the rigid transposition problem, coming from various fields including computer vision and image processing, robotics, astronomy, and computational biology. Those algorithms

differ with respect to the representation of the transformation and the minimization procedure. Some algorithms are based on closed-form solutions, whereas others use iterative solutions. For detailed descriptions of these algorithms, including comparison of their performances, the reader is referred to the surveys of Sabata and Aggarwal,[86] Ferrari and Guerra,[87] and Eggert et al.[88] Here a focus is placed on the representation typically used in computational biology. It is based on the singular-value decomposition (SVD)[89] of a correlation matrix C between the two sets of points.[90–93] This method seems to have been first derived by Schonenman in the context of factor analysis.[94] Other approaches include solutions based on a power decomposition of C[95] or on a representation of rotations with quaternions.[96–98] These methods have been shown to be equivalent.[88,98]

Using the definition of the metric given in Eq. [2], the rigid transformation problem can be restated as finding the rotation R_{min} and the translation T_{min} such that

$$\varepsilon = \frac{1}{n}\sum_{i=1}^{n}(a_i - Rb_i - T)^2 \qquad [5]$$

is minimum.

Considering variations with respect to T first, we find that for an extremum of ε,

$$\frac{\partial \varepsilon}{\partial T} = -\frac{2}{n}\sum_{i=1}^{n}(a_i - Rb_i - T) = 0 \qquad [6]$$

so that

$$T_{min} = \frac{1}{n}\sum_{i=1}^{n} a_i - R_{min}\left(\frac{1}{n}\sum_{i=1}^{n} b_i\right) = \mu_A - R_{min}\mu_B \qquad [7]$$

where μ_A and μ_B are the centers of mass of A and B, respectively.

Note that if the two sets of points are shifted such that their centers of mass coincide at the origin, $T_{min} = 0$. Let $x_i = a_i - \mu_A$ and $y_i = b_i - \mu_B$ be the coordinates of the shifted points, and let $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \ldots y_n]$ be the $3 \times n$ matrices representing the two sets of points A and B, after shifting. The rigid-body transformation problem can then be restated as finding the optimal rotation matrix R_{min} such that

$$\varepsilon = \frac{1}{n}\| X - RY \|^2 \qquad [8]$$

is minimum.

Let C be the correlation matrix of X and Y:

$$C = XY^T \rightarrow C_{ij} = \sum_{k=1}^{n} x_{ik} y_{jk}, i, j = 1, 2, 3 \quad [9]$$

and UDV^T be an SVD[89] of C ($UU^T = VV^T = I$, $D = \mathrm{diag}(d_i)$, $d_1 \geq d_2 \geq d_3 \geq 0$). The minimum value of ε with respect to R is then

$$\varepsilon_{min} = \frac{1}{n}\left(\|X\|^2 + \|Y\|^2 - 2(d_1 + d_2 + \lambda d_3)\right) \quad [10]$$

where $\lambda = \mathrm{sign}(\det(C))$. The optimal rotation is given by

$$R_{\min} = U \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \lambda \end{pmatrix} V^T \quad [11]$$

when $\mathrm{rank}(C) \geq 2$.

This result was first formulated by Schonemann,[94] later refined by Arun et al.,[93] Horn et al.,[95] and Umeyama.[99] Here the proof of Umeyama is given. Finding a rotation matrix R that minimizes ε can be rewritten as finding a matrix R that minimizes the objective function O defined as

$$O = \|X - RY\|^2 + tr\left(L\left(R^T R - I\right)\right) + g(\det(R) - 1) \quad [12]$$

where g is a Lagrange multiplier and L is a symmetric matrix of Lagrange multipliers. The second and third term of O represent the conditions for R to be an orthogonal and proper rotation matrix, respectively. Partial differentiations of O with respect to R, L, and g lead to the following system of equations:[99]

$$\frac{\partial O}{\partial R} = -2XY^T + 2RYY^T + 2RL + gR = 0 \quad [13]$$

$$\frac{\partial O}{\partial L} = R^T R - I = 0 \quad [14]$$

$$\frac{\partial O}{\partial g} = \det(R) - 1 = 0 \quad [15]$$

From Eq. [13],

$$RM = XY^T = C \quad [16]$$

where C is the covariance matrix defined in Eq. [9] and M is a symmetric 3×3 matrix defined by

$$M = YY^T + L + \frac{g}{2}I \qquad [17]$$

Transposing Eq. [16], we obtain

$$MR^T = C^T \qquad [18]$$

and multiplying each side of [16] with each side of [18], Eq. [19] is obtained, as $R^TR = I$ (Eq. [14]):

$$M^2 = C^TC = VD^2V^T \qquad [19]$$

Because M and M^2 are commutative ($MM^2 = M^2M$), both can be reduced to diagonal form by the same orthogonal matrix. Thus,

$$M = VDSV^T \qquad [20]$$

where $S = \text{diag}(s_i)$, $s_i = 1$ or -1.

From Eq. [20],

$$\det(M) = \det\left(VDSV^T\right) = \det(D)\det(S) \qquad [21]$$

and from Eq. [16]

$$\det(M) = \det(R^T)\det(C) = \det(C) \qquad [22]$$

as $\det(R) = \det(R^T) = 1$ (Eq. [15]).

Thus,

$$\det(D)\det(S) = \det(C) \qquad [23]$$

Singular values are non-negative, $\det(D) = d_1d_2d_3 \geq 0$; hence, $\det(S)$ must be equal to 1 if $\det(C) > 0$ and -1 if $\det(C) < 0$.

From the properties of norm and trace of a matrix, we get

$$\begin{aligned}\varepsilon &= \frac{1}{n}tr\left((X - RY)(X - RY)^T\right) = \frac{1}{n}\left(tr\left(XX^T\right) + tr\left((RY)(RY)^T\right) - 2tr\left(XY^TR^T\right)\right) \\ &= \frac{1}{n}\left(\|X\|^2 + \|RY\|^2 - 2tr\left(XY^TR^T\right)\right) = \frac{1}{n}\left(\|X\|^2 + \|Y\|^2 - 2tr(M)\right)\end{aligned} \qquad [24]$$

Substituting Eq. [20] into Eq. [24], we have

$$\begin{aligned}\varepsilon &= \frac{1}{n}\left(\|X\|^2 + \|Y\|^2 - 2tr\left(VDSV^T\right)\right) \\ &= \frac{1}{n}\left(\|X\|^2 + \|Y\|^2 - 2tr(DS)\right) \\ &= \frac{1}{n}\left(\|X\|^2 + \|Y\|^2 - 2(d_1s_1 + d_2s_2 + d_3s_3)\right)\end{aligned} \quad [25]$$

Thus, the minimum value of ε is achieved when $s_1 = s_2 = s_3 = 1$ if $\det(C) > 0$, and $s_1 = s_2 = 1$, $s_3 = -1$ if $\det(C) < 0$.

Next, we determine a rotation matrix R achieving the above minimum value. When $\text{rank}(C) = 3$, M is nonsingular and its inverse is given by

$$M^{-1} = (VDSV^T)^{-1} = VSD^{-1}V^T = VD^{-1}SV^T \quad [26]$$

and

$$R_{min} = CM^{-1} = UDV^TVD^{-1}SV^T = USV^T \quad [27]$$

which completes the proof for Eq. [10]. Note that this expression for R_{min} is also valid when $\text{rank}(C) = 2$ (see Ref. 99).

Weighted Superposition of Sets of Points

It is not always proper to assign the same importance to all points of A and B. Thus, variants of the rigid-body transformation problem have been developed in which each point i is given a weight ω_i. Examples of weighting schemes include (1) considering the mass of the atoms included in the superposition, (2) giving different weights to atoms of the backbone of the protein compared with atoms of the side chains, and (3) giving greater weights to atoms belonging to secondary structures of the protein. Solving the weighted variant of the rigid-body transformation problem amounts to finding the optimal translation T and optimal rotation R such that Eq. [28] is a minimum.

$$\varepsilon' = \frac{1}{n}\sum_{i=1}^{n}\omega_i(a_i - Rb_i - T)^2 \quad [28]$$

Considering variations with respect to T first, we find that for an extremum of ε',

$$\frac{\partial\varepsilon'}{\partial T} = -\frac{2}{n}\sum_{i=1}^{n}\omega_i(a_i - Rb_i - T) = 0 \quad [29]$$

so that

$$T_{min} = \frac{1}{\Omega}\sum_{i=1}^{n} \omega_i a_i - R_{min}\left(\frac{1}{\Omega}\sum_{i=1}^{n} \omega_i b_i\right) = \mu'_A - R_{min}\mu'_B \qquad [30]$$

where Ω is the sum of the weights ($\Omega = \sum_{i=1}^{n} \omega_i$), and μ'_A and μ'_B are the weighted centers of mass of A and B, respectively.

Note again that if the two sets of points are shifted such that their weighted barycenters coincide at the origin, $T_{min} = 0$. Let $x'_i = \sqrt{\omega_i}(a_i - \mu'_A)$ and $y'_i = \sqrt{\omega_i}(b_i - \mu'_B)$ be the weighted coordinates of the shifted points, and let $X' = [x'_1, x'_2, \ldots, x'_n]$ and $Y' = [y'_1, y'_2, \ldots y'_n]$ be the $3 \times n$ matrices representing the two weighted sets of points A and B, after shifting. The rigid-body transformation problem can then be restated as finding the optimal rotation matrix R_{min} such that

$$\varepsilon' = \frac{1}{n}\| X' - RY'\|^2 \qquad [31]$$

is minimum.

Equation [31] is equivalent to Eq. [8], and the same algorithm is used to solve it.

A General Algorithm for Point Set Superposition

The general procedure for superposing two protein structures, when the equivalent atoms are known, can then be summarized as follows:

1. Set input points $A = (a_1, a_2, \ldots a_n)$ for protein 1, $B = (b_1, b_2, \ldots, b_n)$ for protein 2, and weights $(\omega_1, \omega_2, \ldots, \omega_n)$.
2. Compute weighted centers of mass of A and B:

$$\mu'_A = \frac{\sum_{i=1}^{n} \omega_i a_i}{\sum_{i=1}^{n} \omega_i}; \quad \mu'_B = \frac{\sum_{i=1}^{n} \omega_i b_i}{\sum_{i=1}^{n} \omega_i}$$

3. Generate the weighted covariance matrix:

$$C'_{ij} = \sum_{k=1}^{n} \omega_k (a_{ki} - \mu'_{Ai})(b_{kj} - \mu'_{Aj}), \quad i = 1,2,3; \quad j = 1,2,3$$

4. Compute the SVD of C': $C' = UDV^T$ and $\lambda = \text{sign}(\det(C'))$, noting that $D = \text{diag}(d_1,d_2,d_3)$ with $d_1 \geq d_2 \geq d_3 \geq 0$.

5. Define optimal rotation $R_{min} = USV^T$, with $S = \text{diag}(1,1,\lambda)$, and optimal translation

$$T_{min} = \mu'_A - R_{min}\ \mu'_B$$

6. Compute the cRMS between the two structures:

$$cRMS = \sqrt{\varepsilon'} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} \omega_i (a_i - \mu'_A)^2 + \sum_{i=1}^{n} \omega_i (b_i - \mu'_B)^2 - 2(d_1 + d_2 + \lambda d_3)\right)}$$

This algorithm does not take into account the possibility of "noise" in the coordinates of the points. For proteins, the coordinates of atoms are approximations to a "true" position: Proteins are flexible, with fluctuation about a mean position. Moreover, the physical experiments that provide information on the coordinates (usually X-ray crystallography and NMR spectroscopy) have a degree of experimental uncertainty. When superposing two models for the structure of one protein, the cRMS value is therefore a combination of the actual fluctuation between the two models and of the noise level contained in the two models. Noise is even more important for the superposition of two proteins of different lengths.

Protein Structure Superposition

An Ambiguous Problem

The problem of finding an optimal alignment between two proteins is more complex than just solving the rigid-body transformation problem, because the correspondence, i.e., the list of equivalent residues in the two proteins, is often not known. Indeed, the correspondence is part of the desired output, along with the optimal transformation of the position of one protein with respect to the other. The protein structure alignment problem can be stated as finding the maximal substructures of the two proteins that exhibit the highest degree of similarity.

A "substructure" of protein A is a subset of its points, arranged by order of appearance in A. We denote the substructure defined by $P = (p_1, p_2, \ldots, p_k)$, where $1 \leq p_1 < p_2 \ldots < p_k \leq n$, by $A(P) = (a_{p1}, a_{p2}, \ldots, a_{pk})$. The length $|A(P)|$ of A(P) is the number of points it contains, which in this case is k. A "gap" in A(P) is two consecutive indices p_i, p_{i+1} such that $p_i + 1 < p_{i+1}$.

Given two sets of points $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots b_m)$ in three-dimensional (3-D) space, the protein superposition problem is to find the optimal subsets $A(P)$ and $B(Q)$ with $|A(P)| = |B(Q)|$, and the optimal rigid-body transformation G_{opt} between the two subsets $A(P)$ and $B(Q)$ that

minimizes a given distance metric D over all possible rigid-body transformation G, i.e.,

$$\min_G\{D(A(P) - G(B(Q)))\} \quad [32]$$

The two subsets $A(P)$ and $B(Q)$ define a "correspondence," and $p = |A(P)| = |B(Q)|$ is called the correspondence length. Once the optimal correspondence is defined, it is easy to find the optimal rotation and translation using the rigid-body transformation algorithm described earlier. The concept of "optimal correspondence," however, requires more explanation. It is clear that $p = 1$ defines a trivial solution to the protein superposition problem: Any point of A can be aligned with any point of B, with a cRMS of 0. In practice, we are interested in finding the largest possible value for p under the condition that $A(P)$ and $B(Q)$ remain "similar."

A fast, reliable, and convergent method for protein structural alignment is not yet available although significant progress toward this goal has been made over the past decade.[100] Recent developments have focused on both the search algorithm and the definition of the target function to be minimized, which in turn provides a quantitative measure of the "similarity" between two structures. The most direct approach for comparing two protein structures is to move the set of points representing one structure as a rigid body over the set of points of the other structure and look for equivalent residues. It can be achieved only for relatively similar structures and will fail to detect local similarities of structures sharing common substructures. To avoid this problem, the structures can be decomposed into fragments [usually secondary structure elements (SSEs)], but this can lead to situations in which the global alignment is missed. Accordingly, recent work has focused on combining the local and global criteria in a hierarchical and heuristic approach. These methods proceed by first defining a list of equivalent positions in the two structures, from which a structural alignment can be derived. This initial equivalence set is defined by various methods, including dynamic programming,[101,102] comparing distance matrices,[8,103–105] fragment matching,[106,107] geometric hashing,[108–113] maximal common subgraph detection,[114–116] and local geometry matching.[117] Optimization of this equivalence set has been performed with dynamic programming,[102,118–120] Monte Carlo algorithms or simulated annealing,[8] a genetic algorithm,[121] incremental combinatorial extension of the optimal path,[122,123] and mean-field approaches.[124,125] Excellent reviews of these and other methods can be found in Refs. 13, 100, 126, and 127. Many groups involved in developing algorithms for protein structure alignment have generously made their programs available for use over the Internet and the World Wide Web. In some cases, the program is accessible for download, either as an executable or as a full source package (Table 5). These are

Table 5 Websites for Publicly Available Protein Structure Alignment Services and Programs

Program	Web Access (Interface)	Web Access (Program Download)	Method
CE	http://cl.sdsc.edu	ftp://ftp.sdsc.edu/pub/sdsc/biology/CE/src	Extension of the optimal path
DALILIGHT	http://www2.ebi.ac.uk/dali	http://ekhidna.biocenter.helsinki.fi:8080/dali/DaliLite/index.html	Distance matrix alignment
DEJAVU	http://portray.bmc.uu.se/cgi-bin/dejavu/scripts/dejavu.pl		Compare SSE[a]
FATCAT	http://fatcat.burnham.org/fatcatpair.html		Flexible structure alignment based on fragments
FoldMiner	http://dlb4.stanford.edu/FoldMiner/		Structure-database comparison based on motif search
K2 and K2SA	http://zlab.bu.edu/k2		Genetic algorithm (K2) or Simulated annealing (K2SA)
LOCK2	http://motif.stanford.edu/lock2/		Hierarchical protein structure superposition
LSQRMS	http://www.molmovdb.org/align/		STRUCTAL-based program
MATRAS	http://biunit.aist-nara.ac.jp/matras/		Markov transition model of evolution
PRIDE	http://hydra.icgeb.trieste.it/pride/		Probabilistic approach based on CA-CA distance matrix
PRISM	None	http://honiglab.cpmc.columbia.edu/	SSE alignment followed by iterative refinement of the equivalence list
PROSUP	http://lore.came.sbg.ac.at:8080/CAME/CAME_EXTERN/PROSUP		Hierarchical alignment
SARF2	http://123d.ncifcrf.gov/sarf2.html	http://123d.ncifcrf.gov/sarf2.html	Alignment of backbone fragments

Table 5 (*Continued*)

Program	Web Access (Interface)	Web Access (Program Download)	Method
SHEBA	http://rex.nci.nih.gov/RESEARCH/basic/lmb/mms/sheba.html	http://rex.nci.nih.gov/RESEARCH/basic/lmb/mms/SHEBA-download.html	Hierarchical alignment including profiles
SSAP	http://www.biochem.ucl.ac.uk/cgi-bin/cath/GetSsapRasmol.pl		Double dynamic program
SSM	http://www.ebi.ac.uk/msd-srv/ssm/	http://www.ebi.ac.uk/msd-srv/ssm/cgi-bin/ssmdcenter	Secondary Structure Matching
TOPS	http://balabio.dcs.gla.ac.uk/tops/versus.html	http://www.tops.leeds.ac.uk/	Alignment of simplified representations of proteins
TOPSCAN	http://www.bioinf.org.uk/topscan		Fast alignment based on SSE matching
VAST	http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html		Vector alignment

[a]SSE: secondary structure elements

wonderful tools, and the reader is encouraged to test several of these sites. Many of these services have been tested on large datasets with known similarities.[127–129] Those comparison studies do not identify a clear "winner," i.e., a technique that is significantly better than another; apparently an approach that combines existing algorithms performs better than any individual technique alone.[129] The different definitions given to the similarities of two structures are reviewed below, and two methods for protein structure alignment are described, one based on distance matrices (DALI),[8,130] and one based on dynamic programming and comparison of structures in coordinate space (STRUCTAL).[118,119] Recent progress in developing a closed-form protein structure alignment algorithm is then described.

Scoring Functions for Protein Structure Superposition

Because the concept of "optimal" correspondence is ambiguous, the protein structure superposition problem is not uniquely defined. Instead, finding the best superposition of two proteins corresponds to a family of optimization problems, which are specified by the weight given to the similarity (preferably a small deviation between the two subsets), and the correspondence length (preferably large).

Various measures of similarity between two sets of points exist. In the section on rigid-body transformation, the cRMS value, which measures

the root mean square deviation between the coordinates of the points of the two sets, was described. For a given correspondence length, the cRMS can be minimized with a closed-form algorithm (see above). When both the cRMS and the correspondence length must be optimized, no closed-form solutions are known. Approximate solutions that are usually based on heuristics do not minimize the cRMS directly, as a consequence of being very sensitive to outliers (the definition of cRMS, given in Eq. [2], shows that it is a sum of the Euclidian distances between the points of the two structures; if a few of these points are far apart, they will have a significant impact on the value of cRMS). An example of a solution to this outlier problem was given by Levitt et al.[118,119] who introduced a scoring function with a Lorentzian shape:

$$ST(P,Q) = \max_{R,T} \sum_{i=1}^{p} \frac{20}{1 + \| a_{pi} - Rb_{qi} - T \|^2} - 10G_{P,Q} \qquad [33]$$

The summation extends over the length of the correspondence between *A(P)* and $B(Q)$, $G_{P,Q}$ is the total number of gaps in $A(P)$ and $B(Q)$, and R and T are the optimal rotation and translation that maximize the score (as opposed to reaching a minimum for cRMS, as seen in Eq. [1]).

An alternative measure of protein structure similarity is the distance root mean squared deviation (dRMS) that compares corresponding internal distances in the two sets of points:

$$dRMS = \left[\frac{2}{p(p-1)} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} M\big(\| a_i - a_j \|, \| b_i - b_j \|\big) \right]^{1/2} \qquad [34]$$

where p is the cardinality of the two sets.

There is no consensus in the scientific community on the definition of the metric M that should be used to compare the two internal distances $\|a_i - a_j\|$ and $\|b_i - b_j\|$. When comparing two pairs of atoms between two structures, Taylor and Orengo[101] defined a distance or similarity score in the form $e/(D+f)$, where D is the difference between the two intramolecular distances and e and f are arbitrarily defined constant values. Holm and Sander[8] defined a similarity score as $(e - [D/\langle D \rangle])\exp(-[\langle D \rangle / f]^2)$, where $\langle D \rangle$ is the average of the two intramolecular distances. Rossmann and Argos[131] and Russell and Barton[132] used a score defined as $\exp(-[D/e]2)\exp(-[S/e]2)$, where S takes into account local neighbors for each pair of atoms. Currently, no clear evidence exists as to which score performs best.

All scoring functions use geometry for the comparison and ignore similarities in the environment of the residues. Suyama et al.[133] proposed another approach in which they ignored the 3-D geometry altogether and compared

structures on the basis of their 3-D profiles[134] alone, using dynamic programming. These profiles include information on solvent accessibility, hydrogen bonds, local secondary structure states, and side-chain packing. Although this method can align two-domain proteins with different relative orientations of the two domains, it often generates inaccurate alignments.[133] Jung and Lee[135] improved on this method by iteratively refining the initial profile alignment with dynamic programming and 3-D superposition. Their method, which they call SHEBA, was found to be fast and as reliable as other alignment techniques (although it was only tested on a small number of protein pairs). Kawabata and Nishikawa[136] derived a novel scoring scheme for generating structural alignments based on the Markov transition model of evolution. The similarity score between two structures i and j is defined as $\log(P(ji)/P(i))$, where $P(ji)$ is the probability that structure j changes to structure i during evolution and $P(i)$ is the probability that structure i appears by chance. The probabilities are estimated with a Markov transition model that is equivalent to the Dayhoff's substitution model for amino acids. Three types of scores were considered: (1) a score based on accessibility to solvent; (2) a residue-residue distance score; and (3) an SSE score.

Superposition Based on Internal Distance Matrices: DALI

Associated with every protein chain A of n atoms is an $n \times n$ real symmetric matrix D, where $D(i, j)$ is the Euclidian distance between atoms i and j of A. This matrix is the "internal distances matrix" of A, also called the distance map of A. The two representations of a protein (by the coordinates of its atoms and by its internal distances matrix) are closely related. Calculating the distance matrix from the coordinates is easy and takes quadratic time in n. Conversely, it is known that the atomic coordinates of a protein can be recovered from the distances matrix, using the method of distance geometry.[137,138] The recovered atomic coordinates are the original ones, modulo a rigid transformation (and possibly a mirror transformation). This equivalence between coordinates and internal distances has led to two different measures of protein similarities, each based on one of the two representations. The use of the internal distances to compare protein structures has a major advantage of by passing the need to find an optimal rigid transformation that superposes the two structures. As a consequence, most algorithms have been developed that compare the internal distances matrix to align protein structures, the most popular being the Distance ALIgnment algorithm (DALI), which is briefly described below.

Holm and Sander[8] developed a two-stage procedure in DALI that uses simulated annealing to build an alignment of similar hexapeptide backbone fragments between two proteins.

In the first stage, the two protein structures to be compared are divided into overlapping hexapeptides. A contact map, which contains all internal distances, is generated for each hexapeptide. Although residues in the proteins

belong to several overlapping hexapeptides, they are assigned to the hexapeptide with the closest contacts to other fragments. Contact maps of the two proteins are matched by comparing their internal distances, using an "elastic" score of the form $(e - [D/\langle D\rangle])\exp(-[\langle D\rangle/f]2)$, where D is the difference between the two distances to be compared, $\langle D\rangle$ is the average distance, and e and f are parameters. This score is less sensitive to distortion for long-range distances. For the sake of efficiency, only hexapeptide pairs having a similar backbone conformation are compared. Hexapeptides whose contact maps match above a given threshold are stored in lists of fragment equivalences.

In the second stage, an optimization protocol, based on simulated annealing, explores different concatenations of the equivalent hexapeptide pairs. Similarity is assessed by comparing all distances between the aligned substructures. Each step of this second stage consists of addition, replacement, or deletion of residue equivalences (in units of hexapeptides). Because hexapeptides can overlap, each step results in the addition of one to six residues. Once all candidate hexapeptide pairs have been tested, the alignment is processed to remove fragments with a negative contribution to the overall similarity score.

This two-stage method, implemented in DALI,[130] has been used to compare representatives from all nonhomologous (in sequence) families in the protein data bank.[3,4] More details are given in the DALI Domain Classification below.

Superposition Based on cRMS: STRUCTAL

The internal distances matrix is invariant under rigid and mirror transformations of the protein. Although this lead to a simplification of the protein structure superposition problem because algorithms that compare proteins based on internal distances do not need to find the optimal rigid transformation, mirror transformation may introduce errors because mirror images (such as a right-handed helix versus a left-handed helix) will not be detected as being different. Consequently, in addition to the approaches based on the internal distances matrix, methods have been concurrently developed to solve the protein structure alignment problem using coordinates to measure the similarity between proteins. These methods are based on heuristic algorithms that optimize simultaneously the correspondence between the two proteins and the rigid transformation. An example is the algorithm developed by Subbiah et al. and implemented in the program STRUCTAL.[118]

STRUCTAL starts with an arbitrary equivalence of atoms between the two proteins A and B. This equivalence defines a list of corresponding residues (represented by their C_α atoms) that are superimposed with the optimal rigid-body transformation. Once the two proteins are superimposed, the program computes a structure alignment matrix *SA*. $SA(i,j)$ measures the similarity between residue i of protein A and residue j of protein B, based on a function

of the distance d between $C\alpha_i$ and $C\alpha_j$, after optimal superposition. This function is defined such that

$$SA(i,j) = \frac{20}{1 + d(C\alpha_i, C\alpha_j)^2/5} \quad [35]$$

It is simple to compute, and it has the important properties of being positive and of decreasing monotonically with increasing distances. A new alignment is then determined by searching the distance matrix for the alignment with the best score. Dynamic programming rapidly [$O(n^2)$ operations] finds the optimum for the given structure alignment matrix and for a given gap penalty. In STRUCTAL, the gap penalty is set constant, equal to 10. The new alignment leads, in turn, to a new set of equivalencies between the proteins; this set is then used to re-superimpose the two proteins in three dimensions giving rise to a new structure alignment matrix. The procedure is iterated until the alignment matrix does not change with additional iterations.

Because this structural alignment procedure is based on dynamic programming and is iterative, the results depend on the initial equivalence assigned. To account for this potential problem, STRUCTAL starts with five different equivalences. The first three equivalences are simple, corresponding to aligning the chain beginnings, the chain ends, and the chain midpoints of the two structures, respectively, without allowing any gaps. The fourth choice maximizes the sequence identity of pairs of residues considered equivalent, whereas the fifth choice is based on the similarity of the C_α torsion angles between the two chains. After repeating the iterative scheme to find the optimal equivalence and superposition for each of the five initial sets of equivalences, the optimal alignment is chosen as the one with the highest score. Extensive studies have shown that not one of the five initial sets works better than another in converging to the highest score.[119]

An Approximate Polynomial Time Algorithm

A prevailing sentiment among scientists developing algorithms for protein structure alignment is that structure comparison requires exponential computer resources, and so development should focus on heuristic approaches. Consequently, there are no guarantees of finding an optimal alignment with respect to any scoring function with any of the existing methods. In addition, if one of these methods fails to find a satisfactory alignment, it cannot be ruled out that a good alignment exists. There is one interesting, albeit theoretical, exception to this. Kolodny and Linial[139] have developed a polynomial-time algorithm that optimizes simultaneously the correspondence and the rigid transformation leading to a structural alignment. The computation cost of their algorithm is of the order $O(n^{10})$, and as such, it is not practical to implement. This algorithm, however, is not heuristic: it guarantees finding ε-approximations to all solutions of the protein superposition problem, where

these solutions correspond to maxima of the STRUCTAL score *ST* defined in Eq. [35].

For an alignment algorithm to be polynomial, the following conditions must hold:

1. Given a rigid transformation, it should be possible to find an optimal correspondence in polynomial time.
2. The number of rigid transformations under consideration must be bounded by a polynomial.

The STRUCTAL score *ST* is amenable to dynamic programming and therefore can be used to find an optimal correspondence in time and space requirements of order $O(n^2)$, for any given rigid transformation r. The score of this optimal correspondence is denoted as $ST_{opt}(r)$. It validates condition 1. The validity of condition 2 is derived from a lemma given by Kolodny and Linial, which states that for all ε, a finite set $G = G(\varepsilon)$ of rigid transformation exists, such that for every choice of a rigid transformation r, a transformation r_G in $G(\varepsilon)$ exists such that $\| ST_{opt}(r) - ST_{opt}(rG) \| < \varepsilon$, and cardinal$(G) = |G|$ is polynomial in n.

This lemma suggests the following algorithm for the structural alignment problem. For a given value of ε, build $G(\varepsilon)$, the discrete sampling of the space of rigid transformation, and evaluate ST_{opt} over all rigid transformations in $G(\varepsilon)$. The ε-optimal structure alignments of the two proteins are guaranteed to be within ε of the maxima found in the exhaustive search over $G(\varepsilon)$. A major advantage of this exhaustive algorithm is that if it fails to find a good alignment, no good alignment exists. Because the size of $G(\varepsilon)$ is of the order $O(n^{10}/\varepsilon^6)$, the computing time required by this algorithm is still prohibitive. As such, the contribution of Kolodny and Linial should be viewed as mostly theoretical, rather than as practical, but it does provide insights about the complexity of protein structure alignments.

cRMS: An Ambiguous Measure of Similarity

The root mean square deviations (cRMS or dRMS) remain the measures of choice by structural biologists to describe the similarity between two proteins, even though most algorithms for protein structure alignments use scoring schemes that differ significantly from simply taking into account interatomic distances (see above). Both cRMS and dRMS are based on the L_2-norm (i.e., the Euclidian norm), and as such, they suffer from the same drawback as the residual χ^2 in least-squares minimization in which the presence of outliers introduces a bias in the search for an optimal fit and the final measure of the quality of the fit may be artificially poor because of the sole presence of those outliers. Another problem of relying on RMS deviations is that it does not always satisfy the triangular inequality. More precisely, the triangular inequality is satisfied when the correspondences between the

proteins always involve the same points.[140] Generally by varying correspondences, it is easy to find a situation in which the triangular inequality is not satisfied. Consider, for example, two proteins A and B that are dissimilar, and the two-domain protein C, whose subdomains C1 and C2 are very similar to A and B, respectively. In this example, the RMS deviations between A and C and between B and C are low, but the RMS deviation between A and B is large, violating the triangular inequality that states that RMS(A, B) $\leq$ RMS(A, C) + RMS(C, B). As a consequence of these limitations, RMS is a useful measure of structural similarity for only closely related proteins.[141] Several other measures have therefore been proposed to circumvent these problems. The STRUCTAL score S2 (Eq. [34]) was shown to be a more reliable indicator of structure similarity than RMS because it depends on the best-fitting pairs of atoms (thereby removing the weights of outliers). In contrast, RMS gives equal weight to all pairs of atoms. Lesk[142] recently proposed replacing the L_2-norm in the RMS definition by the L norm, also called the Chebyshev norm, to yield the new score:

$$S = \max_{i \in [1,N]} \{ \| x(i) - y(i) \| \} \qquad [36]$$

S reports the worst-fitting pair of atoms (after optimal superposition of the two structures) and, as such, is even more sensitive to outliers than is the RMS. Yang and Honig[120] defined a new protein structure similarity measure, the protein structural distance (PSD). PSD combines a secondary structural alignment score and the RMS deviation of topologically equivalent residue pairs. It thus incorporates the resolution power of both RMS for closely related structures and the secondary structure score for proteins that can be very different. By analyzing the PSD scores obtained from more than one and a half million pairs of proteins, Yang and Honig[120] proposed that a continuum of protein conformation space exists, conflicting with existing ideology in structural classification databases such as SCOP (Structural Classification Of Proteins[6]) and CATH (Class, Architecture, Topology and Homologous Superfamilies[7]). May[143] assessed 37 different protein structure similarity measures for their ability to generate accurate clusters in a hierarchical classification of 24 protein families. It was found that the sum of ranks of distances at aligned positions was a better measure of similarity than the direct sum of distances and that RMS computed over the subset of core-aligned positions performs better than normal RMS computed over all-aligned positions. Variations in the hierarchical classification of protein structures brings into question the validity not only of the measure used for the clustering, but also of the hierarchical clustering. The difficulty associated with defining a similarity score for protein structures reflects the fact that most questions related to structure comparison do not have a unique answer[144–146] and brings to the fore that the problem is ill posed, requiring additional information to provide a well-defined solution. As

an example, consider fold recognition applications, where predictors focus on the well-conserved core region of the protein and pay less attention to the loop geometry. In such cases, it makes sense to define a similarity score that includes only atoms in the core.

Having a quantitative measure of protein structure similarities is essential when assessing the quality of protein structure predictions, such as those generated for the Critical Assessment of techniques for protein Structure Prediction (CASP) project, organized in the form of a meeting held in alternating years at Asilomar, California. For the special case of comparing a predicted structure with its experimental counterpart, the equivalence list is known because the two sequences are identical, which thus reduces the complexity of the problem. On the other hand, each structure prediction may omit different residues depending on how the prediction algorithm works and different parts of the predicted structure may omit geometries of variable quality. Hubbard[147] avoided the problem by generating many superpositions and by calculating the best RMS for each set of equivalent residues (not necessarily contiguous), to provide an RMS/coverage graph, which evaluated predictions at CASP3. The RMS/coverage plot can also be interpreted as defining the number of equivalent residues for a given RMS value.

Differential Geometry and Protein Structure Comparison

The inherent problems of RMS as a measure of protein structure similarities, and the difficulties encountered by the existing heuristic algorithms whose aim is to solve the protein structure superposition problem, have led to the development of a new approach for comparing protein structure, based on differential geometry and the concept of protein shape descriptors. A tutorial on molecular shape descriptors can be found in a previous volume of this series.[148] The idea behind this approach is simple: Represent the protein structure with a vector of geometric properties (GPs) such that the comparison of two protein structures is performed by comparing their GP vectors, usually with a Euclidian metric. Once the GP vectors have been computed, structure comparison with this scheme becomes instantaneous, and it can then be performed over entire databases. The success of this approach depends on the quality of the geometric properties included in GP and on their ability to uniquely capture the geometric properties of the protein. Because there is a growing interest to define such protein shape descriptors, two descriptors derived from knot theory, the writhe and the radius of curvature of a polygonal curve, are reviewed here.

The Writhe of a Protein Chain

The writhe of a polygonal curve is the signed average crossing number of the curve, where the average is taken over the observer's positions, located in all space directions.

Consider a polygonal curve A defined by N line segments i. The writhe of A is computed according to

$$Wr(A) = I_{1,2}(A) = \sum_{0<i_1<i_2<N} W(i_1, i_2) \quad [37]$$

with

$$W(i_1, i_2) = \frac{1}{2\pi} \int_{t_1=i_1}^{i_1+1} \int_{t_2=i_2}^{i_2+1} w(t_1, t_2) dt_1 \; dt_2 \quad [38]$$

where $W(i_1, i_2)$ is the contribution to the writhe of line segments i_1 and i_2. $W(i_1, i_2)$ is the probability of seeing the line segments cross when viewed from an arbitrary direction, multiplied by the sign of the crossing. Computation of $W(i_1,i_2)$ is described in Figure 8. Similarly, the unsigned average number of crossing, usually referred to as the average crossing number, is given by

$$I_{|1,2|}(A) = \sum_{0<i_1<i_2<N} |W(i_1, i_2)| \quad [39]$$

A whole family of structural measures can be envisaged with $W(i_1, i_2)$ and $|W(i_1, i_2)|$ as building blocks,[149] such as

$$I_{(1,3)|2,4|} = \sum_{0<i_1<i_2<i_3<i_4<N} W(i_1, i_3) |W(i_2, i_4)| \quad [40]$$

and

$$I_{(1,4),(2,6),(3,5)} = \sum_{0<i_1<i_2<i_3<i_4<i_5<i_6<N} W(i_1, i_4) W(i_2, i_6) W(i_3, i_5) \quad [41]$$

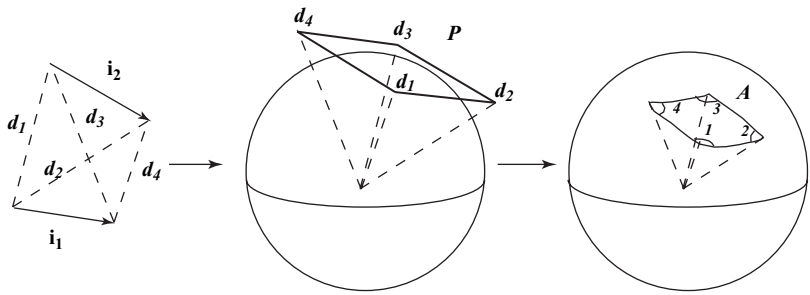


Figure 8 Computing $W(i_1, i_2)$, the writhe of two segments i_1 and i_2. (Left panel) The endpoints of the segments i_1 and i_2 are connected by 4 vectors d_1, d_2, d_3, and d_4 that generate a parallelogram P of directions (center panel). The area A of the projection of P on the surface of the unit sphere is the segment-segment writhe $W(i_1, i_2)$. A is computed as: $A = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) - 2\pi$ (right panel).

These measures are inspired by Vassiliev knot invariants.[150] They form a natural progression of curve descriptors, much as moments of inertia and their correlations define solids.

The writhe and the average crossing number have been used extensively to characterize DNA molecules and more specifically supercoiled DNAs.[151,152] They have also been used to describe proteins. Levitt[153] has used writhe to distinguish different chain threading. Arteca et al. used the writhe as a protein shape descriptor.[154–158] Rogen and Bohr[149] have used the writhe, the average crossing number, and their higher order correlations to define a feature vector that characterizes protein structures. More recently, Rogen and Fain[159] have compared protein structures with feature vectors in $\Re^{30}$ similar to those defined by Rogen and Bohr, using a pseudo-metric, which is the Euclidian distance between the feature vectors. That pseudo-metric is the scaled Gauss metric (SGM) (this name was chosen as the writhe of a continuous curve and is usually computed with a Gauss integral[160]). Rogen and Fain[159] show that SGM performs extremely well as a protein structure classifier, using both CATH and SCOP as test sets. Because both CATH and SCOP include all protein chains in the PDB, they are highly redundant and cannot be considered as discriminative benchmarks. Despite this reservation, the results of Rogen and Fain provide a new way to compare and classify protein structures, using geometric protein shape descriptors.

Protein Chain Thickness and Generalized Radius of Curvature

Any smooth, nonintersecting curve can be thickened to a smooth, nonintersecting tube of constant radius centered on the curve. If the curve is a straight line, there is no upper bound for the radius, but for any other curve, there is a critical radius above which the tube ceases to be smooth or shows self-contact. This critical radius is referred to as the thickness of the curve, and it is used as a shape descriptor in knot theory. Because the geometry of DNA and protein molecules is characterized by the geometry of their backbone chain, it is natural to check whether thickness could be used as a shape descriptor for these molecules. Gonzalez and Maddocks[161] introduced the concept of generalized radius of curvature and used it to characterize the geometry of DNA molecules. Their definition of generalized radius of curvature is based on the fact that any three noncollinear points x, y, z in 3-D space define a unique circle whose radius is given by

$$r(x,y,z) = \frac{|x-y||x-z||y-z|}{4A(x,y,z)} \qquad [42]$$

where $A(x,y,z)$ is the area of the triangle with vertices at x, y and z and $|x-y|$ is the Euclidian distance between x and y. Let us consider a discrete curve C,

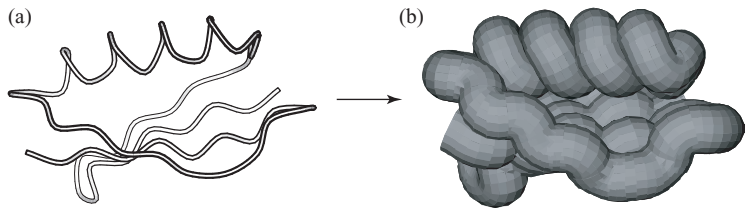


Figure 9 Thickness of a protein. (a) The structure of the B1 immunoglobulin-binding domain of streptococcal protein G is visualized as a thin tube. (b) View of the same tube inflated to its "thickness" (i.e., to a radius above which the tube ceases to be smooth, or shows self contact). Note that no free space exists between consecutive turns of the helices. Figure 9a drawn with MOLSCRIPT[17] and 9b with VMD.[18]

defined by n nodes (c_1, c_2, ... c_n). Gonzalez and Maddocks define the generalized radius of curvature of C at c_i by

$$\rho_C(c_i) = \min_{\substack{1\le j,k\le n \\ i\neq j\neq k\neq i}} r(c_i, c_j, c_k) \qquad [43]$$

$\rho_C(c_i)$ is the radius of the smallest circle passing by c_i and two other distinct nodes of C. $\rho_C(c_i)$ should be distinguished from the local radius of curvature ρ defined at c_i by $\rho(c_i) = \rho(c_{i-1},c_i,c_{i+1})$. The thickness $\Delta(C)$ of the discrete curve C is related to the generalized radius of curvature by

$$\Delta(C) = \min_{1\le i\le n} \rho_C(c_i) \qquad [44]$$

In other words, $\Delta(C)$ is the radius of the smallest circle passing by three points of C.

Figure 9 illustrates the "thickness" of a small globular protein. The concepts of thickness and generalized radius of curvature have been used to characterize the geometry of DNA.[151,152] They have also been used as a "potential" that captures the geometry of a protein,[162–166] which was used, for example, in protein structure prediction computer experiments.[167] Thickness and generalized radius of curvature have not yet been used for protein structure comparison, but it is expected that they would prove useful for detecting protein structure similarities, especially when combined with other features such as writhe.

Upcoming Challenges for Protein Structure Comparison

The most difficult application of protein structure comparison comes in classifying known protein structures into different clusters corresponding to fold families. The role of such classifications is to organize structure databases such as the PDB, in hopes of detecting similarities at the structure level that

cannot be detected at the sequence level, and more generally, to detect evolutionary relationships between proteins. The existing protein structure classification schemes are reviewed in the next section. Multiple challenges must be overcome by a protein structure comparison program in this application. First, it must deal with different levels of structural similarities, it must identify similarities even when those similarities form a small proportion of the proteins being compared, and it must handle insertions of arbitrary size as well as permutations of substructures. Second, it must deal with the fact there may be more than one acceptable solution for the structural alignment of two proteins. These multiple, equivalent solutions (in terms of cRMS and length of the equivalence) may all be viable from a biological perspective,[145] and therefore cannot be ignored. Third, the size of most protein structure databases has grown exponentially in the recent years, and the growth rate is expected to continue as the structural genomics projects enter their productive phases. A need exists for fast techniques to compare and classify these structures, faster than the existing techniques that are too time consuming to be of use.

None of the existing methods, including those described in length above, propose solutions to meet these challenges. Heuristic methods were developed for the sake of efficiency; yet no guarantee exists that they can find the optimal superposition. Also, some of these heuristic methods cannot detect alternative, equally acceptable solutions. The approximate solution developed by Kolodny and Linial[139] resolves some of these issues in the sense that it can detect all maximal solutions with an ε of the optimal solutions, but its computing cost (of the order of $O(n^{10}/\varepsilon^6)$ where n is the size of the proteins considered) makes it unsuitable for large-scale comparisons. There is a need to develop faster, more robust, and exhaustive approaches to solving the myriad of problems associated with protein structure comparison. This field in fact remains an active area of development in structural biology, but solutions may come from interdisciplinary research groups. The problem of comparing two protein structures can be reformalized as the problem of comparing two sets of points in 3-D space. As such, it can be seen as a problem of computational geometry, and it is expected that collaboration between structural biologists well versed in deciphering protein structures and computer scientists who focus on geometric problems could provide the synergy required for significant progress. The recent advances in the application of differential geometry to protein structure (see the sections on writhe and curve thickness above) are signs that these collaborative efforts are working.

PROTEIN STRUCTURE CLASSIFICATION

Perutz et al.[30] showed in 1960 that myoglobin and hemoglobin, the first two protein structures to be solved at atomic resolution using X-ray crystallography, have similar structures even though their sequences differ. These two

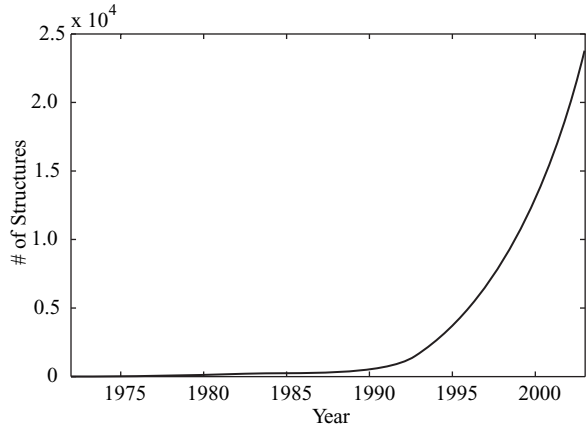


Figure 10 Statistics on the PDB. The number of structures (proteins and nucleic acids) available in the Protein Data Bank (PDB)[3,4] is plotted against time, starting from 1973 when the PDB was created.

proteins are functionally similar, as they are involved with the storage and the transport of oxygen, respectively. Since then, there has been a continued interest in finding structural similarities between proteins, in the hope of revealing shared functionality that could not be detected by sequence information alone. A logical consequence of this interest is the development of systems for classification of protein structures that identify and group proteins sharing the same structure so as to reveal evolutionary relationships. Classifying protein structures has now become essential because of the volume of structural data available (see Figure 10). In parallel with the development of protein structure classification methods are the developments for many classifications of protein sequences described in length in Ref. 168. Table 6 lists some resources available for sequence classification.

All current structural classification methods are based on the same scheme: Protein structures are first divided into discrete, globular domains, which are then classified at the levels of (1) ''class,'' (2) ''folds,'' (3) ''superfamilies,'' and (4) ''families.'' The differences among existing schemes come from the methods that define the domains and the procedures that classify. After reviewing the terms that define a classification, the three main protein structure classifications available, SCOP, CATH, and the DALI Domain Dictionary (DDD), will be described. Links to these databases and related services are listed in Table 7.

The first complication associated with structure classification involves the fact that protein structures are often composed of distinct globular domains. Because these domains can function individually, with distinct functional roles, proteins are usually separated into domains before classification.

Table 6 Resources for Classification of Protein Sequences

Scheme	Description	Web Access
PFam	Domain-level classification of protein sequences	http://www.sanger.ac.uk/Software/Pfam/
PRINTS	Fingerprints information on protein sequences	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
PROSITE	Sequence motif definition	http://www.expasy.org/prosite/
TIGRFAMS	Protein family database	http://www.tigr.org/TIGRFAMs/
PRODOM	Protein domain database	http://protein.toulouse.inra.fr/prodom.html
BLOCKS	Multiple-alignment blocks	http://blocks.fhcrc.org/
eMOTIF	Protein motif database, derived from PRINT and BLOCKS	http://motif.stanford.edu/emotif/
CluSTr	Clusters of related proteins	http://www.ebi.ac.uk/clustr/
COGS	Clusters of orthologous groups	http://www.ncbi.nlm.nih.gov/COG/
ProtoMap	Hierarchical classification of protein sequences	http://protomap.cornell.edu
TRIBES	Protein family databases	http://maine.ebi.ac.uk:8000/services/tribes/
PIR international	Protein sequence databases	http://pir.georgetown.edu/
SYSTERS	Protein family database	http://systers.molgen.mpg.de/
SMART	Small motif database	http://smart.embl-heidelberg.de/
UniProt	Catalog of information on proteins	http://www.expasy.uniprot.org/
InterPro	Databases of protein families and domains	http://www.ebi.ac.uk/interpro/

How to identify and delineate these domains is still an open problem as discussed above. It is important to realize that the existing algorithms for domain identification do not always agree; the corresponding discrepancies in domain definition translate into differences between structural classifications that do not share the same definition.

Table 7 Resources for Protein Structure Classifications

Scheme	Description	Web Access
SCOP	Structural Classification of Protein: manual	http://scop.mrc-lmb.cam.ac.uk/scop/index.html
CATH	Class, Architecture, Topology, Homology: semiautomatic classification of proteins	http://www.biochem.ucl.ac.uk/bsm/cath
DALI Fold Classification	Automatic classification of DALI domain using Dali. Supersedes FSSP	http://www.bioinfo.biocenter.helsinki.fi:8080/dali/index.html
ASTRAL	Databases and tools for analyzing protein structure; derived from SCOP	http://astral.berkeley.edu/
HOMSTRAD	Aligned 3-D structures of homologous proteins	http://www-cryst.bioc.cam.ac.uk/data/align

Once proteins are divided into domains; the domains are then classified hierarchically. At the top of the classification we usually find the "class" of a protein domain, which is generally determined from its overall composition in secondary structure elements. Three main classes of protein domains exist: mainly α domains, mainly β domains, and mixed $\alpha - \beta$ domains (the domains in the $\alpha - \beta$ class are sometimes subdivided into domains with alternating α/β secondary structures and domains with mixed $\alpha + \beta$ secondary structures). In each class, domains are clustered into "folds" according to their topology. A fold is determined from the number, arrangement, and connectivity of the domain's secondary structure elements. The folds are subdivided into "superfamilies." A superfamily contains protein domains with similar functions, which suggests a common ancestry, often in the absence of detectable sequence similarity. Sequence information defines "families," i.e., subclasses of superfamilies that regroup domains whose sequences are similar.

Classification schemes are designated as being curated and automatic. A curated classification is one that is based mainly on human expertise, sometimes guided by computer analyses, to identify similarities between protein structures for organization into groups. An automated classification relies exclusively on the results of a computer procedure to identify the similarities, which are subsequently processed automatically to generate the groups. One advantage of curation is the typically high quality of the clustering results; the disadvantage is that curation is difficult to scale to high volumes of data. Conversely, automatic procedures are fully reproducible and scalable, but they may inaccurately assign similarity. The three most common protein structure classifications illustrate these differences: SCOP is almost completely manually derived, the DALI domain dictionary is based on a fully automated procedure, and CATH is intermediate between these two classification, using automated procedures complemented with human interventions.

The Structure Classification of Proteins (SCOP)

SCOP[6] is a repository that organizes protein structures hierarchically to reflect both structural and evolutionary relatedness. SCOP has been constructed manually, from the delineation of the domains in multidomain proteins to the organization of the levels of the hierarchy. It relies on visual inspection and comparison of protein structures, with the assistance of some automatic computer tools to make the task manageable and to help provide consistency and generality. Since its creation in 1994, SCOP has been updated regularly, with an average frequency of two releases a year. The latest update of SCOP, 1.65, was built from the 20,619 PDB entries (54,745 domains) available on August 1, 2003 and was released in December 2003. Statistics on the growth of SCOP are given in Figure 11.

SCOP is a hierarchic classification with four major levels: classes, folds, superfamilies, and families. As recognized by the authors of SCOP, the exact

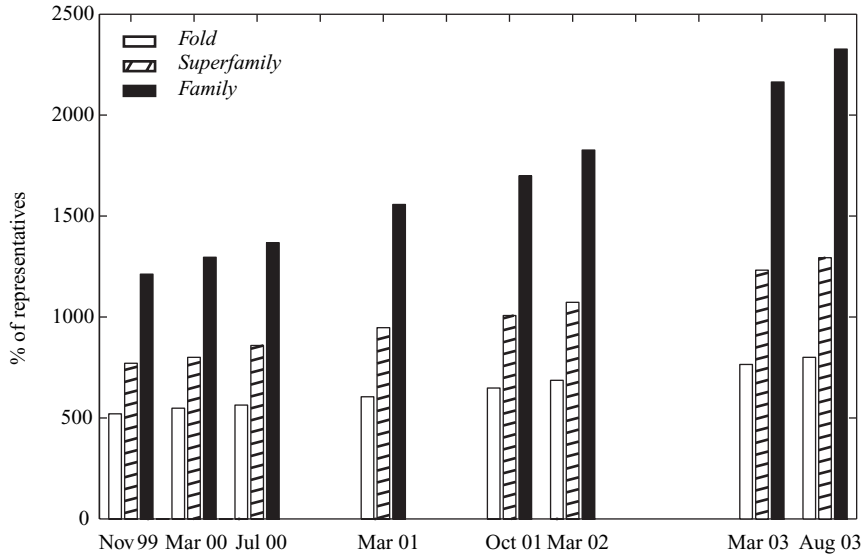


Figure 11 Statistics of the SCOP classification of proteins. The numbers of folds, superfamilies, and families in SCOP are plotted against "time", where time is the timestamp of the PDB used to generate the update of SCOP.

positions of boundaries separating these levels are subjective; where any doubts of similarity existed, they have chosen to create new divisions at the family and superfamily levels.

At the top of the SCOP hierarchy are 11 different classes: alpha, beta, alpha and beta (α/β), alpha plus beta ($\alpha + \beta$), multidomain proteins, membrane and cell-surface proteins, small proteins, coiled coil proteins, low-resolution protein structures, peptides, and designed proteins. Note that only the first seven classes are true classes. The remaining ones serve as place holders for protein domains that cannot (yet) be classified among the major classes and are maintained in SCOP for the sake of completeness and compatibility with the PDB.

In each SCOP class, proteins are clustered into groups based on their structure similarity. Each cluster is referred to by SCOP as a fold. Proteins share a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Proteins with the same fold may differ at the level of their peripheral elements, which can include secondary structures and turn regions. Note that these peripheral elements can represent up to 50% of the structure. Proteins catalogued together in the same fold may have no common evolutionary origin.

SCOP superfamilies identify probable common evolutionary origin. Proteins whose sequences have low similarities, but that share the same fold and have similar functions, suggest that a common evolutionary origin is probable.

Proteins clustered together into families are clearly evolutionary related. The sequences of two proteins of the same family often have a residue identity greater than 30%. In some cases, a high sequence identity is not needed to affirm common origin; many globins, for example, form a family, even though some members of that family have a sequence identity of only 15%.

The CATH Classification

CATH[7] is a protein structure classification in which protein domains are clustered at four major levels: class (C), architecture (A), topology (T), and homologous superfamily (H), each of which are described below. CATH uses a semiautomatic classification procedure that filters out nonprotein, models, and "C_α-only" structures from the PDB. Only crystal structures solved to resolution better than 3.0 Å are considered, together with all NMR structures. The latest update of CATH, v2.5.1, was released January 28, 2004 and includes 48,391 domains.

Multidomain proteins are subdivided into individual domains with a consensus procedure based on three algorithms for domain recognition: DETECTIVE,[82] PUU,[77] and DOMAK.[77] When all three algorithms generate the same result on a multidomain protein, the common solution delineates the domains of that protein. This consensus procedure resulted in 53% of the proteins included in the CATH release 2.5.1 to be subdivided into domains automatically. The remaining structures were assigned domains manually, using one of the assignments made by the automatic procedure, an assignment obtained from the literature, or based on a new assignment defined by visual inspection.

CATH includes four classes (C): alpha, beta, alpha and beta, and few secondary structure (FSS). The alpha–beta class includes both alternating alpha/beta structures and alpha + beta structures, originally defined by Levitt and Chothia.[31] The class of a protein domain is determined according to its secondary structure composition and packing. Ninety percent of the protein domains were automatically assigned to their class in CATH 2.5.1, using the method developed by Michie et al.[169,170] The remaining 10% of domains were assigned to a class by visual inspection.

The architecture (A) level included in CATH describes the overall shape of the domain structures, as determined by the orientation of their secondary structures, ignoring their connectivity. It is assigned manually. This level has no equivalent in SCOP.

Domains are grouped into topologies (T), or fold families, according to their overall shape and the connectivity of their secondary structures. This is done with the structural alignment program SSAP.[101] Proteins belonging to the same class are compared systematically with SSAP, and the corresponding scores are stored in a two-dimensional matrix. Structure pairs that have a sufficiently high SSAP score (>70) are merged into fold families, using single linkage clustering (for a brief description of this clustering technique, see the Appendix).

The Homologous Superfamily level, or H level, groups together protein domains thought to share a common ancestor. This level is equivalent to the superfamily level defined in SCOP. CATH also includes a Sequence Families level (S-level) that is equivalent to the family level of SCOP.

The DALI Domain Dictionary (DDD)

The DDD, also called the DALI domain classification, is derived from a fully automated method of defining and classifying domains.[171,172] DALI domains are defined by a version of the PUU algorithm[77] that has been updated to consider the recurrence of putative domains.[173] When comparing two protein structures, DALI computes a similarity measure, or S score. The mean and standard deviations of the S scores obtained over all pairs of proteins are evaluated. Shifting the S scores by their mean and rescaling by the standard deviation yield the statistically meaningful Z-scores.

The program DALI was used initially to create the Families of Structurally Similar Proteins (FSSP) database.[174] In FSSP, pair-wise structural comparisons are made between proteins of a representative set, in which no two proteins have greater than 25% sequence identity. For each member of the representative set, a file is created that contains all pair-wise structural matches with a Z-score greater than 2.0. The same procedure generates a complete classification of all protein domains in the PDB90 database, the DDD.[172] PDB90 is a representative subset of the PDB, where no two chains share more than 90% sequence identity. An average linkage hierarchical clustering technique (see the Appendix) generates a fold tree covering the PDB90 database. The pair-wise structural alignments are divided with Z-score cutoffs of 2, 4, 8, 16, 32, and 64, creating a six-character index for each domain. The first level ($Z > 2$) is used as an operational definition of folds. Lower levels should not be confused with the superfamily and family levels of CATH and SCOP, as they are not based on direct functional or evolutionary relationships. Both FSSP and the DDD are continuously updated; this is possible as they are both derived from a fully automated procedure.

Comparing SCOP, CATH, and DDD

SCOP, CATH, and DDD agree on most of their classifications, despite differences in the classification methods they have implemented, and in the rules of protein structure and taxonomy they are based on. Hadley and Jones[175] were the first to publish a detailed comparison of the fold classifications produced by SCOP, CATH, and FSSP. They showed that the three classification systems tend to agree in most cases, and that the discrepancies and inconsistencies are accounted for by a small number of problems. Among these, the domain assignment plays a crucial role. As mentioned, the separation of proteins into domains is a difficult and often subjective process. Many

protein structures are assigned different numbers of chains in SCOP, CATH, and FSSP. An obvious domain problem that results is the exclusion of one part of a protein. Hadley and Jones[175] reported the case of papain, a cysteine proteinase from papaya, which was treated as a single domain by SCOP, leaving the catalytic cysteine, histidine, and asparagines together to form the active site, while CATH split the protein into two domains, separating the cysteine from the asparagine and histidine, and rendering each domain effectively functionless. Note that this difference between SCOP and CATH has been corrected since Hadley and Jones published their study, and papain is now treated as a single domain in CATH. Another discrepancy between the structural classifications originates from the "fold overlap" problem, where a fold within one classification encompasses more than one fold within another classification. When a domain is classified in CATH as a three-layer ($\alpha\beta\alpha$) sandwich Rossmann fold, there are several SCOP folds to which it could belong. Although the structures are geometrically similar, SCOP can separate them to reflect an evolutionary distinction. This "fold overlap" problem is observed, for example, for the protein 1phr, and the chain A of the proteins 1gar and 1lfa, corresponding to a phosphotyrosine protein phosphatase, a formyltransferase, and an integrin, respectively. All three structures contain a three-layer sandwich Rossmann fold and are consequently regrouped in the same Topology in CATH (topology index 3.40.50), while they are representatives of their fold class in SCOP (classes c.44, c.65, and c.62, respectively).

Despite these discrepancies, Hadley and Jones[175] recognized the merits of all three classifications and concluded that no one method is distinctly superior to another. They characterize SCOP as a valuable resource for detailed evolutionary information, CATH as a source of geometric information, and FSSP as a raw source of information, which is continually updated.

Divergences in protein structure classifications have triggered the search for a consensus description of the protein structure space. Day et al.[176] recently repeated the comparative study of SCOP, CATH, and DDD previously done by Hadley and Jones, using updated versions of the classifications. Although Day et al. find significant levels of agreement between the three classifications, they highlight disparities whose origins are similar to those found earlier by Hadley and Jones. To remove these disparities, they introduced the concept of consensus folds. Day et al. start from a nonredundant subset of protein domains. To be considered in the analysis, the authors insisted that 80% of the sequence of a domain in SCOP must be present in a DALI domain definition, 80% of the DALI domain must be present in the SCOP definition, and so on for the other pair-wise combinations of the classification systems. Redundant domains were considered as those having >95% sequence identity to a previously counted domain. Each domain in the nonredundant subset was assigned a fold identifier, which corresponds to its classification in SCOP, CATH, and DDD. Domains were then clustered on the basis of their fold identifiers, and the corresponding clusters were referred

to as metafolds. The nonredundant set contained 5720 domains, clustered into 1130 metafolds. About half of these domains are described by one of the top 30 metafolds. These metafolds represent the consensus information contained in SCOP, CATH, and DDD, and as such define a consensus view of the protein fold space.

CONCLUSIONS

Proteins are key molecules in all cellular functions. Nature has extensively explored their sequences and structures to build the library of functions needed for the diversity of life, taking into account all external constraints and the corresponding adaptation. The wealth of information encoded in the protein sequences and structures therefore provides the clues needed to unravel the mysteries of life and its evolution and adaptation over time. Understanding this diversity has become a key topic in recent genetics and molecular biology studies, catalyzed by the development of numerous genomics and structural genomics projects. More than 220 whole genomes have been sequenced and published on the World Wide Web, and more than 1200 are currently under study, which corresponds to databases in excess of one million nonredundant protein sequences. In parallel, the Protein Data Bank contains structural data on more than 27,000 proteins. The challenge now is to organize these data in a way that evolutionary relationships between proteins can be uncovered and used to understand better protein function. The past few years have seen an explosion of techniques in ''bio-informatics'' for organizing and analyzing protein sequence families. Although such approaches can detect homologous proteins, they usually fail to detect remote homologues, i.e., pairs of proteins that have similar structure and function, but that lack easily detectable sequence similarity. Because protein structures are more highly conserved than are protein sequences, there is a growing interest in studying evolution based on an understanding of the protein structure space. The first steps common to the analysis of any large set of data are to group together data points that are similar, and then to identify connections between those elementary groups. These steps are usually performed with classification techniques. In the case of protein structures, this has led to the construction and maintenance of protein structure classifications, which have been reviewed in this tutorial.

Reliable protein structure superposition remains a bottleneck when carrying out a protein structure classification. Comparing and grouping proteins require a definition of the similarity of two structures. Similarity in structural alignment is geometric and captured by the cRMS deviation of the aligned atoms. Other properties of structural alignments that are likely to be significant are the number of positions matched and the number and length of gaps. Good alignments match more positions, have fewer gaps, and are more similar than do poor alignments. Because these properties of alignments

are not independent (shortening the alignment or introducing many gaps can lower the cRMS), researchers have devised alignment scores that attempt to balance their influences. Several measures of similarity have consequently been developed.[143] Perhaps the most significant recent improvements in this area have been in the protocols for assessing the statistical significance of these measures.[119,177] These statistical measures of similarity are now being used in structure comparison algorithms. Ideally, they should detect reliably distant relatives and be fast enough to scan large databases of representative protein structures. Existing methods have been designed to satisfy one or the other, but not both of these two criteria simultaneously (see the section on protein structure superposition). A need exists for a fast, reliable protein structure superposition program.

Critical to the classification of proteins is the definition of domains. It has long been hypothesized that domains are the important evolutionary units. It is supported by recent analyses of the available genome data, which suggest that at least 60% of the genes are multidomain proteins.[178–180] Domain duplications and recombination are thought to have occurred extensively in nature. Protein structure classifications are consequently domain based. Automatic recognition of domains in multidomain proteins can be difficult, although many promising approaches have been developed (see the section on protein domain above). These methods do not always agree with their domain assignments, which in turn leads to discrepancies between the existing protein structure classifications.[175,176]

The three major protein structure classifications are SCOP, CATH, and DDD. SCOP is derived manually and is recognized as a valuable resource of detailed evolutionary information. CATH provides useful geometric information. It also introduces the concept of ''architecture,'' which reveals broad features of the protein structure space. CATH relies on partial automation and as such is subject to inaccuracies introduced by fixed thresholds. The DDD is a fully automatic classification continually updated. It is not as popular as SCOP and CATH, probably because its automatic levels are not as intuitive and require more input from the users to be interpreted.

Protein structure classifications need to be linked with the other genome databases under constructions. Currently, SCOP, CATH, and DDD are valuable resources used mostly for benchmarking of methods and for structural studies. Their impact on biology will be far greater when they are integrated with sequence and function information to present a cohesive picture of the different protein spaces.

ACKNOWLEDGMENTS

Support from the National Science Foundation (Grant CCR-00-86013) and the National Institute of Health (GM 63817) is acknowledged.

APPENDIX: HIERARCHICAL CLUSTERING

The aim of clustering is to group a collection of objects (or observations) into subsets of "clusters," such that those objects within each cluster are more similar to one another than objects assigned to other clusters. Two main elements exist in any clustering technique: the definition of similarity or dissimilarity between objects, and the algorithm that partitions the data into clusters. Here it is assumed that the similarity is known and encoded into a distance ***d*** between the objects. There are two major types of algorithm for portioning objects: k-means clustering and hierarchical clustering.

In hierarchical clustering, the data are regrouped into clusters through a series of partitioning events. Each partition can run from a single cluster containing all n objects to n clusters each containing a single object. Hierarchical clustering techniques are subdivided into two groups: *agglomerative* methods that fuse the objects into groups and *divisive* methods that separate the objects successively into finer groupings. Here the focus is on agglomerative methods, because they are used for generating protein structure classifications.

An agglomerative hierarchical clustering technique involves creating a series of partitions of the n data, P_n, P_{n-1},..., P_1, such that P_n consists of n clusters each containing a single object, and P_1 consists of a single group containing all n objects. At each stage, the procedure joins together the two nearest clusters. Differences between methods are from different ways of defining the distance between clusters. The four main agglomerative hierarchical clustering techniques are as follows:

- ***Single linkage clustering***: The distance between two clusters A and B is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each cluster are considered:

$$D(A,B) = \min\{d(a,b),\ (a,b) \in A \times B\} \qquad [A.1]$$

- ***Complete linkage clustering***: The distance between two clusters A and B is defined as the distance between the most distant pair of objects, one from each cluster:

$$D(A,B) = \max\{d(a,b),\ (a,b) \in A \times B\} \qquad [A.2]$$

- ***Average linkage clustering***: The distance between two clusters A and B is defined as the average of distances between all pairs of objects, where each pair is composed of one object from each cluster:

$$D(A,B) = \frac{\sum_{a \in A} \sum_{b \in B} d(a,b)}{N_A N_B} \qquad [A.3]$$

where N_A and N_B are the sizes of A and B, respectively.

- ***Average group linkage***: The distance between two clusters A and B is defined as the average of distances between all pairs of objects included in the union of *A* and *B*:

$$D(A,B) = Average\left\{ d(a,b),\ (a,b) \in (A \cup B)^2 \right\} \qquad [A.4]$$

There is no answer to the question about which of these techniques performs best. Clustering is an exploratory data analysis procedure; the choice of which technique to be used for clustering often comes from a very good understanding of the objects to be clustered. A tutorial on clustering methods used in computational chemistry has appeared in this series and should be consulted.[181]

REFERENCES

1. J. Monod, *Le Hasard Et La Nécessité*, Seuil, Paris, France, 1973.
2. A. Bernal, U. Ear, and N. Kyrpides, *Nucl. Acids Res.*, **29**, 126 (2001). Genomes Online Database (Gold): A Monitor of Genome Projects World-Wide.
3. F. C. Bernstein, T. F. Koetzle, G. William, D. J. Meyer, M. D. Brice, and J. R. Rodgers, *J. Mol. Biol.*, **112**, 535 (1977). The Protein Databank: A Computer-Based Archival File for Macromolecular Structures.
4. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, and H. Weissig, *Nucl. Acids Res.*, **28**, 235 (2000). The Protein Data Bank.
5. W. Gilbert, *Nature* (*London*), **349**, 99 (1991). Towards a Paradigm Shift in Biology.
6. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, *J. Mol. Biol.*, **247**, 536 (1995). SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures.
7. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, *Structure*, **5**, 1093 (1997). CATH: A Hierarchic Classification of Protein Domain Structures.
8. L. Holm and C. Sander, *J. Mol. Biol.*, **233**, 123 (1993). Protein Structure Comparison by Alignment of Distance Matrices.
9. G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure*, Springer-Verlag, New York, 1979.
10. C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry: The Conformation of Biological Macromolecules*, W. H. Freeman Company, New York, 1980.
11. C. Branden and J. Tooze, *Introduction to Protein Structure*, Garland Publishing, New York, 1991.
12. T. E. Creighton, *Proteins*, W. H. Freeman & Co., New York, 1993.
13. W. R. Taylor, A. C. W. May, N. P. Brown, and A. Aszodi, *Reports Prog. Phys.*, **64**, 517 (2001). Protein Structure: Geometry, Topology and Classification.
14. R. B. Corey and L. Pauling, *Rev. Sci. Instr.*, **24**, 621 (1953). Molecular Models of Amino Acids, Peptides and Proteins.
15. W. L. Koltun, *Biopolymers*, **3**, 665 (1965). Precision Space-Filling Atomic Models.
16. J. Kendrew, R. Dickerson, B. Strandberg, R. Hart, D. Davies, and D. Philips, *Nature* (*London*), **185**, 422 (1960). Structure of Myoglobin: A Three Dimensional Fourier Synthesis at 2 Angstrom Resolution.
17. P. J. Kraulis, *J. Appl. Cryst.*, **24**, 946 (1991). Molscript: A Program to Produce Both Detailed and Schematic Plots of Protein Structures.

18. W. Humphrey, A. Dalke, and K. Schulten, *J. Molec. Graphics*, **14**, 33 (1996). VMD - Visual Molecular Dynamics.
19. K. C. Timberlake, *General, Organic, and Biological Chemistry: Structures of Life*, Benjamin Cummings, San Francisco, CA, 2004.
20. G. M. Crippen, *J. Mol. Biol.*, **126**, 315 (1978). Tree Structural Organization of Proteins.
21. A. V. Efimov, *FEBS Lett.*, **224**, 372 (1987). Pseudo-Homology of Protein Standard Structures Formed by 2 Consecutive Beta-Strands.
22. A. V. Efimov, *FEBS Lett.*, **284**, 288 (1991). Structure of Coiled Beta-Beta Hairpins and Beta-Beta-Corners.
23. A. V. Efimov, *Protein Eng.*, **4**, 245 (1991). Structure of Alpha-Alpha Hairpins with Short Connections.
24. A. V. Efimov, *Prog. Biophys. Mol. Biol.*, **60**, 201 (1993). Standard Structures in Proteins.
25. C. Brooks, M. Karplus, and M. Pettitt, *Adv. Chem. Phys.*, **71**, 1 (1988). Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics.
26. C. B. Anfinsen, *Science*, **181**, 223 (1973). Principles That Govern Protein Folding.
27. P. Koehl and M. Levitt, *Nature Struct. Biol.*, **6**, 108 (1999). A Brighter Future for Protein Structure Prediction.
28. D. Baker and A. Šali, *Science*, **294**, 93 (2001). Protein Structure Prediction and Structural Genomics.
29. J.-E. Shea, M. R. Friedel, and A. Baumketner in *Rev. Comput. Chem.*, K. B. Lipkowitz, T. R. Cundari, and V. Gillet, Eds., Wiley, New York, 2006, Vol. **22**, pp. 169–228. Predicting Protein Folding.
30. M. Perutz, M. Rossman, A. Cullis, G. Muirhead, G. Will, and A. North, *Nature* (*London*), **185**, 416 (1960). Structure of Haemoglobin: A Three-Dimensional Fourier Synthesis at 5.5 Angstrom Resolution, Obtained by X-Ray Analysis.
31. M. Levitt and C. Chothia, *Nature* (*London*), **261**, 552 (1976). Structural Patterns in Globular Proteins.
32. A. M. Lesk and C. Chothia, *J. Mol. Biol.*, **136**, 225 (1980). How Different Amino-Acid Sequences Determine Similar Protein Structures: The Structure and Evolutionary Dynamics of the Globins.
33. C. Chothia and J. Janin, *Proc. Natl. Acad. Sci. (USA)*, **78**, 4146 (1981). Relative Orientation of Close Packed Beta Pleated Sheets in Proteins.
34. F. E. Cohen, M. J. E. Sternberg, and W. R. Taylor, *J. Mol. Biol.*, **148**, 253 (1981). Analysis of the Tertiary Structure of Protein Beta Sheet Sandwiches.
35. C. Chothia and J. Janin, *Biochemistry*, **21**, 3955 (1982). Orthogonal Packing of Beta Pleated Sheets in Proteins.
36. F. E. Cohen, M. J. E. Sternberg, and W. R. Taylor, *J. Mol. Biol.*, **156**, 821 (1982). Analysis and Prediction of the Packing of Alpha Helices against a Beta Sheet in the Tertiary Structure of Globular Proteins.
37. K. C. Chou, *Proteins: Struct. Func. Genet.*, **21**, 319 (1995). A Novel Approach to Predicting Protein Structural Classes in a (20-1)-D Amino Acid Composition Space.
38. K. C. Chou and C. T. Zhang, *Critical Rev. Biochem. Molec. Biol.*, **30**, 275 (1995). Prediction of Protein Structural Classes.
39. I. Bahar, A. R. Atilgan, R. L. Jernigan, and B. Erman, *Proteins: Struct. Func. Genet.*, **29**, 172 (1997). Understanding the Recognition of Protein Structural Classes by Amino Acid Composition.
40. W. M. Liu and K. C. Chou, *J. Protein Chem.*, **17**, 209 (1998). Prediction of Protein Structural Classes by Modified Mahalanobis Discriminant Algorithm.
41. K. C. Chou, W. M. Liu, G. M. Maggiora, and C. T. Zhang, *Proteins: Struct. Func. Genet.*, **31**, 97 (1998). Prediction and Classification of Domain Structural Classes.
42. Y. D. Cai, Y. X. Li, and K. C. Chou, *Biochim. Biophys. Acta*, **1476**, 1 (2000). Using Neural Networks for Prediction of Domain Structural Classes.

43. G. P. Zhou and N. Assa-Munt, *Proteins: Struct. Func. Genet.*, **44**, 57 (2001). Some Insights into Protein Structural Class Prediction.
44. R. Y. Luo, Z. P. Feng, and J. K. Liu, *Eur. J. Biochem.*, **269**, 4219 (2002). Prediction of Protein Structural Class by Amino Acid and Polypeptide Composition.
45. E. G. Hutchinson and J. M. Thornton, *Protein Eng.*, **6**, 233 (1993). The Greek Key Motif: Extraction, Classification and Analysis.
46. J. S. Richardson, *Nature* (*London*), **268**, 495 (1977). β-Sheet Topology and the Relatedness of Proteins.
47. D. W. Banner, A. C. Bloomer, G. A. Petsko, D. C. Phillips, C. I. Pogson, I. A. Wilson, P. H. Corran, A. J. Furth, J. D. Milman, R. E. Offord, J. D. Priddle, and S. G. Waley, *Nature* (*London*), **255**, 609 (1975). Structure of Chicken Muscle Triose Phosphate Isomerase Determined Crystallographically at 2.5 Å Resolution Using Amino-Acid Sequence Data.
48. A. G. Murzin, A. M. Lesk, and C. Chothia, *J. Mol. Biol.*, **236**, 1369 (1994). Principles Determining the Structure of Beta-Sheet Barrels in Proteins. 1. A Theoretical-Analysis.
49. A. G. Murzin, A. M. Lesk, and C. Chothia, *J. Mol. Biol.*, **236**, 1382 (1994). Principles Determining the Structure of Beta-Sheet Barrels in Proteins. 2. The Observed Structures.
50. G. Pujadas and J. Palau, *Biologia* (*Bratislava*), **54**, 231 (1999). Tim Barrel Fold: Structural, Functional and Evolutionary Characteristics in Natural and Designed Molecules.
51. R. K. Wierenga, *FEBS Lett.*, **492**, 193 (2001). The Tim-Barrel Fold: A Versatile Framework for Efficient Enzymes.
52. N. Nagano, C. A. Orengo, and J. M. Thornton, *J. Mol. Biol.*, **321**, 741 (2002). One Fold with Many Functions: The Evolutionary Relationships between Tim Barrel Families Based on Their Sequences, Structures and Functions.
53. M. C. Vega, E. Lorentzen, A. Linden, and M. Wilmanns, *Curr. Opin. Chem. Biol.*, **7**, 694 (2003). Evolutionary Markers in the (Beta/Alpha)8-Barrel Fold.
54. J. A. Gerlt and F. M. Raushel, *Curr. Opin. Chem. Biol.*, **7**, 252 (2003). Evolution of Function in (Beta/Alpha)8-Barrel Enzymes.
55. E. L. Wise and I. Rayment, *Acc. Chem. Res.*, **37**, 149 (2004). Understanding the Importance of Protein Structure to Nature's Routes for Divergent Evolution in Tim Barrel Enzymes.
56. G. D. Rose, *J. Mol. Biol.*, **134**, 447 (1979). Hierarchic Organization of Domains in Globular Proteins.
57. J. S. Richardson, *Adv. Protein Chem.*, **34**, 167 (1981). The Anatomy and Taxonomy of Protein Structure.
58. J. Janin and C. Chothia, *Meth. Enzymol.*, **115**, 420 (1985). Domains in Proteins - Definitions, Location, and Structural Principles.
59. C. P. Ponting and R. B. Russell, in *Annu. Rev. Biophys. Biomol. Struct.*, R. M. Stroud, W. K. Olson, and M. P. Sheetz, Eds., Annual Reviews, Palo Alto, CA, 2002, Vol. **31**, pp. 45–71. The Natural History of Protein Domains.
60. S. Veretnik, P. E. Bourne, N. N. Alexandrov, and I. N. Shindyalov, *J. Mol. Biol.*, **339**, 647 (2004). Toward Consistent Assignment of Structural Domains in Proteins.
61. R. A. Laskowski, E. G. Hutchinson, A. D. Michie, A. C. Wallace, M. L. Jones, and J. M. Thornton, *Trends Biochem. Sci.*, **22**, 488 (1997). PDBSum: A Web-Based Database of Summaries and Analyses of All PDB Structures.
62. R. A. Laskowski, *Nucl. Acids Res.*, **29**, 221 (2001). PDBSum: Summaries and Analyses of PDB Structures.
63. W. Kabsch and C. Sander, *Biopolymers*, **22**, 2577 (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features.
64. G. Wang and R. L. Dunbrack, *Bioinformatics* (*Oxford*), **19**, 1589 (2003). Pisces: A Protein Sequence Culling Server.
65. S. E. Brenner, P. Koehl, and M. Levitt, *Nucl. Acids Res.*, **28**, 254 (2000). The Astral Compendium for Protein Structure and Sequence Analysis.

66. J. M. Chandonia, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner, *Nucl. Acids Res*., **30**, 260 (2002). Astral Compendium Enhancements.

67. J. M. Chandonia, G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner, *Nucl. Acids Res*., **32**, D189 (2004). The Astral Compendium in 2004.

68. M. G. Rossman and A. Liljas, *J. Mol. Biol*., **85**, 177 (1974). Recognition of Structural Domains in Globular Proteins.

69. M. H. Zehfus and G. D. Rose, *Biochemistry*, **25**, 5759 (1986). Compact Units in Proteins.

70. S. A. Islam, J. C. Luo, and M. J. E. Sternberg, *Protein Eng*., **8**, 513 (1995). Identification and Analysis of Domains in Proteins.

71. A. S. Siddiqui and G. J. Barton, *Protein Sci*., **4**, 872 (1995). Continuous and Discontinuous Domains: An Algorithm for the Automatic Generation of Reliable Protein Domain Definitions.

72. N. Alexandrov and I. Shindyalov, *Bioinformatics* (*Oxford*), **19**, 429 (2003). PDB: Protein Domain Parser.

73. S. J. Wodak and J. Janin, *Biochemistry*, **20**, 6544 (1981). Location of Structural Domains in Proteins.

74. A. A. Rashin, *Nature* (*London*), **291**, 85 (1981). Location of Domains in Globular Proteins.

75. N. Go, *Proc. Natl. Acad. Sci.* (*USA*), **80**, 1964 (1983). Modular Structural Units, Exon and Function in Chicken Lysozyme.

76. M. H. Zehfus, *Protein Eng*., **7**, 335 (1994). Binary Discontinuous Compact Protein Domains.

77. L. Holm and C. Sander, *Proteins: Struct. Func. Genet*., **19**, 256 (1994). Parser for Protein Folding Units.

78. Y. Xu, D. Xu, and H. N. Gambow, *Bioinformatics* (*Oxford*), **16**, 1091 (2000). Protein Domain Decomposition Using a Graph-Theoretic Approach.

79. J. T. Guo, D. Xu, D. Kim, and Y. Xu, *Nucl. Acids Res*., **31**, 944 (2003). Improving the Performance of Domainparser for Structural Domain Partition Using Neural Network.

80. W. R. Taylor, *Protein Eng*., **12**, 203 (1999). Protein Structural Domain Identification.

81. M. B. Swindells, *Protein Sci*., **4**, 93 (1995). A Procedure for the Automatic Determination of Hydrophobic Cores in Protein Structures.

82. M. B. Swindells, *Protein Sci*., **4**, 103 (1995). A Procedure for Detecting Structural Domains in Proteins.

83. A. S. Siddiqui and G. J. Barton, *Protein Sci*., **4**, 872 (1995). Continuous and Discontinuous Domains: An Algorithm for the Automatic-Generation of Reliable Protein Domain Definitions.

84. C. Guerra and S. Istrail, *Mathematical Methods for Protein Structure Analysis and Design, Advanced Lectures*, Springer, Berlin, Germany, 2003.

85. C. Chen and Q. Li, *Acta Cryst. A*, **A60**, 201 (2004). A Strict Solution for the Optimal Superposition of Protein Structures.

86. B. Sabata and J. K. Aggarwal, *Comput. Vis. Graph. Image Proc: Image Understanding*, **54**, 309 (1991). Estimation of Motion from a Pair of Range Images: A Review.

87. C. Ferrari and C. Guerra, in *Mathematical Methods for Protein Structure Analysis and Design*, C. Guerra and S. Istrail, Eds., Springer, Berlin, Germany, 2003, pp. 57–82. Geometric Methods for Protein Structure Comparison.

88. D. W. Eggert, A. Lorusso, and R. B. Fisher, *Mach. Vis. and Applic*., **9**, 272 (1997). Estimating 3D Rigid Body Transformations: A Comparison of Four Major Algorithms.

89. G. H. Golub and C. F. V. Loan, *Matrix Computation*, John Hopkins University Press, Baltimore, MD, 1996.

90. W. Kabsch, *Acta Cryst. A*, **32**, 922 (1976). Solution for Best Rotation to Relate 2 Sets of Vectors.

91. W. Kabsch, *Acta Cryst. A*, **34**, 827 (1978). Discussion of Solution for Best Rotation to Relate 2 Sets of Vectors.
92. A. D. McLachlan, *J. Mol. Biol.*, **128**, 49 (1979). Gene Duplications in the Structural Evolution of Chymotrypsin.
93. K. S. Arun, T. S. Huang, and S. D. Blostein, *IEEE Trans. Pattern Anal. & Machine Intel.*, **9**, 698 (1987). Least-Sqaure Fitting of Two 3D Point Sets.
94. P. H. Schonemann, *Psychometrica*, **31**, 1 (1966). A Generalized Solution of the Orthogonal Procrustes Problem.
95. B. Horn, H. Hilden, and S. Negahdaripour, *J. Opt. Soc. Am.*, **5**, 1127 (1988). Closed-Form Solution of Absolute Orientation Using Orthonormal Matrices.
96. B. Horn, *J. Opt. Soc. Am.*, **4**, 629 (1987). Closed-Form Solution of Absolute Orientation Using Unit Quaternions.
97. M. W. Walker, L. Shao, and R. A. Voltz, *CVGIP: Image Understanding*, **54**, 358 (1991). Estimating 3D Location Parameters Using Dual Number Quaternions.
98. E. A. Coutsias, C. Seok, and K. A. Dill, *J. Comput. Chem*, **25**, 1849 (2004). Using Quaternions to Calculate RMSD.
99. S. Umeyama, *IEEE Trans. Pattern Anal. & Machine Intel.*, **13**, 376 (1991). Least-Squares Estimation of Transformation Parameters between 2-Point Patterns.
100. P. Koehl, *Curr. Opin. Struct. Biol.*, **11**, 348 (2001). Protein Structure Similarities.
101. W. R. Taylor and C. A. Orengo, *J. Mol. Biol.*, **208**, 1 (1989). Protein Structure Alignment.
102. W. R. Taylor, *Protein Sci.*, **8**, 654 (1999). Protein Structure Comparison Using Iterated Double Dynamic Programming.
103. K. Nishikawa and T. Ooi, *J. Theor. Biol.*, **43**, 351 (1974). Comparison of Homologous Tertiary Structures of Proteins.
104. L. Holm, C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend, *Protein Sci.*, **1**, 1691 (1992). A Database of Protein Structure Families with Common Folding Motifs.
105. L. Holm and C. Sander, *Nature* (*London*), **361**, 309 (1993). Globin Fold in a Bacterial Toxin.
106. G. Vriend and C. Sander, *Proteins: Struct. Func. Genet.*, **11**, 52 (1991). Detection of Common 3-Dimensional Substructures in Proteins.
107. N. N. Alexandrov, K. Takahashi, and N. Go, *J. Mol. Biol.*, **225**, 5 (1992). Common Spatial Arrangements of Backbone Fragments in Homologous and Nonhomologous Proteins.
108. R. Nussinov and H. J. Wolfson, *Proc. Natl. Acad. Sci.* (*USA*), **88**, 10495 (1991). Efficient Detection of 3-Dimensional Structural Motifs in Biological Macromolecules by Computer Vision Techniques.
109. R. Nussinov, D. Fischer, and H. Wolfson, *FASEB J.*, **6**, A349 (1992). A Computer Vision Based 3-Dimensional Approach for the Comparison of Protein Structures.
110. D. Fischer, O. Bachar, R. Nussinov, and H. Wolfson, *J. Biomol. Struct. Dyn.*, **9**, 769 (1992). An Efficient Automated Computer Vision Based Technique for Detection of 3-Dimensional Structural Motifs in Proteins.
111. D. Fischer, H. Wolfson, and R. Nussinov, *J. Biomol. Struct. Dyn.*, **11**, 367 (1993). Spatial, Sequence-Order-Independent Structural Comparison of Alpha/Beta Proteins: Evolutionary Implications.
112. D. Fischer, H. Wolfson, S. L. Lin, and R. Nussinov, *Protein Sci.*, **3**, 769 (1994). 3-Dimensional, Sequence Order-Independent Structural Comparison of a Serine-Protease against the Crystallographic Database Reveals Active-Site Similarities - Potential Implications to Evolution and to Protein-Folding.
113. H. J. Wolfson and I. Rigoutsos, *IEEE Comput. Sci. & Eng.*, **4**, 10 (1997). Geometric Hashing: An Overview.
114. P. J. Artymiuk, A. R. Poirrette, H. M. Grindley, D. W. Rice, and P. Willett, *J. Mol. Biol.*, **243**, 327 (1994). A Graph-Theoretic Approach to the Identification of 3-Dimensional Patterns of Amino-Acid Side-Chains in Protein Structures.

115. P. J. Artymiuk, A. R. Poirrette, D. W. Rice, and P. Willett, *Topics Curr. Chem.*, **174**, 73 (1995). The Use of Graph-Theoretical Methods for the Comparison of the Structures of Biological Macromolecules.

116. E. J. Gardiner, P. J. Artymiuk, and P. Willett, *J. Mol. Graph. & Modelling*, **15**, 245 (1997). Clique-Detection Algorithms for Matching Three-Dimensional Molecular Structures.

117. T. D. Wu, S. C. Schmidler, T. Hastie, and D. L. Brutlag, *J. Comput. Biol.*, **5**, 585 (1998). Regression Analysis of Multiple Protein Structures.

118. S. Subbiah, D. V. Laurents, and M. Levitt, *Curr. Biol.*, **3**, 141 (1993). Structural Similarity of DNA-Binding Domains of Bacteriophage Repressors and the Globin Core.

119. M. Gerstein and M. Levitt, *Protein Sci.*, **7**, 445 (1998). Comprehensive Assessment of Automatic Structural Alignment against a Manual Standard; the SCOP Classification of Proteins.

120. A. S. Yang and B. Honig, *J. Mol. Biol.*, **301**, 665 (2000). An Integrated Approach to the Analysis and Modeling of Protein Sequences and Structures. I. Protein Structural Alignment and a Quantitative Measure for Protein Structural Distance.

121. J. D. Szustakowski and Z. P. Weng, *Proteins: Struct. Func. Genet.*, **38**, 428 (2000). Protein Structure Alignment Using a Genetic Algorithm.

122. I. N. Shindyalov and P. E. Bourne, *Protein Eng.*, **11**, 739 (1998). Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path.

123. M. E. Ochagavia and S. J. Wodak, *Proteins: Struct. Func. Bioinfo.*, **55**, 436 (2004). Progressive Combinatorial Algorithm for Multiple Structural Alignments: Application to Distantly Related Proteins.

124. R. Blankenbecler, M. Ohlson, C. Peterson, and M. Ringler, *Proc. Natl. Acad. Sci. (USA)*, **100**, 11936 (2003). Matching Protein Structures with Fuzzy Alignments.

125. L. Chen, T. Zhou, and Y. Tang, *Bioinformatics* (*Oxford*), **21**, 51 (2005). Protein Structure Alignment by Deterministic Annealing.

126. B. W. Matthews and M. G. Rossman, *Meth. Enzymol.*, **115**, 397 (1985). Comparison of Protein Structures.

127. M. L. Sierk and W. R. Pearson, *Protein Sci.*, **13**, 773 (2004). Sensitivity and Selectivity in Protein Structure Comparison.

128. M. Novotny, D. Madsen, and G. J. Kleywegt, *Proteins: Struct. Func. Genet.*, **54**, 260 (2004). Evaluation of Protein Fold Comparison Servers.

129. R. Kolodny, P. Koehl, and M. Levitt, *J. Mol. Biol.*, **346**, 1173 (2005). Comprehensive Evaluation of Structural Alignment Method: Scoring by Geometric Match Measures.

130. L. Holm and C. Sander, *Trends Biochem. Sci.*, **20**, 478 (1995). DALI - a Network Tool for Protein-Structure Comparison.

131. M. G. Rossmann and P. Argos, *J. Mol. Biol.*, **105**, 75 (1976). Exploring Structural Homology of Proteins.

132. R. B. Russell and G. J. Barton, *Proteins: Struct. Func. Genet.*, **14**, 309 (1992). Multiple Protein Sequence Alignment from Tertiary Structure Comparison Assignment of Global and Residue Confidence Levels.

133. M. Suyama, Y. Matsuo, and K. Nishikawa, *J. Mol. Evol.*, **44**, S163 (1997). Comparison of Protein Structures Using 3D Profile Alignment.

134. J. U. Bowie, R. Lüthy, and D. Eisenberg, *Science*, **253**, 164 (1991). A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure.

135. J. Jung and B. Lee, *Protein Eng.*, **13**, 535 (2000). Protein Structure Alignment Using Environmental Profiles.

136. T. Kawabata and K. Nishikawa, *Proteins: Struct. Func. Genet.*, **41**, 108 (2000). Protein Structure Comparison Using the Markov Transition Model of Evolution.

137. I. D. Kuntz, G. M. Crippen, P. A. Kollman, and D. Kimelman, *J. Mol. Biol.*, **106**, 983 (1976). Calculation of Protein Tertiary Structure.

138. G. M. Crippen and T. F. Havel, *Distance Geometry and Molecular Conformation*, Research Studies, Somerset, United Kingdom, 1988.

139. R. Kolodny and N. Linial, *Proc. Natl. Acad. Sci. (USA)*, **101**, 12201 (2004). Approximate Protein Structural Alignment in Polynomial Time.

140. K. Kaindl and B. Steipe, *Acta Cryst. A*, **53**, 809 (1997). Metric Properties of the Root-Mean-Square Deviation of Vector Sets.

141. K. Mizuguchi and N. Go, *Curr. Opin. Struct. Biol.*, **5**, 377 (1995). Seeking Significance in 3-Dimensional Protein-Structure Comparisons.

142. A. M. Lesk, *Proteins: Struct. Func. Genet.*, **33**, 320 (1998). Extraction of Geometrically Similar Substructures: Least-Squares and Chebyshev Fitting and the Difference Distance Matrix.

143. A. C. W. May, *Proteins: Struct. Func. Genet.*, **37**, 20 (1999). Toward More Meaningful Hierarchical Classification of Protein Three-Dimensional Structures.

144. C. A. Orengo, M. B. Swindells, A. D. Michie, M. J. Zvelebil, P. C. Driscoll, M. D. Waterfield, and J. M. Thornton, *Protein Sci.*, **4**, 1977 (1995). Structure Similarity between the Pleckstrin Homology Domain and Verotoxin: The Problem of Measuring and Evaluating Structural Similarity.

145. Z. K. Feng and M. J. Sippl, *Folding & Design*, **1**, 123 (1996). Optimum Superimposition of Protein Structures: Ambiguities and Implications.

146. A. Godzik, *Protein Sci.*, **5**, 1325 (1996). The Structural Alignment between Two Proteins: Is There a Unique Answer?

147. T. J. P. Hubbard, *Proteins: Struct. Func. Genet.*, **Suppl 3**, 15 (1999). RMS/Coverage Graphs: A Qualitative Method for Comparing Three-Dimensional Protein Structure Predictions.

148. G. A. Arteca, in *Rev. Comput. Chem.*, K. B. Lipkowitz and D. B. Boyd, Eds., VCH Publishers, New York, 1996, Vol. **9**, pp. 191–254. Molecular Shape Descriptors.

149. P. Rogen and H. Bohr, *Math. Biosci.*, **182**, 167 (2003). A New Family of Global Protein Shape Descriptors.

150. D. Barnatan, *Topology*, **34**, 423 (1995). On the Vassiliev Knot Invariants.

151. A. Stasiak and J. H. Maddocks, *Nature* (*London*), **406**, 251 (2000). Mathematics - Best Packing in Proteins and DNA.

152. K. A. Hoffman, R. S. Manning, and J. H. Maddocks, *Biopolymers*, **70**, 145 (2003). Link, Twist, Energy, and the Stability of DNA Minicircles.

153. M. Levitt, *J. Mol. Biol.*, **170**, 723 (1983). Protein Folding by Restrained Energy Minimization and Molecular Dynamics.

154. G. A. Arteca, *Biopolymers*, **33**, 1829 (1993). Overcrossing Spectra of Protein Backbones - Characterization of 3-Dimensional Molecular Shape and Global Structural Homologies.

155. G. A. Arteca, *Phys. Rev. E*, **49**, 2417 (1994). Scaling Behavior of Some Molecular Shape Descriptors of Polymer Chains and Protein Backbones.

156. G. A. Arteca, *Phys. Rev. E*, **51**, 2600 (1995). Scaling Regimes Self-Entanglements in Very Compact Proteins.

157. G. A. Arteca and O. Tapia, *J. Chem. Inf. Comput. Sci.*, **39**, 642 (1999). Characterization of Fold Diversity among Proteins with the Same Number of Amino Acid Residues.

158. C. T. Reimann, G. A. Arteca, and O. Tapia, *Phys. Chem. Chem. Phys.*, **4**, 4058 (2002). Proteins in Vacuo. A Connection between Mean Overcrossing Number and Orientationally-Averaged Collision Cross Section.

159. P. Rogen and B. Fain, *Proc. Natl. Acad. Sci. (USA)*, **100**, 119 (2003). Automatic Classification of Protein Structure by Using Gauss Integrals.

160. J. H. White, *Am. J. Math.*, **91**, 693 (1969). Self-Linking and Gauss-Integral in Higher Dimensions.

161. O. Gonzalez and J. H. Maddocks, *Proc. Natl. Acad. Sci. (USA)*, **96**, 4769 (1999). Global Curvature, Thickness, and the Ideal Shapes of Knots.

162. A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar, *Nature* (*London*), **406**, 287 (2000). Optimal Shapes of Compact Strings.

163. J. R. Banavar, A. Maritan, C. Micheletti, and A. Trovato, *Proteins: Struct. Func. Genet*., **47**, 315 (2002). Geometry and Physics of Proteins.

164. J. R. Banavar, A. Maritan, and F. Seno, *Proteins: Struct. Func. Genet*., **49**, 246 (2002). Anisotropic Effective Interactions in a Coarse-Grained Tube Picture of Proteins.

165. J. R. Banavar and A. Maritan, *Rev. Modern Phys*., **75**, 23 (2003). Colloquium: Geometrical Approach to Protein Folding: A Tube Picture.

166. J. R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan, and A. Trovato, *J. Phys: Cond. Matter*, **15**, S1787 (2003). Tubes near the Edge of Compactness and Folded Protein Structures.

167. T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, *Proc. Natl. Acad. Sci.* (*USA*), **101**, 7960 (2004). Geometry and Symmetry Presculpt the Free-Energy Landscape of Proteins.

168. C. A. Ouzounis, R. M. R. Coulson, A. J. Enright, V. Kunin, and J. B. Pereira-Leal, *Nature Rev. Genet*., **4**, 508 (2003). Classification Schemes for Protein Structure and Function.

169. A. D. Michie, C. A. Orengo, and J. M. Thornton, *J. Mol. Biol*., **262**, 168 (1996). Analysis of Domain Structural Class Using an Automated Class Assignment Protocol.

170. S. Jones, M. Stewart, A. Michie, M. B. Swindells, C. Orengo, and J. M. Thornton, *Protein Sci*., **7**, 233 (1998). Domain Assignment for Protein Structures Using a Consensus Approach: Characterization and Analysis.

171. L. Holm and C. Sander, *Science*, **273**, 595 (1996). Mapping the Protein Universe.

172. S. Dietmann and L. Holm, *Nature Struct. Biol*., **8**, 953 (2001). Identification of Homology in Protein Structure Classification.

173. L. Holm and C. Sander, *Nucl. Acids Res*., **26**, 316 (1998). Touring Protein Fold Space with DALI/FSSP.

174. L. Holm and C. Sander, *Nucl. Acids Res*., **22**, 3600 (1994). The FSSP Database of Structurally Aligned Protein Fold Families.

175. C. Hadley and D. T. Jones, *Structure*, **7**, 1099 (1999). A Systematic Comparison of Protein Structure Classifications: SCOP, CATH and FSSP.

176. R. Day, D. A. C. Beck, R. S. Armen, and V. Daggett, *Protein Sci*., **12**, 2150 (2003). A Consensus View of Fold Space: Combining SCOP, CATH, and the DALI Domain Dictionary.

177. A. Harrison, F. Pearl, R. Mott, J. Thornton, and C. Orengo, *J. Mol. Biol*., **323**, 909 (2002). Quantifying the Similarities within Fold Space.

178. G. Apic, J. Gough, and S. Teichmann, *J. Mol. Biol*., **310**, 311 (2001). Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes.

179. S. Teichmann, A. G. Murzin, and C. Chothia, *Curr. Opin. Struct. Biol*., **11**, 354 (2001). Determination of Protein Function, Evolution and Interactions by Structural Genomics.

180. B. Rost, *Curr. Opin. Struct. Biol*., **12**, 409 (2002). Did Evolution Leap to Create the Protein Universe?

181. G. M. Downs and J. M. Barnard, in *Rev. Comput. Chem*., K. B. Lipkowitz, R. Larter, and T. Cundari, Eds., Wiley-VCH, New York, 2002, Vol. **18**, pp. 1–40. Clustering Methods and Their Uses in Computational Chemistry.