Protein Structure Classification

Patrice Koehl, Department of Computer Science and Genome Center,
University of California, Davis,
One Shields Avenue, Davis, 95616, USA

e-mail: koehl@cs.ucdavis.edu
URL: http://www.cs.ucdavis.edu/~koehl

Abstract

Years of research in biology have established that all cellular functions are deeply connected to the shape of their molecular actors. As a response, structural molecular biology has emerged as a new line of experimental research focused on revealing the structure of 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. While these studies are generating a wealth of information, stored into protein structure databases, the key to their success lies in our ability to organize and analyze the information contained in these databases, and integrate it with other biological efforts aimed at solving the mysteries behind cell functions. In this survey, I focus on the first step behind any such organization scheme, namely the classification of protein structures. I review the properties of protein structures, with a special interest on their geometry. Computer methods for the automatic comparison and classification of these structures are then reviewed. In parallel, I describe the existing classifications of protein structures, and their applications in biology, with a special focus on computational biology. I conclude the review with a discussion on the future of these classifications.

Introduction

The molecular basis of life rests on the activity of large biological macro-molecules, including nucleic acids (DNA and RNA), carbohydrates, lipids and proteins. While each play an essential role, there is something special about proteins, as they are the active actors of cellular functions. In this paper, I describe the growing interest in unraveling the mysteries behind their functions, focusing on the effort of organizing the information obtained from structural studies of proteins. Firstly, I briefly relate this effort to the continuous developments of scientific classification in biology.

Classification and biology.

Classification is a very broad term which simply means putting things in classes. Any organizational scheme is a classification: objects can be sorted with respect to size, colors, origins, ... Classification is one of the most basic activities in any science, probably because it is easier to think about a few groups than it is to think about a whole population. Scientific classification in biology probably started with Aristotle, in the 4th century B.C. He divided all livings things into two groups, animal and plants. Animals were themselves divided into two groups, those with blood, and those without (at least no red blood), while 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, though logical way that made it easy to convey scientific information. Among these biologists, it is worth citing the Swedish naturalist Corolus Linnaeus from the 18th century who set formal rules for a two name system called the binomial system of nomenclature, which is still used today. However, 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, that is species should be related based on a shared history. *Systematic classifications* were introduced accordingly, whose aims 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 the genome sequencing projects and the tools of bio-informatics are obviously 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 takes into account the products of these genes, and their contributions to the mechanisms of life. As function is related to shape, this is where protein structure classification will play a significant role in our understanding of the organization of life. Paraphrasing Jacques Monod, it is in the protein that lies the secret of life (1).

The biomolecular revolution.

All living organisms can be described as arrangements of cells, the smallest units capable of carrying functions important for life. Cells can be divided into organelles, which are themselves assemblies of bio-molecules. These bio-molecules are usually polymers 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 form, from unicellular organisms to complex multi-cellular species like us. Unraveling the secrets behind this hierarchy has become one of the major challenges of the twentieth and now twenty-first centuries. While physics and chemistry have provided significant insight into the structure of the atoms and their arrangements in small chemical structures, the focus now is set on understanding the structure and function of bio-molecules. These usually large molecules serve as storage

for the genetic information (the nucleic acids such as DNA and RNA), and as key actors of cellular functions (the proteins). Biochemistry, the field that studies these bio-molecules, is currently experiencing a major revolution. In hope of deciphering the rules that define cellular functions, large scale experimental projects are performed as collaborative efforts involving many laboratories in many countries. The main aims of these projects are 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, 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, corresponding to a database of over a million gene sequences (see http://www.genomesonline.org/ (2)) , and more than a thousand other genomes are currently being sequenced. The need to store this data efficiently and to analyze its contents has led to the emergence of a collaborative effort between computer science and biology, referred to as bio-informatics. In parallel, the repository of bio-molecular structures (3, 4) contains more than 27,600 structures of proteins and nucleic acids. The similar 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 bio-geometry. The combined efforts of 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 seen as a consequence of a paradigm shift in molecular biology (5), as the classical approach of hypothesis-driven research in biochemistry is being replaced with a data-driven discovery approach. I believe that in fact the two approaches co-exist, and that both benefit from these computer-based disciplines.

Outline. 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. I then describe the existing protein structure classifications, focusing on SCOP, CATH, and the DALI domain classification. Finally I conclude the paper with a discussion of the future of protein structure classifications.

Basic principles of protein structure

While all bio-molecules play an important part in life, there is something special about proteins, which are the products of the information contained in the genes. A perhaps surprising finding that crystallized over the last handful of 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 text books, such as those of Schulz and Schirmer (6), Cantor and Schimmel (7), of Branden and Tooze (8) and of Creighton (9). I also refer the reader to the excellent review of Taylor and collaborators (10).

Visualization.

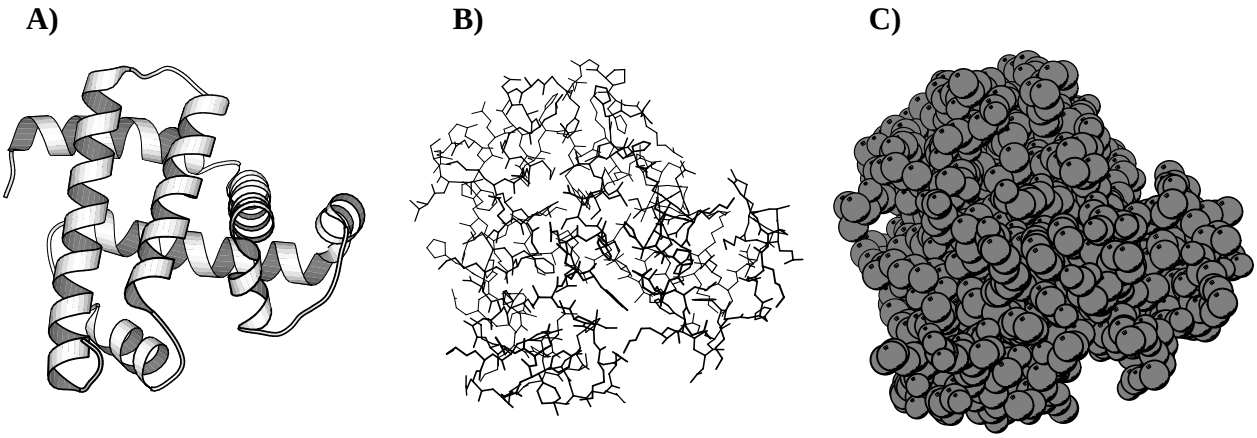


Figure 1: **Visualizing protein structures**. Myoglobin is a small protein very common in muscle cells, where it serves as oxygen storage. Its structure was determined by X-ray crystallography as early as 1960 by John Kendrew and his collaborators (13). It was in fact the first protein structure available. Here I show the structure of sperm whale myoglobin using three different types of visualization. For simplicifity, I do not show the heme. The coordinates are taken from the PDB file 1mbd. (A) **Cartoon**. This representation 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. It emphasizes the chemical nature of the molecule (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 (14).

The need for visualizing bio-molecules is based on the 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 have a feeling for these molecules. These models, usually made out of painted wood, plastic, rubber and/or metal were designed to highlight different properties of the molecule under study. In the space-filling models, such as those of Corey-Pauling-Koltun (CPK) (11, 12), 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 the skeletal models, chemical bonds are represented by rods, whose junctions define the position of the atoms. These models were used for example by

Kendrew and colleagues in their studies of myoglobin (13). They are useful to the chemists by highlighting the chemical reactivity of the bio-molecules and, consequently, their potential activity. With the introduction of computer graphics to structural biology, the principles of these models have been translated into software such that molecules could be visualized on the computer screen. Figure 1 shows examples of computer visualizations of myoglobin, including space-filling and skeletal representations. Many computer programs are now available that visualize bio-molecules. I only cite here MOLSCRIPT (14) and VMD (15), which have been used to generate most of the figures of this paper.

Protein Building blocks.

Proteins are heteropolymer chains of amino acids, often referred to as *residues*. This term comes from chemistry and describes the material found at the bottom of a reaction tube once a protein has been cut into pieces in order to determine its composition. There are twenty naturally occurring amino acids that make up proteins. With the exception of proline, amino acids have a common structure, shown in figure 2A. Naturally occurring amino acids that are incorporated into proteins are, for the most part, the levorotary (L) isomer. Substituants on the alpha carbon, i.e. *side-chains*, range in size from a single hydrogen atom to large aromatic rings and can be charged or include only non-polar saturated hydrocarbons (16); see table 1 and figure 2B.

Classification	**Amino acid**
Non polar	glycine (G), 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)

Table 1: **Classification of the 20 amino acids** based on their interaction with water (16). The one-letter code of each amino acid is given in parenthesis. Non polar amino acids do not have 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 NH2 accepts a proton to become NH3+ and the carboxyl group COOH donates a proton and becomes COO-.

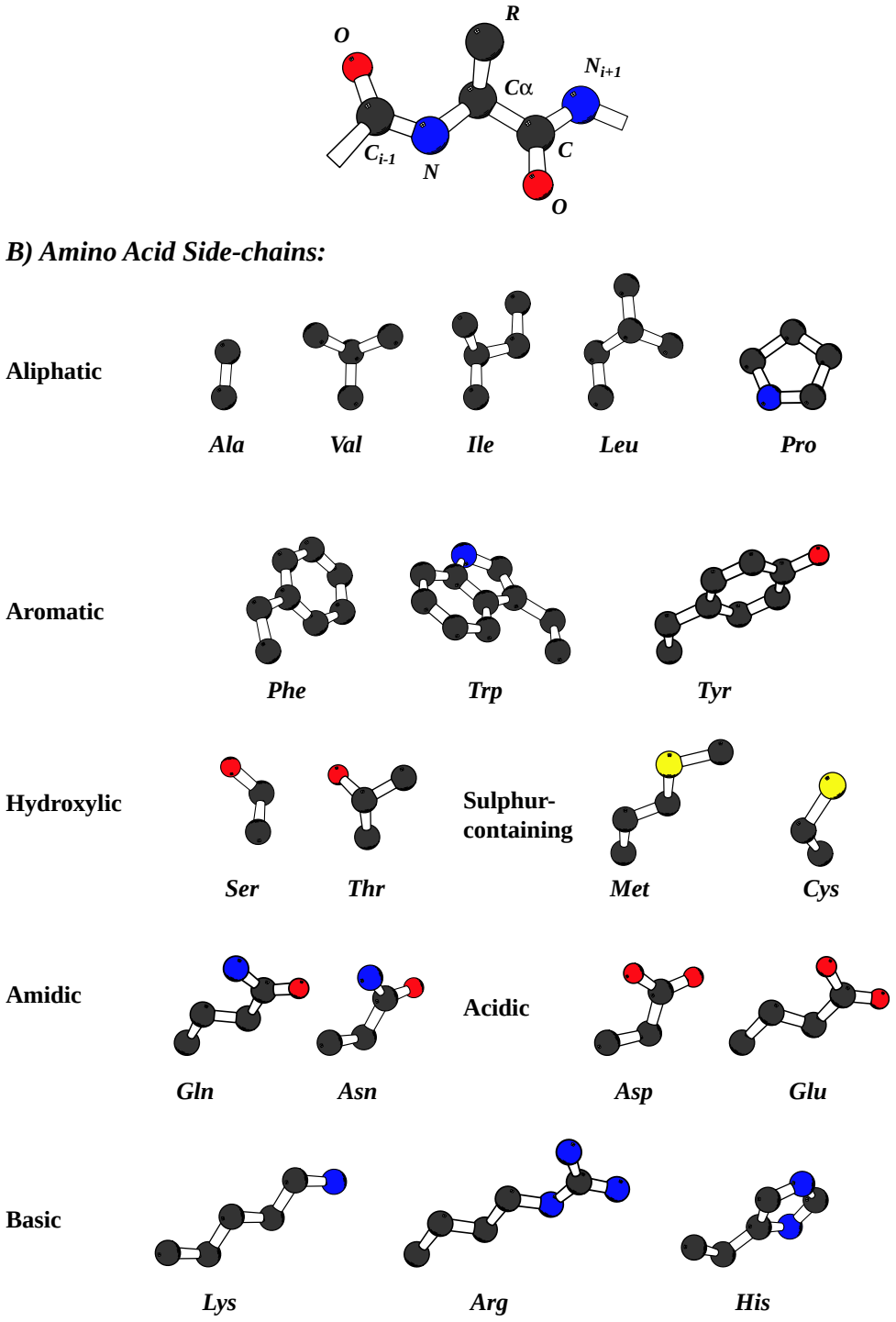


Figure 2: **The twenty natural amino acids that make up proteins.** (A) Each amino acid has a main-chain (N, Cα, C and O) on which is attached a side-chain schematically represented as R. Amino acids in proteins are attached through planar peptide bonds, connecting atom C of the current residue to atom N of the following residue. For sake of simplicity, I omit the hydrogens. (B) Classification of the amino acids side-chains R according to their chemical properties. Glycine (Gly) is omitted, as its side-chain is a single H atom. Figure drawn using Molscript (14).

Protein Structure Hierarchy.

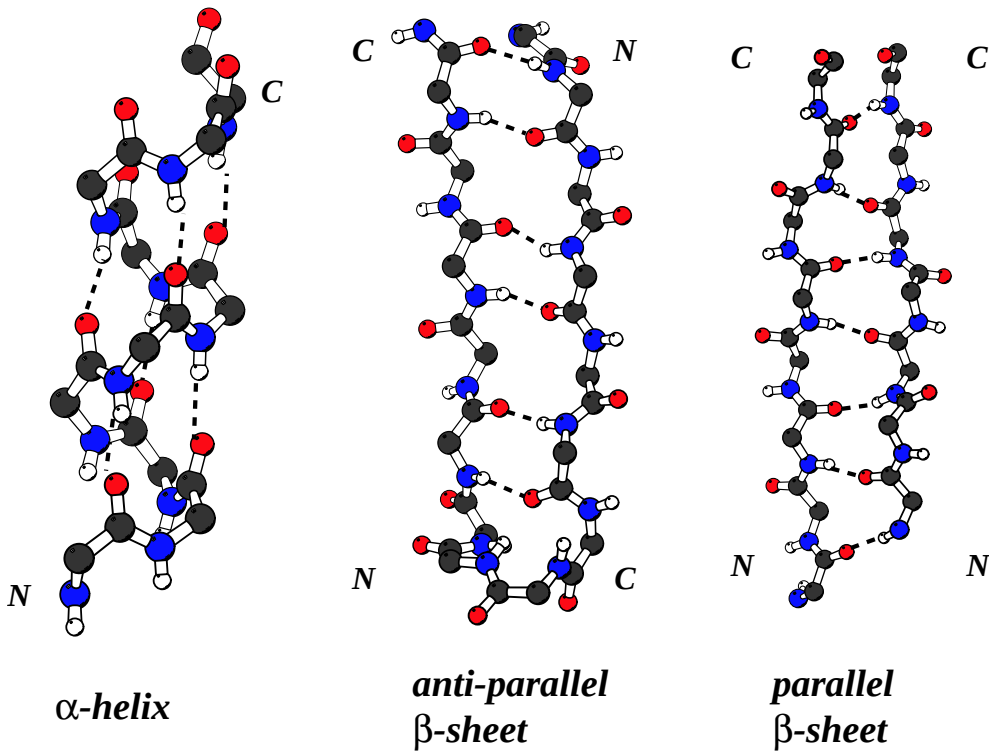


Figure 3: **The three main secondary structure elements (SSE) found in proteins**. For simplicity, side-chains and non-polar hydrogens are ignored. The protein backbone is shown with balls and sticks, and hydrogen bonds are shown as discontinuous lines. (A) The regular α-helix is a right handed helix, in which all residues adopt similar conformations, with the backbone torsion angles φ and ϕ close to -60 and -40, respectively. 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 bonds C=O and N-HN are parallel to the main axis of the helix. (B) An anti-parallel β-sheet. Two strands (stretches of extended backbone segments, with φ and ϕ close to -120 and 120, respectively) are running in an anti-parallel geometry. The atoms HN and O of residue *i* in the first strand are involved in hydrogen bonds with the atoms O and HN of residue *j* in the opposite strand, respectively, while residues i+1 and j+1 face outwards. (C) A parallel β-sheet. The two strands are parallel, and the atoms HN and O of residue *i* in the first strand are involved in hydrogen bonds 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 outwards is observed in parallel and anti-parallel β-sheets. A strand can therefore be involved in two different sheets. Figure drawn using Molscript (14).

Condensation between the -NH3+ and the -COO- groups of two amino acids generates a peptide bond and results in the formation of a dipeptide. Protein chains correspond to an extension of this chemistry, resulting in long chains of many amino acids bonded together. The order in which amino acids appear defines the *primary sequence* or *primary structure* of the protein. In its native environment, the polypeptide chain adopts a unique three-dimensional shape, referred to as the *tertiary* or *native structure* of the protein (17). 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). The former is a right handed helix with 3.6 aminoacids per turn, while the latter is an approximately planar layout the backbone. Helices often pack together to form a hydrophobic core, while β-strands pair together to form parallel, or antiparallel β-sheets . Note that in addition to these two types

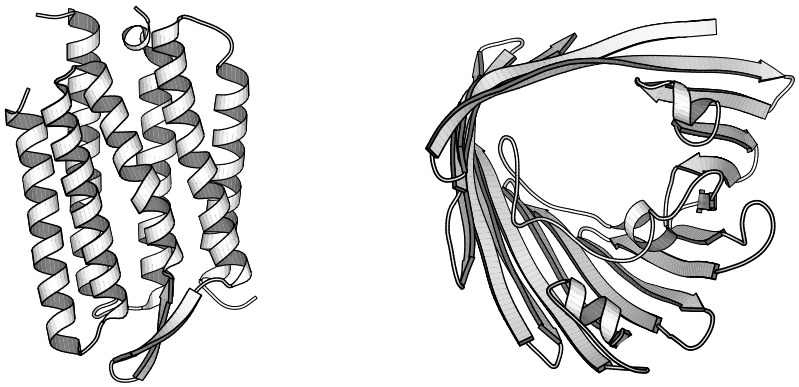
of secondary structures, there is a wide variety of other commonly occurring sub-structures, referred to as *super-secondary structure*. More information on these sub-structures can be found in the work of Efimov (18-21).

Three types of proteins.

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

Fibrous proteins are elongated molecules in which the secondary structure forms the dominant structure. They are insoluble, play a structural or supportive role in the body, and are also involved in movement (such as in muscle and ciliary proteins). Fibrous proteins often 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 alanine and serines. In collagen, the major protein component of connective tissue, every third residue is a glycine, and many of the others are prolines.

Membrane proteins are restricted to the phospho-lipid bilayer membrane that surrounds the cell and many of its organelles. These proteins cover a large range, 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 through the membrane, ranging from simple ions to nutrients. The structures of fully embedded membrane proteins can be classified into two major categories: the all helical structures, such as bacteriorhodopsin, and the all beta structures, such as porins (see figure 4). Note that as of October 2004, there are 158 structures of membrane proteins in the PDB, out of which 86 are unique (see http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html).



(a) Bacteriorhodopsin (b) Porin

Figure 4: **Two examples of membrane proteins**. (a) Bacteriorhodopsin (PDB code 1C3W) is a mainly α-protein, containing seven helices. It is a membrane protein serving as an ion pump, and found in bacteria that can survive in high salt concentration. (b) Porin (PDB code 2por) 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 (14).

Globular proteins have a unique structure derived from a non repetitive sequence. They range in size from hundred to several hundred residues, and adopt a compact structure. In globular proteins, non-polar amino acids have a tendency to re-group and form the core of the proteins, while polar amino acids remain accessible to the solvent. In the tertiary structure, β-strands are usually paired in parallel or anti-parallel arrangements, to form β-sheets. On average, the protein main-chain consists of about 25% of residues in α-helix formation, 25% of residues in β-strands, with the rest of the residues adopting less regular structural arrangements (22).

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:800/srs5/
DSSP	Database of secondary structures of proteins (available through SRS)	http://srs.embl-heidelberg.de:800/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/

Table 2: Resources on protein structures

Geometry of globular proteins.

From the seminal work of Anfinsen (23), we know that the sequence fully determines the three-dimensional structure of the protein, which itself defines its function. While the key to the decoding of the

information contained in genes was found more than fifty years ago (the genetic code), we have not yet found the rules that relate a protein sequence to its structure (24, 25). Our knowledge of protein structure therefore comes from years of experimental studies, either using X-ray crystallography or NMR spectroscopy. The first protein structures to be solved were those of myoglobin and hemoglobin (13, 26). Currently (October 2004), there are nearly 27,700 protein structures in the PDB database (3, 4) of bio-molecular structures; see http://www.rcsb.org. (Note that this numbers overestimates the number of different structures available as 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). Table 2 lists the web addresses of protein structure databases and the resources available for analyzing these structures.

As there are only two types of secondary structures (α and β), proteins can be divided into three main structural classes (27): mainly α proteins (28), mainly β proteins (29-31), and mixed α - β proteins (32). A fourth class includes proteins with little or no secondary structures at all, which are stabilized by metal ions and/or disulphide bridges. There has been significant effort put into classifying protein structures into their main folding class automatically: these efforts will be reviewed in the next section. In parallel, there has been significant work on predicting a protein folding class based on its sequence. More details can be found in (33-40).

The mainly α class, the smallest of all 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 is the four helix bundle structure (see figure 5). The most studies α structure is the globin fold, which as 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 (13).

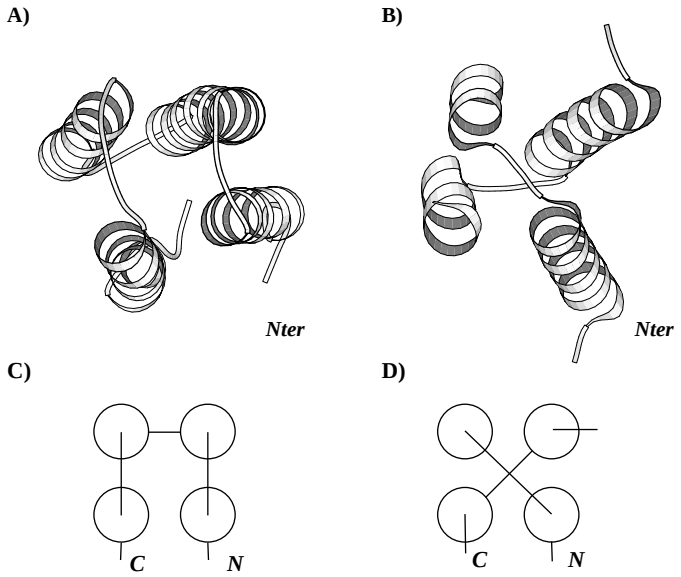


Figure 5: **Two different topologies of four helix bundles**. A bundle is an array of α-helices, each oriented roughly along the same (bundle) axis. A and C show a four helical, up-and-down bundle with a left handed twist, observed in hemerythrin from a sipunculid worm (PDB code 2hmz). B and D show a four helix bundle with a right handed twist, observed in a fragment of the dimerization domain of a liver transcription factor (PDB code 1g2y). A and B are cartoon representations of the proteins obtained with MOLSCRIPT (14), while C and D show the schematic topologies produced by TOPS (http://www.tops.leed.ac.uk/).

The mainly β class contains the parallel and antiparallel β structures. In these, the β strands are usually arranged in two β sheets that pack against each other and form a distorted barrel structure. There are three major types of β barrels, the up-and-down barrels, the Greek key barrels (41), and the jelly roll barrels (see figure 6). Most of the known antiparallel β structures, including the immunoglobulins have barrels that include at least one Greek key motif. The two other motifs are observed in proteins of quite diverse function, where functional diversity is obtained by differences in the loop regions that connect 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. This leads to a fairly rigid classification scheme (42), which is quite sensitive to the definition of hydrogen bonds and β-strands.

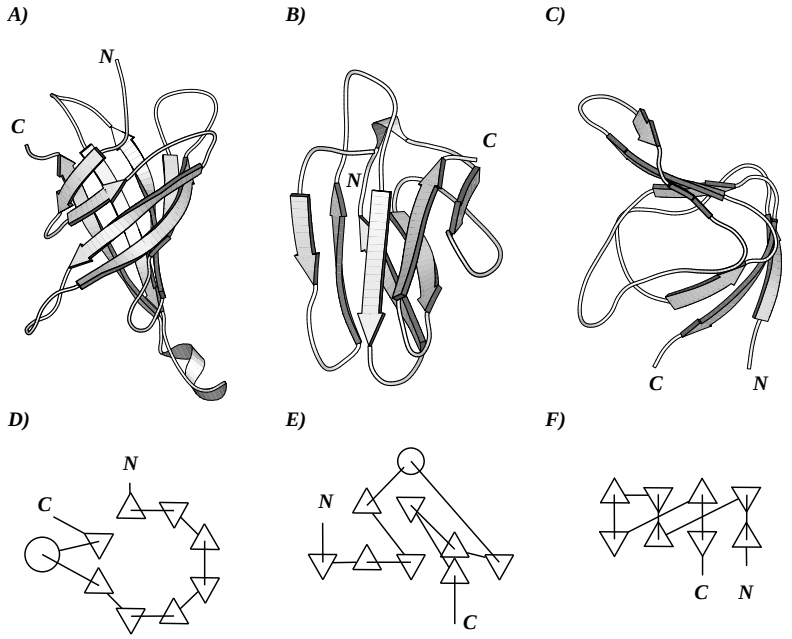


Figure 6: **Three common sandwich topologies of beta proteins**: a meander (A and D) observed in a glycoprotein from chicken (PDB code 2cam), a Greek key (B and E) observed in an α-amylase (PDB code 1bli), and a jelly roll (C and F) observed in a gene activator protein from *E. Coli* (PDB code 1g6n). A meander (or up-and-down) is a simple topology in which any two consecutive strands are adjacent and anti parallel. A Greek key motif is a topology of a small number of b-sheet strands in which some inter-strand connection exist between b-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 (14), while D, E and F show the schematic topologies produced by TOPS (http://www.tops.leed.ac.uk/).

The α-β protein class is the largest of all three classes. It can be subdivided into proteins that have a mainly alternating arrangement of α helices and β strands along the sequence, and those that have more segregated secondary structures. The former class can be itself divided into two groups: one with a central core of often eight parallel β strands arranged together into a barrel surrounded by α helices, and a second group that comprises an open, twisted parallel or mixed β sheet, with α helices on both side (see figure 7). A particularly striking example of α-β barrel is seen in the eight-fold β-α barrel $(\beta\alpha)_8$ which was found originally in the triose phosphate isomerase of chicken (43), and is consequently often referred to as the TIM-barrel (for a complete analysis, see(44-51)). Many of the proteins adopting a TIM barrel

structure have completely different amino acid sequences and different functions. The open α/β-sheet structures vary considerably is size, number of β strands, and their strand order.

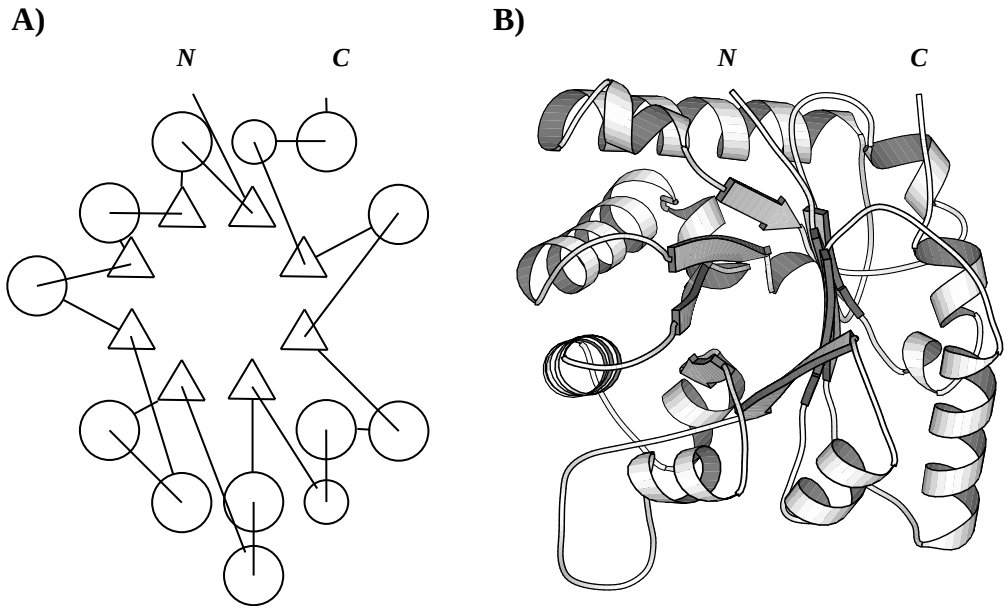


Figure 7: **Topology (A) and cartoon representation (B) of the TIM barrel**. The protein chain alternates between β and α secondary structure type, giving rise to a barrel β-sheet in the center surrounded by a large ring of a-helix on the outside. This structure, first seen in the triose phosphate isomerase of chicken ((PDB code 1tim, after which it is often name TIM barrel), 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 (14).

Protein domains.

Large proteins do not contain a single large hydrophobic core, probably because of limitations in the folding kinetics and stability. Single compact units of more than 500 amino acids are rare. Large proteins in fact are organized into "units" with sizes around 200-300 residues, referred to as *domains* (52-54). For a detailed analysis of domains in proteins, see (55). Domains are defined simultaneously as: (a) regions that display a significant level of sequence similarity; (b) the minimal part of a gene that is capable of performing a function; (c) a region of a protein with an experimentally assigned function; (d) region of a structure that recurs in different contexts in different proteins; and (e) compact, spatially distinct units of protein structure. As more structures of proteins are solved, contradictions in these definitions appear. Some domains are compact while others are clearly not globular. Some are too small to form a stable domain, and lack a hydrophobic core. Currently, we are in the awkward situation in which the concept of structural domain is well accepted, yet its definition remains ambiguous (56). This will be discussed in details in the next section.

Resources on protein structures

All experimental protein structures available today are stored in the Protein Databank (PDB) (3), 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 in order to ease access to the information in contains. For example, the services "PDB at a glance" and "Molecules to Go" were designed as easy-to-use interfaces to the PDB with simple search engines. The MSD search 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 (57, 58) and the Biotech Validation Suite are services from which quality control programs can be run to check the quality of a protein structure. NRL, Entrez and SRS are integrated services that regroup the PDB with other databases on proteins. For example, SRS includes DSSP (59), a database of secondary structures of proteins. PISCES (60) and ASTRAL (61-63) can generate subsets of the PDB database, based on the user's criteria. Table 2 lists the web addresses of all these services.

Protein structure comparison

Any attempts to study a large collection of objects will usually start with classifying them according to a given measure of similarity. This is probably a consequence of the fact that it is easier to deal with a few representatives than to deal with a whole population. Protein structure similarity is most often detected and quantified by a protein structure alignment program, applied to the different domains of the proteins considered. In this section, I review existing techniques for automatically detecting domains in protein structures, as well as techniques for finding the optimal alignment between two structural domains. I conclude with a brief description of new techniques for comparing protein structural domains that do not rely on a structural alignment, but on a direct comparison of the topology of the domains.

Automatic identification of protein structural domain.

Decomposition of multi-domain protein structures into individual domains has been traditionally done manually. As the rate of protein structure determination has increased drastically in the past few years, this manual process has become a bottleneck in maintaining and updating protein structure classifications. There is a need consequently for automation. Automatic decomposition of proteins into structural domains can be traced back to the work of Rossman and Liljas in 1974 (64), 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. Analysis of the distance plot however required human intervention. Crippen (17) generalized this concept, using hierarchical cluster analysis to protein fragment-fragment contacts. This procedure generates a tree of protein fragments, from small, locally compact region to the complete protein. Several methods have been subsequently proposed, that follows this concept of identifying domains based on a difference between intra-domain and inter-domain properties. These properties often refer to distances (intra domain distances between residues are usually shorter than inter domain distances (65-68), contact surface area between domains (69, 70), "compactness" (52, 71, 72), or dynamics (73). To find the cutting points in a protein chain that delineate domains, recursive algorithms have been developed which either scan the chain to find single cuts such that the two resulting fragments very a given protein domain definition based on one of the properties enumerated above, or directly look for multiple cuts (see for example (68)). This problem has also been formulated as en eigenvalue problem on the $C\alpha$-$C\alpha$ distance matrix (73), or as a network flow problem (74, 75). The methods described above 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 to recognize 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 (76) 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. Briefly, his procedure works as follows. Each residue in the protein chain is assigned a numeric label, usually the sequential residue number itself. If a residue i with label s_i is surrounded by neighbors with, on average, a higher label, then its label increases, otherwise it decreases. This procedure is iterated until the system reaches equilibrium. Special care is taken to ensure that the protein chain does not pass too frequently between domains, that secondary structures, in particular β-sheets are not broken, and that small domains are either ignored or avoided. For full details, see (76). Swindells developed an alternative "bottom-up" approach, in which he first identifies core regions in the protein (77), which are then extended to define the different domains in the proteins (78). Most of these methods include a refinement scheme to assess the quality of the domains that have been identified, based on their accessible surface area , hydrophobic moment profile, size of the

domain, dynamics between domains, compactness, number of protein segments (75), and presence of intact β sheets (76).

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

Table 3: Web sites for publicly available services and/or programs for protein domain assignment

The diversity in the definitions of protein structural domains these domains is a serious issue for the generation of protein structure classifications. Many programs have been developed to delineate domains automatically in multi-domain proteins. In table 3, I list the programs that are currently accessible on the web, either as a web service, or available for download. While these programs agree on most cases, the existence of discrepancies still prevents consistent assignments of protein domains (56). The absence of quality control on the results of the protein domain assignment programs has led the developers of protein structure classifications to use a combination of automatic and manual methods. For example, CATH (79) defines domains in multi-domain proteins based on a consensus of three automatic programs, namely PUU (73), DOMAK (80) and Detective (78). 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 and/or on the web. In fact, several structural domain databases are available on the web to assist manual assignments of domains (see table 4).

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

Table 4: Databases of protein structural domains

The rigid body transformation problem

Definition

I start with the (relatively) easier problem of comparing two protein structures with the same number of atoms and a known correspondence table between these atoms (for review, see (81)). 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 it still creates some confusion on how it can be solved (82), I present here a full mathematical description of the problem, as well as a proof for one of its closed form solution. The problem of comparing two different models of a protein can be formalized as:

Rigid Body Transformation Problem: given two sets of points $A=(a_1, a_2, \ldots, a_n)$ and $\mathbf{B}=(b_1,b_2,\ldots b_m)$ in three dimensional space and assume 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, i.e.

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

When comparing two proteins, the sets of points can include the Cα 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, or cRMS, defined as follows:

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

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: $\Re^3 \to \Re^3$ that satisfies the properties:

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

and

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

where $\wedge$ is the cross product.

Equation [3] states that distances are conserved, while equation [4] says that internal reflection are 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 minimum.

A closed form solution based on SVD

In the literature, there exist a large number of algorithms that solve the rigid transposition problem, coming from various fields including computer vision and image processing, robotics, astronomy and computational biology. They differ with respect to the representation of the transformation, and the minimization procedure. Some of these algorithms are based on closed form solutions, while others use iterative solutions. For detailed descriptions of these algorithms, including comparison of their performances, I refer the readers to the surveys of Sabata and Aggarwal (83), Ferrari and Guerra (84), and

Eggert and colleagues (85). Here I focus on the representation classically used in computational biology, and briefly describe its background. It is based on the singular value decomposition (86) of a correlation matrix *C* between the two sets of points (87-90). This method appears to have been first derived by Schonenman in the context of factor analysis (91). Other approaches include solutions based on a power decomposition of *C* (92), or on a representation of rotations with quaternions (93-95). These methods have been shown to be equivalent (85, 95).

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

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

is minimum.

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

$$\frac{\partial \varepsilon}{\partial T} = -\frac{2}{n}\sum_{i=1}^{n}(a_i - Rb_i - T) = 0 \qquad [7]$$

so that

$$T_{min} = \frac{1}{n}\sum_{i=1}^{n}a_i - R_{min}\left(\frac{1}{n}\sum_{i=1}^{n}b_i\right) = \mu_A - R_{min}\mu_B \qquad [8]$$

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

Note that if the two sets of points are shifted such that their barycenters 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 $X = [x_1,x_2,\ldots,x_n]$ and $Y=[y_1,y_2,\ldots y_n]$ the 3xn 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

$$\varepsilon = \frac{1}{n}\left\|X - RY\right\|^2 \qquad [9]$$

is minimum.

Let *C* be the correlation matrix of *X* and *Y*:

$$C = XY^T \quad \rightarrow \quad C_{ij} = \sum_{k=1}^{n} x_{ik}y_{jk},\ i,j = 1,2,3, \qquad [10]$$

and UDV^T a singular value decomposition (86) of *C* ($UU^T=VV^T=I$, D= diag(d_i), $d_1 \geq d_2 \geq d_3 \geq 0$). Then the minimum value of ε with respect to *R* is

$$\varepsilon_{min} = \frac{1}{n}\left(\left\|X\right\|^2 + \left\|Y\right\|^2 - 2\left(d_1 + d_2 + \lambda d_3\right)\right) \qquad [11]$$

where λ = sign(det(*C*)). The optimal rotation is given by

$$R_{\min} = U \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \lambda \end{pmatrix} V^T \qquad [12]$$

when rank(*C*) ≥2.

This result was first formulated by Schöneman (91), later refined by Arun et al (90), Horn et al (92), and Umeyama (96). Here I follow the proof of Umeyama.

Finding a rotation matrix *R* that minimizes ε can be rewritten as finding a matrix *R* that minimizes the objective function *O* defined as:

$$O = \left\| X - RY \right\|^2 + tr\left(L\left(R^T R - I\right)\right) + g\left(\det(R) - 1\right), \qquad [13]$$

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 (96):

$$\frac{\partial O}{\partial R} = -2XY^T + 2RYY^T + 2RL + gR = 0 \qquad [14]$$

$$\frac{\partial O}{\partial L} = R^T R - I = 0 \qquad [15]$$

$$\frac{\partial O}{\partial g} = \det(R) - 1 = 0 \qquad [16]$$

From equation [14],

$$RM = XY^T = C \qquad [17]$$

where *C* is the covariance matrix defined in equation [10], and *M* is a symmetric 3x3 matrix defined by:

$$M = YY^T + L + \frac{g}{2} I \qquad [18]$$

Transposing equation [17], we obtain:

$$MR^T = C^T \qquad [19]$$

and multiplying each side of [17] with each side of [19], equation [20] is obtained, as *RTR=I* (equation [15]).

$$M^2 = C^T C = VD^2V^T \quad [20]$$

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

$$M = VDSV^T \quad [21]$$

where S = diag(s_i), s_i=1 or -1.

From equation [21],

$$\det(M) = \det\left(VDSV^T\right) = \det(D)\det(S) \quad [22]$$

and from equation [17]

$$\det(M) = \det(R^T)\det(C) = \det(C) \quad [23]$$

as det(R)=det(R^T)=1 (equation [16]).

Thus,

$$\det(D)\det(S) = \det(C) \quad [24]$$

Since 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:

$$\begin{aligned}\varepsilon &= \frac{1}{n} tr\left((X - RY)(X - RY)^T\right) = \frac{1}{n}\left(tr\left(XX^T\right) + tr\left((RY)(RY)^T\right) - 2tr(XY^TR^T)\right) \\ &= \frac{1}{n}\left(\|X\|^2 + \|RY\|^2 - 2tr\left(XY^TR^T\right)\right) = \frac{1}{n}\left(\|X\|^2 + \|Y\|^2 - 2tr(M)\right)\end{aligned} \quad [25]$$

Substituting equation [21] into equation [25], we have

$$\begin{aligned}\varepsilon &= \frac{1}{n}\left(\|X\|^2 + \|Y\|^2 - 2tr\left(VDSV^T\right)\right) \\ &= \frac{1}{n}\left(\|X\|^2 + \|Y\|^2 - 2tr(DS)\right) = \frac{1}{n}\left(\|X\|^2 + \|Y\|^2 - 2(d_1s_1 + d_2s_2 + d_3s_3)\right)\end{aligned} \quad [26]$$

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. This concludes the proof for equation [11].

Next, we determine a rotation matrix *R* achieving the above minimum value. When rank(*C*)=3, M is non singular, and its inverse is given by:

$$M^{-1} = (VDSV^T)^{-1} = VSD^{-1}V^T = VD^{-1}SV^T \quad [27]$$

and

$$R_{min} = CM^{-1} = UDV^TVD^{-1}SV^T = USV^T, \quad [28]$$

which completes the proof for equation [12]. Note that this expression for R_{min} is also valid when rank(*C*)=2 (see (96)).

Weighted superposition of sets of points.

It is not always judicious to give the same importance to all points of A and B. This has led to a variant of the rigid body transformation problem, in which each point i is given a weight ω_i. Examples of weighting schemes include considering the mass of the atoms included in the superposition, giving different weights to atoms of the backbone of the protein compared to atoms of the side-chains, and giving more 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 as

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

is minimum.

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

$$\frac{\partial\varepsilon'}{\partial T} = -\frac{2}{n}\sum_{i=1}^{n}\omega_i(a_i - Rb_i - T) = 0 \quad [30]$$

so that

$$T_{min} = \frac{1}{\Omega}\sum_{i=1}^{n}\omega_i a_i - R_{min}\left(\frac{1}{\Omega}\sum_{i=1}^{n}\omega_i b_i\right) = \mu'_A - R_{min}\mu'_B \quad [31]$$

where Ω is the sum of the weights ($\Omega = \sum_{i=1}^{n}\omega_i$), and μ'_A and μ'_B are the weighted barycenters 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 $x'_i = \sqrt{\omega_i}\left(a_i - \mu'_A\right)$ and $y'_i = \sqrt{\omega_i}\left(b_i - \mu'_B\right)$ be the weighted coordinates of the shifted points, and $X' = [x'_1, x'_2, \ldots, x'_n]$ and $Y=[y'_1, y'_2, \ldots y'_n]$ the 3xn 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

$$\varepsilon' = \frac{1}{n}\left\| X' - RY' \right\|^2 \qquad [32]$$

is minimum.

Equation [32] is equivalent to equation [9], and the same algorithm described above 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:

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

2) Compute weighted barycenters of A and B:

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

3) Generate weighted covariance matrix:

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

4) Compute SVD of C': $C'=UDV^T$ and λ = sign(det(C')); note that D = 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 = diag(1,1,λ), and optimal translation

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

6) Compute the cRMS between the two structures:

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

Note that this algorithm does not take into account the possible presence of noise in the coordinates of the points. In the case of proteins, the coordinates of atoms are approximations to a "true" position: proteins are flexible, fluctuation about a mean position. In addition, the physical experiments that provide information on the coordinates (usually X-ray crystallography and NMR spectroscopy) are noisy. 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 in the two models. The presence of noise will be even more important for the superposition of two proteins that can have different lengths.

Protein structure superposition

An ambiguous problem

The problem of finding an optimal alignment between two proteins is more complex than the rigid body transformation problem, as the correspondence, i.e. the list of equivalent residues in the two proteins is not known and in fact is part of the desired output, with the optimal transformation of the position of one protein with respect to the other. The protein structure alignment problem can be stated in fact as finding the maximal substructures of the two proteins that exhibit the highest degree of similarity.
A "substructure" of a 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,…,p_k) where $1 \leq p_1 < p_2 \ldots < p_k \leq n$, by A(P)=($a_{p1}$,$a_{p2}$,…,$a_{pk}$). The length |A(P)| of A(P) is the number of points it contains, i.e. *k*. A "gap" in A(P) is two consecutive indices p_i, p_{i+1} such that $p_i + 1 < p_{i+1}$.

Protein Structure Superposition Problem: 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, find the optimal subsets *A(P)* and *B(Q)* with |A(P)|=|B(Q)|, and find 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.

$$\min_G \{D(A(P) - G(B(Q)))\} \qquad [33]$$

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: this is the rigid body transformation problem, described in detail above. The concept of optimal correspondence however requires more attention. 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".

Though significant progress has been made over the past decade, a fast, reliable and convergent method for protein structural alignment is not yet available (97). Recent developments have focused both on the search algorithm and on defining the target function to be minimized, that is, a quantitative measure of the "similarity" between two structures. The most direct approach to the comparison of two protein structures is to move the set of points representing one structure as a rigid body over the other, and look for equivalent residues. This can only be achieved for relatively similar structures and will fail to detect local similarities of structures sharing common substructures. To avoid this problem, the structures can be broken into fragments (usually secondary structure elements [SSEs]), but this can lead to situations in which the global alignment can be missed. 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 methods such as dynamic programming (98, 99), comparison of distance matrices (100-103), fragment matching (104, 105), geometric hashing (106-111), maximal common subgraph detection (112-114) and local geometry matching (115). Optimization of this equivalence set is performed using dynamic programming (99, 116-118), Monte Carlo algorithms or simulated annealing (119), a genetic algorithm (120), incremental combinatorial extension of the optimal path (121, 122) and mean-field approaches (123, 124). Excellent reviews of these and other methods can be found in refs (10, 97, 125, 126).

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		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
SHEBA	http://rex.nci.nih.gov/RESEARCH/basic/lmb/mms/sheba.htm	http://rex.nci.nih.gov/RESEARCH/basic/lmb/mms/SHEBA-download.htm	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

Table 5: Web sites for publicly available protein structure alignment services and programs

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 itself is accessible for download, either as an executable or as a full source package (table 5). These are wonderful tools and I do encourage the reader to test several of these sites. Many of these services have been tested on large datasets with known similarities (126-128). These comparison studies do not identify a clear "winner", i.e. a technique that is significantly better than the other. In fact, it appears that a technique that combines existing algorithms performs better than the individual techniques (128). In the following I will review the different definitions given to the similarities of two structures, and will describe in detail two methods for protein structure alignment, one based on distance matrices (DALI) (102, 129), and one based on dynamic programming and comparison of structures in coordinate space (STRUCTAL) (116, 117). Finally I will describe recent progress in developing a closed form protein structure alignment algorithm.

Scoring functions for protein structure superposition

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

There are various measures of similarity between two sets of points. In the section on rigid body transformation, I have mentioned the cRMS value, which measures the root mean square deviation between the coordinates of the points of the two sets. For a given correspondence length, the cRMS can be minimized using a closed form algorithm (see above). When both cRMS and correspondence length need be optimized, there are no known closed form solutions. Approximate solutions usually based on heuristics do not in fact minimize the cRMS directly, as it is very sensitive to outliers (since it is based on the L2 norm). For example, Levitt and co-workers (116, 117) have introduced a scoring function with a Lorentzian shape:

$$ST(P,Q) = \max_{R,T} \sum_{i=1}^{p} \frac{20}{1 + \left\| a_{pi} - Rb_{qi} - T \right\|^2} - 10G_{P,Q} \qquad [34]$$

where the summation extends over the length of the correspondence between *A(P)* and *B(Q)*, and $G_{P,Q}$ is the total number of gaps in *A(P)* and *B(Q)*. R and T are the optimal rotation and translation that achieve a maximum of the score (as opposed to reaching a minimum for cRMS, see equation [2]).

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

$$dRMS = \left[\frac{2}{p(p-1)} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} M\left(\left\| a_i - a_j \right\|, \left\| b_i - b_j \right\| \right) \right]^{1/2} \qquad [35]$$

where p is the cardinality of the two sets.

Interestingly, there is no consensus on the definition of the metric *M* used to compare the two internal distances $\|a_i\text{-}a_j\|$ and $\|b_i\text{-}b_j\|$. When comparing two pairs of atoms between two structures, Taylor and Orengo (98) 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 (102 defined a similarity score as (e-[D/<D>])exp(-[<D>/f]2), where <D> is the average of the two intramolecular distances. Rossmann and Argos {Rossmann, 1976 #135), and Russell and Barton (130) used a score exp*(-[D/e]2)*exp*(-[S/e]2)*, where *S* takes into account local neighbors for each pair of atoms. At this stage, there is no clear evidence as to which score performs best.

All the scores cited above use geometry for the comparison, ignoring similarities in the environment of the residues. Suyama et al. (131) proposed another approach in which they ignored the 3D geometry altogether and compared structures on the basis of 3D profiles (132) alone, using dynamic programming. These profiles include information on solvent accessibility, hydrogen bonds, local secondary structure states and side-chain packing. Although this method is able to align two-domain proteins with different relative orientations of the two domains, it often generates inaccurate alignments (131). Jung and Lee (133) recently improved upon this method by iteratively refining the initial profile alignment using dynamic programming and 3D superposition. Their method, referred to as SHEBA, was found to be fast and as reliable as other alignment techniques (though it was only tested on a small number of protein pairs). Kawabata and Nishikawa (134) 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 using a Markov transition model that is equivalent to the Dayhoff's substitution model for amino acids. Three types of scores were considered: a score based on accessibility to solvent; a residue-residue distance score; and an SSE score.

Superposition based on internal distance matrices: DALI

Associated with every protein chain A of n atoms is an n x 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 distance map of A. The two representation 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. Reversely, it is known that the coordinates of the atoms of the protein can be recovered from the distances matrix, using distance geometry (135, 136). The recovered atomic coordinates are the original ones, modulo a rigid transformation (and possible a mirror transformation). This equivalence between coordinates and internal distances has lead 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, in that it bypasses the need to find an optimal rigid transformation that superposes the two structures. As a consequence, many algorithms have been proposed that compare internal distances matrix to align protein structures. The most commonly used of these algorithms is DALI, which is briefly described below.

Holm and Sander (102) developed a two stage procedure, DALI (Distance ALIgnment algorithm) which 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 is generated for each hexapeptide, which contains all its internal distances. 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/<D>])*exp*(-[<D>/f]2)* where *D* is the difference between the two distances to be compared, *<D>* is the average distance, and *e* and *f* are parameters. This score is less sensitive to distortion for long range distances. For sake of efficiency, only

hexapeptide pairs having similar backbone conformation are compared. Hexapeptides whose contact maps match above a given threshold are stored in lists of fragment equivalences.

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

This method, implemented in the program DALI (129), has been used to compare representatives from all the non-homologous (in sequence) families in the protein data bank (3, 4). See the section on the Dali Domain Classification below for details.

Superposition based on cRMS: STRUCTAL

The internal distances matrix is invariant under rigid and mirror transformations of the protein. While the first property leads to a simplification of the protein structure superposition problem as algorithms which compare proteins based on internal distances do not need to find the optimal rigid transformation, the second property may introduce errors as mirror images (such as a right handed helix and a left handed helix) will not be detected as being different. Consequently, methods have been concurrently developed to solve the protein structure alignment problem using coordinates to measure the similarity between two proteins. These methods are based on heuristic algorithms that optimize the correspondence between the two proteins and the rigid transformation simultaneously. Here I review the algorithm proposed by Michael Levitt, implemented in the program STRUCTAL (116).

STRUCTAL starts with an arbitrary equivalence between the two proteins A and B. This equivalence defines a list of corresponding residues (represented by their Cα) which are superimposed using 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:

$$SA(i,j) = \frac{20}{1 + 5d(C\alpha_i, C\alpha_j)^2} \qquad [36]$$

It is simple to compute, and has the important properties of being positive and of decreasing monotonically with increasing distances. A new alignment is then determined by searching in the distance matrix the alignment with the best score. Dynamics programming rapidly ($O(n^2)$ operations) find the optimum for the given structure alignment matrix, and a gap penalty. The gap penalty is set constant, equal to 10. This 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. This allows the computation of a new structure alignment matrix, and the procedure is iterated until the alignment matrix does not change anymore.

This structural alignment procedure based on dynamic programming is iterative, and as such may depend on the choice of the initial equivalence. STRUCTAL starts with five different equivalences. The first three equivalences are simple, and correspond to aligning the chain beginnings, the chain ends and the chain mid-points of the two structures, respectively, without allowing any gaps. The fourth choice maximizes sequence identity of the pairs of residues considered equivalent, while the fifth choice is based

on similarity of $C\alpha$ torsion angles between the two chains. After repeating the iterative scheme of finding the optimal equivalence and superposition for each of the five initial set of equivalences, the optimal alignment is chosen as the one with the highest score. Extensive studies have shown that no one of the five initial sets work better than another (117).

An approximate polynomial time algorithm

A prevailing sentiment in the community developing algorithms for protein structure alignment is that structure comparison requires exponential computer resources, and thus, investigations should concentrate on heuristic approaches. As a consequence, none of the existing methods guarantees finding an optimal alignment with respect to any scoring function. In addition, if one of these methods fails to find a good alignment, there is no guarantee that such an alignment does not exist. There is one interesting, though theoretical exception: Kolodny and Linial (137) have developed a polynomial-time algorithm that optimizes simultaneously the correspondence and the rigid transformation that leads to a structural alignment. The computation cost of their algorithm is of the order of $O(n^{10})$, and, as such, it is not practical. 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 equation [34].

For an algorithm for aligning two protein structures to be polynomial, the two following conditions must hold:

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

The STRUCTAL score *ST* is "separable" and an optimal correspondence can be found using dynamic programming in $O(n^2)$ in time and space requirements, for any given rigid transformation r. The score of this optimal correspondence is denoted $ST_{opt}(r)$. This validates condition 1. The validity of condition 2 is derived from a lemma given by Kolodny and Linial, which states that for all ε, there exists a finite set $G=G(\varepsilon)$ of rigid transformation, such that for every choice of a rigid transformation r, there exists a transformation r_G in $G(\varepsilon)$ 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, it is certain that it does not exist. As the size of $G(\varepsilon)$ is of the order of $O(n^{10}/\varepsilon^6)$, 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 practical. It does provide insights however on the complexity of protein structure alignments.

cRMS: an ambiguous measure of similarity

Though most of the algorithms for protein structure alignments use scoring schemes that differ significantly from simply taking into account interatomic distances (see above), the root mean square deviations (cRMS or dRMS) remain the measures of choice to describe the similarity between two proteins. 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, 2, in least-squares minimization: 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 these outliers. Another problem of RMS 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 (138). In general however, with varying correspondences it is possible to build a case where the triangular inequality is not satisfied. Consider for example two proteins A and B that are dissimilar, and the two-domain protein C, whose sub-domains C1 and C2 are strongly similar to A and B, respectively. In this example, the RMS values between A and C and between B and C are low, but the RMS between A and B is large, violating the triangular inequality that would have stated that RMS(A,B) ≤ RMS(A,C)+RMS(C,B). As a consequence of these limitations, RMS is a useful measure of structural similarity only for closely related proteins (139). Several other measures have therefore been proposed to circumvent these problems. The STRUCTAL score S2 (equation [35]) was defined as a more reliable indicator of structure similarity than RMS because it depends most strongly on the best-fitting pairs of atoms (thereby removing the weights of outliers), whereas RMS gives equal weight to all pairs of atoms. Interestingly, Lesk (140) recently proposed replacing the L2-norm in the RMS definition by the L norm, also called the Chebyshev norm, yielding a new score:

$$S = \max_{i \in [1,N]} \left\{ \left\| \boldsymbol{x}(i) - \boldsymbol{y}(i) \right\| \right\} \qquad [37]$$

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 the RMS. Yang and Honig (118) 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 (118) proposed that there is a continuous aspect of protein conformation space, in apparent disagreement with structural classification databases such as SCOP (Structural Classification Of Proteins (141)) and CATH (Class, Architecture, Topology and Homologous Superfamilies (79)). May (142) assessed 37 different protein structure similarity measures in terms of their robustness in generating accurate clusters in a hierarchical classification of 24 protein families. It was found in this study that the sum of ranks of distances at aligned positions was a better measure than the direct sum of distances and that RMS computed over the subset of core-aligned positions performs better than normal RMS. Variations in the hierarchical classification of protein structures raise the question of the validity not only of the measure used for the clustering, but also of the hierarchical clustering itself. The difficulty of defining a similarity score between protein structures is most probably a reflection of the fact that the problem of structure comparison does not have a unique answer (143-145). This could also reflect the fact that the problem is ill posed and that additional information is required to characterize a problem with a well-defined solution. For example, in fold recognition applications, predictors will 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 only includes atoms in the core.

A quantitative measure of the similarities of protein structures is essential for a critical assessment of the quality of protein structure predictions, such as those generated for CASP (a community-wide experiment on the Critical Assessment of techniques for protein Structure Prediction, organized in the form of a meeting held in alternating years at Asilomar, California). In the special case of comparing a predicted structure with the corresponding experimental structure, the equivalence list is known because the two sequences are identical, which reduces the complexity of the problem. On the other hand, each prediction may omit different residues and different parts of the structure may have different accuracies. Hubbard (146) solved the problem by generating a large number of superpositions and calculating the best RMS for each number of equivalent residues (not necessarily contiguous). The result is the RMS/coverage graph, which was used for the evaluation of predictions at CASP3. This plot can also be interpreted as defining

the number of equivalent residues for a given RMS value (the Adam Zemla's global distance test, GDT, used in CASP4).

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 lead to the development of a new approach for comparing protein structure, based on differential geometry and the concept of protein shape descriptors. The idea behind this approach is relatively simple: represent the protein structure with a vector of geometric properties, GP, such that the comparison of two protein structures is performed through a comparison of their GP vectors, usually using a Euclidian metric. Once the GP vectors have been computed, structure comparison using this scheme becomes instantaneous, and can then be performed over whole databases. Success of this approach obviously depends on the quality of the geometric properties included in GP, and their ability to uniquely capture the geometric properties of the protein. There has been a growing interest in the recent years to define such protein shape descriptors. Here I briefly review two descriptors derived from knot theory, namely the writhe and the radius of curvature of a polygonal curve.

The writhe of a protein chain

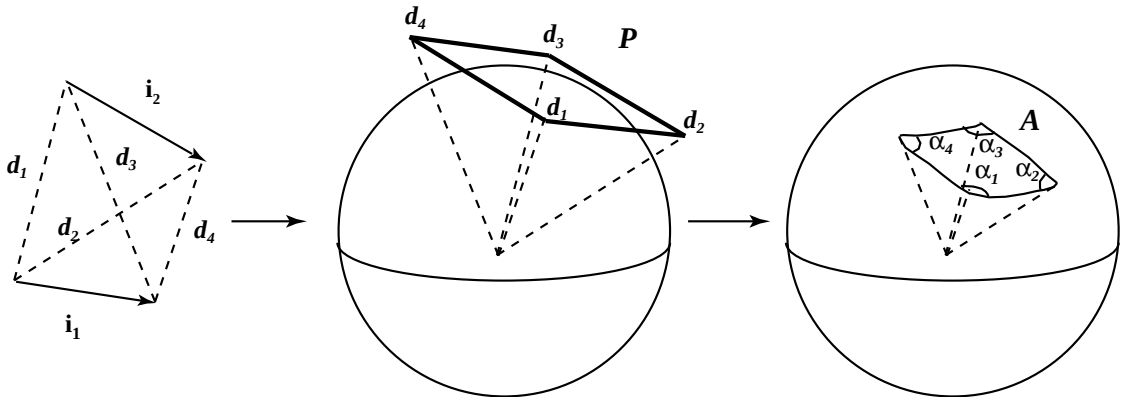


Figure 8: Computing ***W(i1,i2),* the writhe of two segments *i1* and *i2*.** The two segments i_1 and i_2 generates a parallelogram *P* of directions, with vertices d_1, d_2, d_3 and d_4. The area *A* of the projection of *P* on the surface of the unit sphere is the segment-segment writhe $W(i_1,i_2)$. *A* is easily computed as: $A = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - 2\pi$.

Geometrically, 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:

$$Wr(A) = I_{1,2}(A) = \sum_{0<i_1<i_2<N} W(i_1, i_2) \qquad [38]$$

with

$$W(i_1,i_2)=\frac{1}{2\pi}\int_{t_1=i_1}^{i_1+1}\int_{t_2=i_2}^{i_2+1}w(t_1,t_2)dt_1dt_2 \qquad [39]$$

where $W(i_1,i_2)$ is the contribution to writhe of the line segments i_1 and i_2. $W(i_1,i_2)$ is the probability to see the line segments cross 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:

$$I_{|1,2|}(A)=\sum_{0<i_1<i_2<N}\left|W(i_1,i_2)\right| \qquad [40]$$

A whole family of structural measures can be build using $W(i_1,i_2)$ and $|W(i_1,i_2)|$ as building blocks (147), such as:

$$I_{(1,3)|2,4|}=\sum_{0<i_1<i_2<i_3<i_4<N}W(i_1,i_3)\left|W(i_2,i_4)\right| \qquad [41]$$

and

$$I_{(1,4),(2,6),(3,5)}=\sum_{0<i_1<i_2<i_3<i_4<i_5<i_6<N}W(i_1,i_4)W(i_2,i_6)W(i_3,i_5) \qquad [42]$$

These measures are inspired from the Vassiliev knot invariants (148). 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 (149, 150). They have also been used to describe proteins. Levitt (151) has used writhe to distinguish different chain threading. Arteca and co-workers used the writhe as a protein shape descriptor (152-156). Rogen and Bohr (147) have used the writhe, the average crossing number and their higher order correlations to define a feature vector that characterize protein structures. More recently, Fain and Rogen (157) have compared protein structures using feature vectors in $\Re^{30}$ similar to those defined by Rogen and Bohr, and a pseudo metric, which is simply the Euclidian distance between the feature vectors. This pseudo metric is name SGM for scaled Gauss metric, as the writhe of a continuous curve is usually computed using a Gauss integral (158). Fain and Rogen (157) show that SGM performs extremely well as a protein structure classifier, using both CATH and SCOP as test sets. As both CATH and SCOP include all protein chains in the PDB, they are highly redundant and can not be considered as discriminative benchmarks. Despite this reserve, the results of Fain and Rogen are very promising, and open the door to a new way to compare and classify protein structures, using geometric protein shape descriptors.

Thickness and generalized radius of curvature of a protein chain

Any smooth, non-intersecting curve can be thickened into a smooth, non intersecting 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.

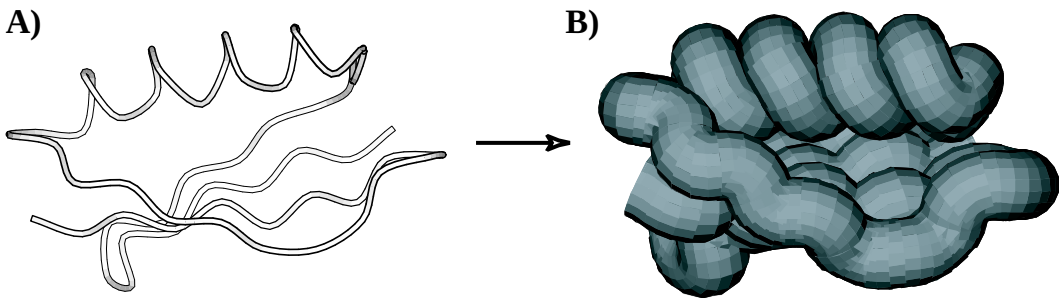


Figure 9: **Thickness of a protein**. (A) The structure of the B1 immunoglobulin-binding domain of streptococcal protein G (PDB code 1pgb) is visualized as a thin tube. (B) View of the same tube for 1pgb inflated to its "thickness", i.e. to a radius above which the tube ceases to be smooth, or shows self contact. Note that there is no free space between consecutive turns of the helices. Figure 9A drawn with MOLSCRIPT (14) and 9B with VMD (15).

This critical radius is referred to as the thickness of the curve, and it is used as a shape descriptor in knot theory. As the geometry of both DNA and protein molecules is characterized by the geometry of their chain, it is quite natural to check if thickness can be used as a descriptor of the shape of these molecules. This was pioneered by Gonzales and Maddocks (159), who introduced the concept of generalized radius of curvature, and applied this concept to characterize the geometry of DNA molecule. Their definition of generalized radius of curvature is based on the fact that any three non collinear points *x, y, z* in three dimensional space define a unique circle whose radius is given by:

$$r(x,y,z) = \frac{|x-y||x-z||y-z|}{4A(x,y,z)} \qquad [43]$$

where $A(x,y,z)$ is the area of the triangle whose vertices are *x,y* and *z* and $|x\text{-}y|$ is the Euclidian distance between x and y. Let us consider a discrete curve C, defined by n nodes ($c_1,c_2,\ldots c_n$). Gonzales and Maddocks define the generalized radius of curvature of C at c_i by:

$$\rho_C(c_i) = \min_{\substack{1\le j,k\le n \\ i\ne j\ne k\ne i}} r(c_i,c_j,c_k) \qquad [44]$$

$\rho_C(c_i)$ 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:

$$\Delta(C) = \min_{1\le i\le n} \rho_C(c_i) \qquad [45]$$

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 applied to the characterization of the geometry of DNA molecule (149, 150). They have also been used as basis for a "potential" that captures the geometry of a protein (160-164), which has been used for example in protein structure prediction computer experiments (165). They have not yet been used for protein structure comparison, but it is expected that they would prove quite useful for detecting protein structure similarities, when combined to other features such as writhe.

Upcoming challenges for protein structure comparison?

The most difficult application of protein structure comparison arises in the classification of the known protein structures into different clusters corresponding to fold families. The role of such classifications is to rationalize the organization of the protein structure databases such as the PDB, in hope to detect similarities at the structure level that could not be detected at the sequence level, and more generally to detect evolutionary relationships between proteins. The existing protein structure classifications are reviewed in the next section. The challenges met by a protein structure comparison program in this application are multiple. Firstly, it must be able to deal with different levels of structural similarities, must identify similarities even when these form a small proportion of the proteins being compared, and must be able to handle insertions of arbitrary size as well as permutations of substructures. Secondly, it must deal with the fact the 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 (166), and therefore cannot be ignored. Thirdly, the size of the protein structure databases has experienced exponential growth in the recent years, and the growth rate is expected to increase as the structural genomics projects enter their productive phases. This generates the need for fast techniques to compare and classify these structures, faster than the existing techniques that are usually time consuming.

None of the existing methods, including those described in length above, propose solutions to all these challenges. Heuristic methods were developed for sake of efficiency; there is no guarantee however that they find the optimal superposition. Some of these methods also cannot detect alternative, equally acceptable solutions. The approximate solution developed by Kolodny and Linial (137) solves some of these issues in the sense that it is able to 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 unusable for large scale comparisons. There is therefore still a need to develop faster, robust and exhaustive approaches to the problem of protein structure comparison. This field in fact remains an active area of development in structural biology. Solutions may in fact come from interdisciplinary research. The problem of comparing two protein structures can be reformalized as the problem of comparing two sets of points in 3D 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 should 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 building up.

Protein structure classification

In 1960 Perutz *et al.* (26) showed 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 proteins are functionally similar, as they are involved with the storage and transport of oxygen, respectively. Since then, there has been a continuing interest in finding structural similarities between proteins, in hope of revealing shared functionality that could not be detected by sequence information only. A logical consequence of this interest is the development of systems of classification of protein structures, whose aims are centered on the identification and regrouping of proteins sharing the same structure, in hope to reveal evolutionary relationships. Classifying protein structures has now become essential as the volume of structural data available grows exponentially (see figure 10). Note that in parallel to protein structure classification, there are many classifications of protein sequences available. I will not review those, as they are described in length in (167). For sake of completeness, I list in table 6 some of the resources available for sequence classification.

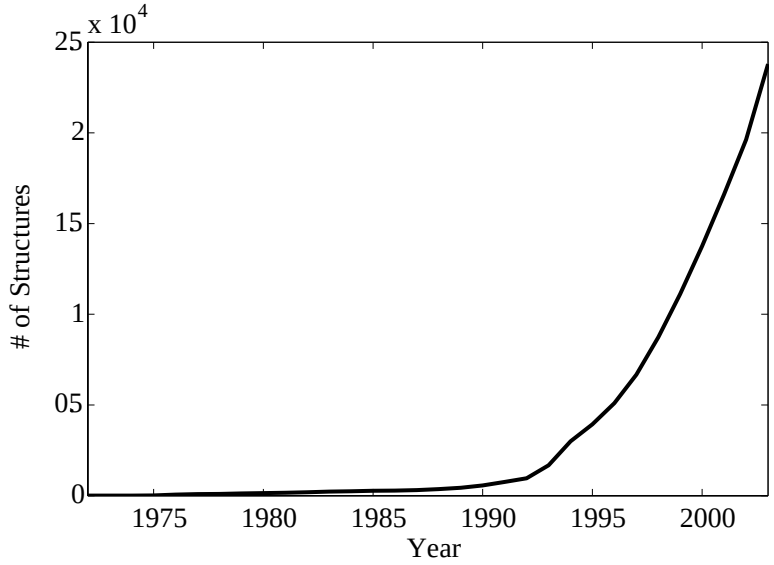


Figure 10: **Statistics on the PDB**. The number of structures (proteins and nucleic acids) available in the Protein Data Bank (PDB) (3, 4) is plotted versus time, starting from 1973 when the PDB was created.

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/

Table 6: Resources for classification of protein sequences

All current structural classification of proteins are based on the same scheme: protein structures are first divided into discrete, globular domains, which are then classified at the levels of “class”, “folds”, “superfamilies” and “families”; the differences arise from the methods used to define the domains, and on the procedures used to classified. After reviewing all the terms that define a classification, I will describe in details the three main protein structure classifications available, SCOP, CATH, and the DALI Domain Dictionary (DDD). Links to these databases and related services are listed in table 7.

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: semi-automatic 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 3D structures of homologous proteins	http://www-cryst.bioc.cam.ac.uk/data/align

Table 7: Resources for protein structure classifications

The first complication for structure classification is 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 prior to classification. The identification and delineation of these domains is still an open problem, which was discussed in length in the section on protein domain above. It is important to realise 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.

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

Classification schemes can be divided into curated, and automatic. A curated classification is based on human expertise, sometimes guided by computer analyses, to identify similarities between protein structures and organize them into groups. An automated classification relies on the results of the execution of a computer procedure to identify the similarities, which are subsequently processed automatically to generate the groups. One advantage of curation is the usual high quality of the clustering; the disadvantage is that it is difficult to scale to high volumes of data. Conversely, automatic procedures are fully reproducible and scalable, but may generate inaccurate assignment of similarity. The three main 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, using automated procedures complemented with human interventions.

The Structure Classification Of Proteins (SCOP)

SCOP (141) organizes protein structures hierarchically, to reflect both structural and evolutionary relatedness. SCOP has been constructed manually, from the delineation of the domains in multi-domain proteins to the organization of the levels of the hierarchy by visual inspection and comparison of protein structures, with the assistance of some automatic computer tools to make the task manageable and help provide consistency and generality. Since its creation in 1994, SCOP has been regularly updated, with an average frequency of two releases a year. The latest update of SCOP, 1.65 was built from the 20,619 PDB entries (54745 domains) available on August 1st, 2003, and was released in December 2003. Statistics on the growth of SCOP are given in figure 11.

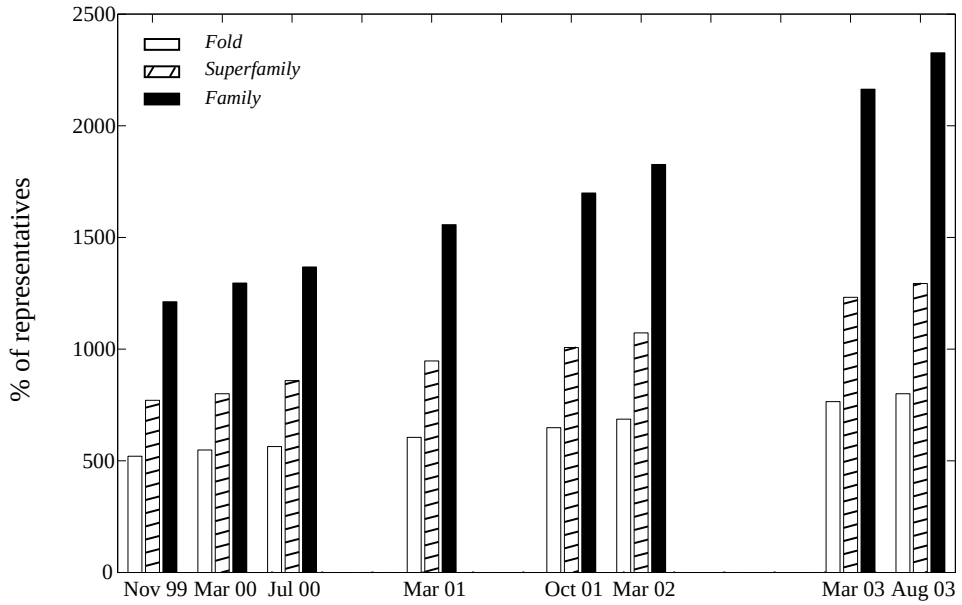


Figure 11: **Statistics of the SCOP classification of proteins**. The numbers of folds, superfamilies and families in SCOP are plotted versus "time", where time is the timestamp of the PDB used to generate the update of SCOP.

SCOP is a hierarchic classification, with four major levels, namely classes, folds, superfamilies and families, described below. As recognized by the authors of SCOP, the exact positions of boundaries between these levels are to some degree 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 hierarchy are 11 different classes: alpha, beta, alpha and beta (α/β), alpha plus beta ($\alpha+\beta$), multi-domain 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 could not (yet) be classified among the major classes, and are maintained in SCOP for sake of completeness and compatibility with the PDB.

SCOP **folds** identify structural similarities. 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 placed 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 who share the same fold and have similar functions, suggesting that a common evolutionary origin is probable, are placed together in superfamilies.

Proteins clustered together into **families** are clearly evolutionary related. In general, the sequences of two proteins placed in the same family have a residue identity greater than 30%. In some cases, a high sequence identity is not need to affirm common origin: many globins form a family, even though some of the members of that family have sequence identity of only 15%.

The CATH classification.

CATH (79) clusters protein domains at four major levels: Class (C), Architecture (A), Topology (T), and Homologous superfamily (H), described below. CATH is derived from a semi-automatic procedure. CATH filters out non-protein, models and "C-alpha 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 28th, 2004 and includes 48,391 domains.
Multi-domain proteins are subdivided into domains using a consensus procedure, based on three algorithms for domain recognition, DETECTIVE (78), PUU (73) and DOMAK (73). When all three algorithms agree on a protein, the common solution is used to delineate the domains of that protein. This procedure allowed 53% of the proteins included into the CATH release 2.5.1 to be subdivided into domains automatically. The remaining structures are assigned domains manually, using one of the assignments made by the automatic procedure, an assignment obtained from the literature, or a new assignment defined by visual inspection.

CATH includes 4 classes (C): alpha, beta, alpha and beta, and few secondary structures. The alpha-beta class includes both alternating alpha/beta structures and alpha+beta structures, originally defined by Levitt and Chothia (27). The class of a protein domain is determined according to its secondary structure composition and packing. In release 2.5.1 of CATH, it is assigned automatically for over 90% of the domains, using the method developed by Michie et al (168, 169). The 10% remaining domains are assigned to a class using 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 using the structural alignment program SSAP (98). Proteins belonging to the same class are compared systematically using 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 appendix).

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

The DALI Domain Dictionary

The DALI Domain Dictionary, also called DALI domain classification is derived from a fully automated method of defining and classifying domains ((170), Dietmann and Holm (171). DALI domains are defined by a version of the PUU algorithm (73) that has been updated to consider the recurrence of putative domains (172). Structural similarities between domains are defined by the program DALI (see the section on DALI above for an overview of that program). 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 the 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 initially used to create the FSSP database (173). FSSP is known as "Families of Structurally Similar Proteins". In FSSP, pair-wise structural comparisons are made between proteins of a representative set, where no two proteins have greater than 25% sequence identity. For each member of the representative set, a file a created that contains all pair-wise structural matches with a Z-score greater than 2.0.

The same procedure was extended to generate a complete classification of all protein domains in the PDB90 database, the DALI Domain Dictionary or DALI Domain Classification (171). 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 appendix) was used to generate a fold tree covering the PDB90 database. The pair-wise structural alignments are divided using 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 DALI Domain Dictionary are continuously updated; this is set up easily, as they are both derived from a fully automated procedure.

Comparing SCOP, CATH and DDD

Despite differences in the classification methods they have implemented, and in the rules of protein structure and taxonomy they are based on, SCOP, CATH and DDD agree on the majority of their classifications. Hadley and Jones (174) 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 explanations. Among these, the domain assignment plays a crucial role. As mentioned above, 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 in differences in classification is the exclusion of one part of a protein. Hadley and Jones (174) reports the case of papain (PDB code 1ppo), 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 splitted the protein into two domains, separating the cysteine from the asparagine and histidine, and rendering each domain effectively functionless. Note that this difference has been corrected since Hadley and Jones published their study, and papain is now a single domain in CATH. Another discrepancy between the structural classifications arises from the 'fold overlap' problem, where a fold within one classification encompasses more than one fold within another classification. When a domain is classified within CATH as a three layer ($\alpha\beta\alpha$) sandwich Rossmann fold, there are several SCOP folds to which it could conceivably belong: although the structures are geometrically similar, SCOP can separate them to reflect an evolutionary distinction. This 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 consequently regrouped in the same Topology in CATH (3.40.50), while they are representatives of their own fold class in SCOP (c.44, c.65 and c.62, respectively).

Despite these discrepancies, Hadley and Jones (174) recognize the merits of all three classifications, and conclude that no one method is distinctly superior. 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, continually updated.

Divergences in protein structure classifications have triggered the search of a consensus description of the protein structure space. Day et al (175) recently repeated the comparative study of SCOP, CATH and DDD, based on updated versions of the classifications compared to Hadley and Jones' work. While Day et al find significant levels of agreement between the three classifications, they highlight disparities whose origins are similar than those found earlier. To average out these disparities, they have introduced the concept of consensus folds. They start from a non redundant subset of protein domains. To be considered, 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 to be those having >95% sequence identity to a previously counted domain. Each domain in the non redundant subset is assigned a fold identifier, corresponding to its classifications in SCOP, CATH and DDD. Domains are then clustered on the basis of their fold identifiers, and the corresponding clusters are referred to as metafolds. The non redundant 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 the key molecules to all cellular functions. Nature has extensively explored their sequences and structures in order 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 diversity of protein sequences and structures therefore provides invaluable clues to unravel the mysteries of life and its evolution and adaptation over time. This opportunity has become a leitmotiv in recent genetics and molecular biology studies, crystallized by the development of numerous genomics and structural genomics projects. At time of writing, more than 220 whole genomes have been sequenced and published on the world wide web, and more than 1200 are currently under studies, corresponding to databases of more than one million non redundant protein sequences. In parallel, the protein structure database contains structural data on more than 27000 proteins. The challenge now is to organize all these data in a way that evolutionary relationship 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. While such approaches detect homologous proteins, they usually fail to detect remote homologues, i.e. pairs of proteins that have similar structure and function, but lack easily detectable sequence similarity. As protein structures are more highly conserved than 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 identify connections between the elementary groups. These steps are usually performed using classification techniques. In the case of protein structures, this has lead to the construction and maintenance of protein structure classifications, which I have reviewed in this survey.

Reliable protein structure superposition remains a bottleneck when building 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. Clearly, better alignments match more positions, have fewer gaps and are more similar. Since 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 these values. Several measures of similarity have consequently been developed (142). Perhaps the most significant recent improvements in this area have been in the protocol assessing the statistical significance of these measures (117, 176)). These measures of similarity are used in structure comparison algorithms. Ideally, these methods 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, but not these two criteria (see section on protein structure superposition). There is therefore a need 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. This is supported by recent analyses of the available genome data, which suggest that at least 60% of the genes are multi-domain proteins (177-179). Domain duplications and recombination are thought to have occurred extensively. Protein structure classifications are consequently domain based, and only contain multi-domain proteins when the domains of the later have not yet been assigned. Automatic recognition of domains in multi-domain proteins can be very difficult, although many promising approaches have been developed (see section on protein domain above). These methods do not always agree in their domain assignments; this leads to discrepancies between the existing protein structure classifications (174, 175).

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 DALI Domain Dictionary, 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 integrated with the other genome databases under constructions. Currently, SCOP, CATH and DDD are valuable resources mostly for benchmarking of methods, and for structural studies. Their impact in biology in general will be far greater when they will be integrated with sequence and function information, in order to present a cohesive picture of the different protein spaces.

Acknowledgments

PK acknowledges support from the National Science Foundation (Grant CCR-00-86013) and from the National Institute of Health (GM 63817).

References

1 J. Monod, *Le Hasard Et La Necessite*, Seuil, Paris, France, 1973.

2 A. Bernal, U. Ear and N. Kyrpides, *Nucl. Acids. Res.*, **29**, 126 -127 (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 -542 (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 -242 (2000). The Protein Data Bank.

5 W. Gilbert, *Nature (London)*, **349**, 99 (1991). Towards a Paradigm Shift in Biology.

6 G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure*, Springer-Verlag, New York, 1979.

7 C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry: The Conformation of Biological Macromolecules*, W. H. Freeman Company, New York, 1980.

8 C. Branden and J. Tooze, *Introduction to Protein Structure*, Garland Publishing, New York, 1991.

9 T. E. Creighton, *Proteins*, W. H. Freeman & Co., New York, 1993.

10 W. R. Taylor, A. C. W. May, N. P. Brown and A. Aszodi, *Reports Prog. Phys.*, **64**, 517-90 (2001). Protein Structure: Geometry, Topology and Classification.

11 R. B. Corey and L. Pauling, *Rev. Sci. Instr.*, **24**, 621 -627 (1953). Molecular Models of Amino Acids, Peptides and Proteins.

12 W. L. Koltun, *Biopolymers*, **3**, 665 -679 (1965). Precision Space-Filling Atomic Models.

13 J. Kendrew, R. Dickerson, B. Strandberg, R. Hart, D. Davies and D. Philips, *Nature (London)*, **185**, 422 -427 (1960). Structure of Myoglobin: A Three Dimensional Fourier Synthesis at 2 Angstrom Resolution.

14 P. J. Kraulis, *J. Appl. Crystallo.*, **24**, 946 -950 (1991). MOLSCRIPT: A Program to Produce Both Detailed and Schematic Plots of Protein Structures.

15 W. Humphrey, A. Dalke and K. Schulten, *J. Molec. Graphics.*, **14**, 33-38 (1996). VMD - Visual Molecular Dynamics.

16 K. C. Timberlake, (1992). Chemistry, 5th Edition.

17 G. M. Crippen, *J. Mol. Biol.*, **126**, 315-332 (1978). Tree Structural Organization of Proteins.

18 A. V. Efimov, *FEBS Lett.*, **224**, 372-376 (1987). Pseudo-Homology of Protein Standard Structures Formed by 2 Consecutive Beta-Strands.

19 A. V. Efimov, *FEBS Lett.*, **284**, 288-292 (1991). Structure of Coiled Beta-Beta Hairpins and Beta-Beta-Corners.

20 A. V. Efimov, *Protein Eng.*, **4**, 245-250 (1991). Structure of Alpha-Alpha Hairpins with Short Connections.

21 A. V. Efimov, *Prog. Biophys. Mol. Biol.*, **60**, 201-39 (1993). Standard Structures in Proteins.

22 C. Brooks, M. Karplus and M. Pettitt, *Adv. Chem. Phys.*, **71**, 1 -259 (1988). Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics.

23 C. B. Anfinsen, *Science*, **181**, 223 -230 (1973). Principles That Govern Protein Folding.

24 P. Koehl and M. Levitt, *Nature Struct. Biol.*, **6**, 108 -111 (1999). A Brighter Future for Protein Structure Prediction.

25 D. Baker and A. Sali, *Science*, **294**, 93 -96 (2001). Protein Structure Prediction and Structural Genomics.

26 M. Perutz, M. Rossmann, A. Cullis, G. Muirhead, G. Will and A. North, *Nature (London)*, **185**, 416 -422 (1960). Structure of Haemoglobin: A Three-Dimensional Fourier Synthesis at 5.5 Angstrom Resolution, Obtained by X-Ray Analysis.

27 M. Levitt and C. Chothia, *Nature (London)*, **261**, 552-558 (1976). Structural Patterns in Globular Proteins.

28 A. M. Lesk and C. Chothia, *J. Mol. Biol.*, **136**, 225-270 (1980). How Different Amino-Acid Sequences Determine Similar Protein Structures the Structure and Evolutionary Dynamics of the Globins.

29 C. Chothia and J. Janin, *Proc. Natl. Acad. Sci. (USA)*, **78**, 4146-4150 (1981). Relative Orientation of Close Packed Beta Pleated Sheets in Proteins.

30 C. Chothia and J. Janin, *Biochemistry*, **21**, 3955-3965 (1982). Orthogonal Packing of Beta Pleated Sheets in Proteins.

31 F. E. Cohen, M. J. E. Sternberg and W. R. Taylor, *J. Mol. Biol.*, **148**, 253-272 (1981). Analysis of the Tertiary Structure of Protein Beta Sheet Sandwiches.

32 F. E. Cohen, M. J. E. Sternberg and W. R. Taylor, *J. Mol. Biol.*, **156**, 821-862 (1982). Analysis and Prediction of the Packing of Alpha Helices against a Beta Sheet in the Tertiary Structure of Globular Proteins.

33 K. C. Chou, *Proteins: Struct. Func. Genet.*, **21**, 319-344 (1995). A Novel Approach to Predicting Protein

Structural Classes in a (20-1)-D Amino Acid Composition Space.
34 Y. D. Cai, Y. X. Li and K. C. Chou, *Biochim. Biophys. Acta.*, **1476**, 1-2 (2000). Using Neural Networks for Prediction of Domain Structural Classes.
35 W. M. Liu and K. C. Chou, *J. Prot. Chem.*, **17**, 209-217 (1998). Prediction of Protein Structural Classes by Modified Mahalanobis Discriminant Algorithm.
36 I. Bahar, A. R. Atilgan, R. L. Jernigan and B. Erman, *Proteins: Struct. Func. Genet.*, **29**, 172-185 (1997). Understanding the Recognition of Protein Structural Classes by Amino Acid Composition.
37 R. Y. Luo, Z. P. Feng and J. K. Liu, *Eur. J. Biochem.*, **269**, 4219-4225 (2002). Prediction of Protein Structural Class by Amino Acid and Polypeptide Composition.
38 K. C. Chou and C. T. Zhang, *Critical Rev. Biochem. Molec. Biol.*, **30**, 275-349 (1995). Prediction of Protein Structural Classes.
39 G. P. Zhou and N. Assa-Munt, *Proteins: Struct. Func. Genet.*, **44**, 57-59 (2001). Some Insights into Protein Structural Class Prediction.
40 K. C. Chou, W. M. Liu, G. M. Maggiora and C. T. Zhang, *Proteins: Struct. Func. Genet.*, **31**, 97-103 (1998). Prediction and Classification of Domain Structural Classes.
41 E. G. Hutchinson and J. M. Thornton, *Protein Eng.*, **6**, 233-245 (1993). The Greek Key Motif: Extraction, Classification and Analysis.
42 J. S. Richardson, *Nature (London)*, **268**, 495-500 (1977). β–Sheet Topology and the Relatedness of Proteins.
43 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**, 609-614 (1975). Structure of Chicken Muscle Triose Phosphate Isomerase Determined Crystallographically at 2.5a Resolution Using Amino-Acid Sequence Data.
44 A. G. Murzin, A. M. Lesk and C. Chothia, *J. Mol. Biol.*, **236**, 1369-1381 (1994). Principles Determining the Structure of Beta-Sheet Barrels in Proteins .1. A Theoretical-Analysis.
45 A. G. Murzin, A. M. Lesk and C. Chothia, *J. Mol. Biol.*, **236**, 1382-1400 (1994). Principles Determining the Structure of Beta-Sheet Barrels in Proteins .2. The Observed Structures.
46 R. K. Wierenga, *FEBS Lett.*, **492**, 193-198 (2001). The Tim-Barrel Fold: A Versatile Framework for Efficient Enzymes.
47 G. Pujadas and J. Palau, *Biologia (Bratislava)*, **54**, 231-254 (1999). Tim Barrel Fold: Structural, Functional and Evolutionary Characteristics in Natural and Designed Molecules.
48 N. Nagano, C. A. Orengo and J. M. Thornton, *J. Mol. Biol.*, **321**, 741-765 (2002). One Fold with Many Functions: The Evolutionary Relationships between Tim Barrel Families Based on Their Sequences, Structures and Functions.
49 E. L. Wise and I. Rayment, *Acc. Chem. Res.*, **37**, 149-158 (2004). Understanding the Importance of Protein Structure to Nature's Routes for Divergent Evolution in Tim Barrel Enzymes.
50 M. C. Vega, E. Lorentzen, A. Linden and M. Wilmanns, *Curr. Opin. Chem. Biol.*, **7**, 694-701 (2003). Evolutionary Markers in the (Beta/Alpha)8-Barrel Fold.
51 J. A. Gerlt and F. M. Raushel, *Curr. Opin. Chem. Biol.*, **7**, 252-264 (2003). Evolution of Function in (Beta/Alpha)8-Barrel Enzymes.
52 G. D. Rose, *J. Mol. Biol.*, **134**, 447-470 (1979). Hierarchic Organization of Domains in Globular Proteins.
53 J. S. Richardson, *Adv. Protein. Chem.*, **34**, 167-339 (1981). The Anatomy and Taxonomy of Protein Structure.
54 J. Janin and C. Chothia, *Methods in Enzymology*, **115**, 420-430 (1985). Domains in Proteins - Definitions, Location, and Structural Principles.
55 C. P. Ponting and R. R. Russell, in *Annual Review of Biophysics and Biomolecular Structure*, R. M. Stroud, W. K. Olson and M. P. Sheetz, Eds, Annual Reviews, Palo Alto, 2002, pp 45-71. The Natural History of Protein Domains.
56 S. Veretnik, P. E. Bourne, N. N. Alexandrov and I. N. Shindyalov, *J. Mol. Biol.*, **339**, 647-678 (2004). Toward Consistent Assignment of Structural Domains in Proteins.
57 R. A. Laskowski, E. G. Hutchinson, A. D. Michie, A. C. Wallace, M. L. Jones and J. M. Thornton, *Trends Biochem. Sci.*, **22**, 488-490 (1997). PDBSUM: A Web-Based Database of Summaries and Analyses of All PDB Structures.
58 R. A. Laskowski, *Nucl. Acids. Res.*, **29**, 221-222 (2001). PDBSUM: Summaries and Analyses of PDB Structures.
59 W. Kabsch and C. Sander, *Biopolymers*, **22**, 2577-2637 (1983). Dictionary of Protein Secondary Structure:

Pattern Recognition of Hydrogen Bonded and Geometrical Features.
60 G. Wang and R. L. Dunbrack, *Bioinformatics (Oxford)*, **19**, 1589-1591 (2003). Pisces: A Protein Sequence Culling Server.
61 S. E. Brenner, P. Koehl and R. Levitt, *Nucl. Acids. Res.*, **28**, 254-256 (2000). The Astral Compendium for Protein Structure and Sequence Analysis.
62 J. M. Chandonia, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt and S. E. Brenner, *Nucl. Acids. Res.*, **30**, 260-263 (2002). Astral Compendium Enhancements.
63 J. M. Chandonia, G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt and S. E. Brenner, *Nucl. Acids. Res.*, **32**, D189-D192 (2004). The Astral Compendium in 2004.
64 M. G. Rossman and A. Liljas, *J. Mol. Biol.*, **85**, 177-181 (1974). Recognition of Structural Domains in Globular Proteins.
65 M. H. Zehfus and G. D. Rose, *Biochemistry*, **25**, 5759-5765 (1986). Compact Units in Proteins.
66 S. A. Islam, J. C. Luo and M. J. E. Sternberg, *Protein Eng.*, **8**, 513-525 (1995). Identification and Analysis of Domains in Proteins.
67 N. Alexandrov and I. Shindyalov, *Bioinformatics (Oxford)*, **19**, 429-430 (2003). PDP: Protein Domain Parser.
68 A. S. Siddiqui and G. J. Barton, *Protein Sci.*, **4**, 872-884 (1995). Continuous and Discontinuous Domains: An Algorithm for the Automatic Generation of Reliable Protein Domain Definitions.
69 S. J. Wodak and J. Janin, *Biochemistry*, **20**, 6544-6552 (1981). Location of Structural Domains in Proteins.
70 A. A. Rashin, *Nature (London)*, **291**, 85-86 (1981). Location of Domains in Globular Proteins.
71 N. Go, *Proc. Nat. Acad. Sci. (USA)*, **80**, 1964-1968 (1983). Modular Structural Units, Exon and Function in Chicken Lysozyme.
72 M. H. Zehfus, *Protein Eng.*, **7**, 335-340 (1994). Binary Discontinuous Compact Protein Domains.
73 L. Holm and C. Sander, *Proteins: Struct. Func. Genet.*, **19**, 256-268 (1994). Parser for Protein Folding Units.
74 Y. Xu, D. Xu and H. N. Gambow, *Bioinformatics (Oxford)*, **16**, 1091-1104 (2000). Protein Domain Decomposition Using a Graph-Theoretic Approach.
75 J. T. Guo, D. Xu, D. Kim and Y. Xu, *Nucl. Acids. Res.*, **31**, 944-952 (2003). Improving the Performance of Domainparser for Structural Domain Partition Using Neural Network.
76 W. R. Taylor, *Protein Eng.*, **12**, 203-216 (1999). Protein Structural Domain Identification.
77 M. B. Swindells, *Protein Sci.*, **4**, 93-102 (1995). A Procedure for the Automatic Determination of Hydrophobic Cores in Protein Structures.
78 M. B. Swindells, *Protein Sci.*, **4**, 103-112 (1995). A Procedure for Detecting Structural Domains in Proteins.
79 C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells and J. M. Thornton, *Structure*, **5**, 1093-1108 (1997). Cath : A Hierarchic Classification of Protein Domain Structures.
80 A. S. Siddiqui and G. J. Barton, *Protein Sci.*, **4**, 872-884 (1995). Continuous and Discontinuous Domains : An Algorithm for the Automatic-Generation of Reliable Protein Domain Definitions.
81 C. Guerra and S. Istrail, *Mathematical Methods for Protein Structure Analysis and Design, Advanced Lectures*, Lecture Notes in Computer Science, Springer, 2003.
82 C. Chen and Q. Li, *Acta Cryst. A*, **A60**, 201-203 (2004). A Strict Solution for the Optimal Superposition of Protein Structures.
83 B. Sabata and J. K. Aggarwal, *Comput. Vis. Graph. Image Proc: Image Understanding*, **54**, 309-324 (1991). Estimation of Motion from a Pair of Range Images: A Review.
84 C. Ferrari and C. Guerra, in *Mathematical Methods for Protein Structure Analysis and Design*, C. Guerra and S. Istrail, Eds, Springer, 2003, pp 57-82. Geometric Methods for Protein Structure Comparison.
85 D. W. Eggert, A. Lorusso and R. B. Fisher, *Mach. Vis. and Applic.*, **9**, 272-290 (1997). Estimating 3d Rigid Body Transformations: A Comparison of Four Major Algorithms.
86 G. H. Golub and C. F. V. Loan, *Matrix Computation*, John Hopkins University Press, 1996.
87 W. Kabsch, *Acta Cryst. A*, **32**, 922-923 (1976). Solution for Best Rotation to Relate 2 Sets of Vectors.
88 W. Kabsch, *Acta Cryst. A*, **34**, 827-828 (1978). Discussion of Solution for Best Rotation to Relate 2 Sets of Vectors.
89 A. D. Mclachlan, *J. Mol. Biol.*, **128**, 49-80 (1979). Gene Duplications in the Structural Evolution of Chymotrypsin.
90 K. S. Arun, T. S. Huang and S. D. Blostein, *IEEE Trans. Pattern Anal. & Machine Intel.*, **9**, 698-700 (1987). Least-Sqaure Fitting of Two 3d Point Sets.

91 P. H. Schonemann, *Psychometrica*, **31**, 1-10 (1966). A Generalized Solution of the Orthogonal Procrustes Problem.
92 B. Horn, H. Hilden and S. Negahdaripour, *J. Opt. Soc. Am.*, **5**, 1127-1135 (1988). Closed-Form Solution of Absolute Orientation Using Orthonormal Matrices.
93 B. Horn, *J. Opt. Soc. Am.*, **4**, 629-642 (1987). Closed-Form Solution of Absolute Orientation Using Unit Quaternions.
94 M. W. Walker, L. Shao and R. A. Voltz, *CVGIP: image understanding*, **54**, 358-367 (1991). Estimating 3d Location Parameters Using Dual Number Quaternions.
95 E. A. Coutsias, C. Seok and K. A. Dill, *J. Comp. Chem*, **25**, 1849-1857 (2004). Using Quaternions to Calculate Rmsd.
96 S. Umeyama, *IEEE Trans. Pattern Anal. & Machine Intel.*, **13**, 376-380 (1991). Least-Squares Estimation of Transformation Parameters between 2-Point Patterns.
97 P. Koehl, *Curr. Opin. Struct. Biol.*, **11**, 348-353 (2001). Protein Structure Similarities.
98 W. R. Taylor and C. A. Orengo, *J. Mol. Biol.*, **208**, 1-22 (1989). Protein Structure Alignment.
99 W. R. Taylor, *Protein Sci.*, **8**, 654-665 (1999). Protein Structure Comparison Using Iterated Double Dynamic Programming.
100 K. Nishikawa and T. Ooi, *J. Theo. Biol.*, **43**, 351-374 (1974). Comparison of Homologous Tertiary Structures of Proteins.
101 L. Holm, C. Ouzounis, C. Sander, G. Tuparev and G. Vriend, *Protein Sci.*, **1**, 1691-1698 (1992). A Database of Protein Structure Families with Common Folding Motifs.
102 L. Holm and C. Sander, *J. Mol. Biol.*, **233**, 123-128 (1993). Protein Structure Comparison by Alignment of Distance Matrices.
103 L. Holm and C. Sander, *Nature (London)*, **361**, 309 (1993). Globin Fold in a Bacterial Toxin.
104 G. Vriend and C. Sander, *Proteins: Struct. Func. Genet.*, **11**, 52-58 (1991). Detection of Common 3-Dimensional Substructures in Proteins.
105 N. N. Alexandrov, K. Takahashi and N. Go, *J. Mol. Biol.*, **225**, 5-9 (1992). Common Spatial Arrangements of Backbone Fragments in Homologous and Nonhomologous Proteins.
106 D. Fischer, O. Bachar, R. Nussinov and H. Wolfson, *J. Biomol. Struct. Dyn.*, **9**, 769-789 (1992). An Efficient Automated Computer Vision Based Technique for Detection of 3-Dimensional Structural Motifs in Proteins.
107 D. Fischer, H. Wolfson and R. Nussinov, *J. Biomol. Struct. Dyn.*, **11**, 367-380 (1993). Spatial, Sequence-Order-Independent Structural Comparison of Alpha/Beta Proteins: Evolutionary Implications.
108 D. Fischer, H. Wolfson, S. L. Lin and R. Nussinov, *Protein Sci.*, **3**, 769-778 (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.
109 R. Nussinov and H. J. Wolfson, *Proc. Natl. Acad. Sci. (USA)*, **88**, 10495-10499 (1991). Efficient Detection of 3-Dimensional Structural Motifs in Biological Macromolecules by Computer Vision Techniques.
110 R. Nussinov, D. Fischer and H. Wolfson, *Faseb J.*, **6**, A349-A349 (1992). A Computer Vision Based 3-Dimensional Approach for the Comparison of Protein Structures.
111 H. J. Wolfson and I. Rigoutsos, *IEEE Comput. Sci. & Eng.*, **4**, 10-21 (1997). Geometric Hashing: An Overview.
112 P. J. Artymiuk, A. R. Poirrette, H. M. Grindley, D. W. Rice and P. Willett, *J. Mol. Biol.*, **243**, 327-344 (1994). A Graph-Theoretic Approach to the Identification of 3-Dimensional Patterns of Amino-Acid Side-Chains in Protein Structures.
113 P. J. Artymiuk, A. R. Poirrette, D. W. Rice and P. Willett, *Topics Curr. Chem.*, **174**, 73-103 (1995). The Use of Graph-Theoretical Methods for the Comparison of the Structures of Biological Macromolecules.
114 E. J. Gardiner, P. J. Artymiuk and P. Willett, *J. Mol. Graph. & Modelling*, **15**, 245-253 (1997). Clique-Detection Algorithms for Matching Three-Dimensional Molecular Structures.
115 T. D. Wu, S. C. Schmidler, T. Hastie and D. L. Brutlag, *J. Comput. Biol.*, **5**, 585-595 (1998). Regression Analysis of Multiple Protein Structures.
116 S. Subbiah, D. V. Laurents and M. Levitt, *Curr. Biol.*, **3**, 141-148 (1993). Structural Similarity of DNA-Binding Domains of Bacteriophage Repressors and the Globin Core.
117 M. Gerstein and M. Levitt, *Protein Sci.*, **7**, 445-456 (1998). Comprehensive Assessment of Automatic Structural Alignment against a Manual Standard ; the Scop Classification of Proteins.
118 A. S. Yang and B. Honig, *J. Mol. Biol.*, **301**, 665-678 (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.
119 L. Holm and C. Sander, *J. Mol. Biol.*, **233**, 123-138 (1993). Protein-Structure Comparison by Alignment of Distance Matrices.
120 J. D. Szustakowski and Z. P. Weng, *Proteins: Struct. Func. Genet.*, **38**, 428-440 (2000). Protein Structure Alignment Using a Genetic Algorithm.
121 I. N. Shindyalov and P. E. Bourne, *Protein Eng.*, **11**, 739-747 (1998). Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path.
122 M. E. Ochagavia and H. Wodak, *Proteins: Struct. Func. Bioinfo.*, **55**, 436-454 (2004). Progressive Combinatorial Algorithm for Multiple Structural Alignments: Application to Distantly Related Proteins.
123 R. Blankenbecler, M. Ohlson, C. Peterson and M. Ringler, *Proc. Nat. Acad. Sci. (USA)*, **100**, 11936-11940 (2003). Matching Protein Structures with Fuzzy Alignments.
124 L. Chen, T. Zhou and Y. Tang, *Bioinformatics (Oxford)*, (in press) (2004). Protein Structure Alignment by Deterministic Annealing.
125 B. W. Matthews and M. G. Rossmann, *Methods in Enzymology*, **115**, 397-420 (1985). Comparison of Protein Structures.
126 M. L. Sierk and W. R. Pearson, *Protein Sci.*, **13**, 773-785 (2004). Sensitivity and Selectivity in Protein Structure Comparison.
127 M. Novotny, D. Madsen and G. J. Kleywegt, *Proteins: Struct. Func. Genet.*, **54**, 260-270 (2004). Evaluation of Protein Fold Comparison Servers.
128 R. Kolodny, P. Koehl and M. Levitt, *J. Mol. Biol.*, (in press) (2004). Comprehensive Evaluation of Structural Alignment Method: Scoring by Geometric Match Measures.
129 L. Holm and C. Sander, *Trends Biochem. Sci.*, **20**, 478-480 (1995). Dali - a Network Tool for Protein-Structure Comparison.
130 R. B. Russell and G. J. Barton, *Proteins: Struct. Func. Genet.*, **14**, 309-323 (1992). Multiple Protein Sequence Alignment from Tertiary Structure Comparison Assignment of Global and Residue Confidence Levels.
131 M. Suyama, Y. Matsuo and K. Nishikawa, *J. Mol. Evol.*, **44**, S163-S173 (1997). Comparison of Protein Structures Using 3d Profile Alignment.
132 J. U. Bowie, R. Lüthy and D. Eisenberg, *Science*, **253**, 164-170 (1991). A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure.
133 J. Jung and B. Lee, *Protein Eng.*, **13**, 535-543 (2000). Protein Structure Alignment Using Environmental Profiles.
134 T. Kawabata and K. Nishikawa, *Proteins: Struct. Func. Genet.*, **41**, 108-122 (2000). Protein Structure Comparison Using the Markov Transition Model of Evolution.
135 I. D. Kuntz, G. M. Crippen, P. A. Kollman and D. Kimelman, *J. Mol. Biol.*, **106**, 983-994 (1976). Calculation of Protein Tertiary Structure.
136 G. M. Crippen and T. F. Havel, *Distance Geometry and Molecular Conformation*, Research Studies, Somerset, UK, 1988.
137 R. Kolodny and N. Linial, *Proc. Natl. Acad. Sci. (USA)*, **101**, 12201-12206 (2004). Approximate Protein Structural Alignment in Polynomial Time.
138 K. Kaindl and B. Steipe, *Acta Cryst. A*, **53**, 809 (1997). Metric Properties of the Root-Mean-Square Deviation of Vector Sets.
139 K. Mizuguchi and N. Go, *Curr. Opin. Struct. Biol.*, **5**, 377-382 (1995). Seeking Significance in 3-Dimensional Protein-Structure Comparisons.
140 A. M. Lesk, *Proteins: Struct. Func. Genet.*, **33**, 320-328 (1998). Extraction of Geometrically Similar Substructures: Least-Squares and Chebyshev Fitting and the Difference Distance Matrix.
141 A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, *J. Mol. Biol.*, **247**, 536-540 (1995). Scop : A Structural Classification of Proteins Database for the Investigation of Sequences and Structures.
142 A. C. W. May, *Proteins: Struct. Func. Genet.*, **37**, 20-29 (1999). Toward More Meaningful Hierarchical Classification of Protein Three-Dimensional Structures.
143 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**, 1977-1983 (1995). Structure Similarity between the Pleckstrin Homology Domain and Verotoxin: The Problem of Measuring and Evaluating Structural Similarity.
144 Z. K. Feng and M. J. Sippl, *Folding & Design*, **1**, 123-132 (1996). Optimum Superimposition of Protein Structures: Ambiguities and Implications.
145 A. Godzik, *Protein Sci.*, **5**, 1325-1338 (1996). The Structural Alignment between Two Proteins: Is There a

Unique Answer?
146 T. J. P. Hubbard, *Proteins: Struct. Func. Genet.*, 15-21 (1999). Rms/Coverage Graphs: A Qualitative Method for Comparing Three-Dimensional Protein Structure Predictions.
147 P. Rogen and H. Bohr, *Math. Biosci.*, **182**, 167-181 (2003). A New Family of Global Protein Shape Descriptors.
148 D. Barnatan, *Topology*, **34**, 423-472 (1995). On the Vassiliev Knot Invariants.
149 A. Stasiak and J. H. Maddocks, *Nature (London)*, **406**, 251-253 (2000). Mathematics - Best Packing in Proteins and DNA.
150 K. A. Hoffman, R. S. Manning and J. H. Maddocks, *Biopolymers*, **70**, 145-157 (2003). Link, Twist, Energy, and the Stability of DNA Minicircles.
151 M. Levitt, *J. Mol. Biol.*, **170**, 723-764 (1983). Protein Folding by Restrained Energy Minimization and Molecular Dynamics.
152 G. A. Arteca, *Biopolymers*, **33**, 1829-1841 (1993). Overcrossing Spectra of Protein Backbones - Characterization of 3-Dimensional Molecular Shape and Global Structural Homologies.
153 G. A. Arteca, *Phys. Rev. E*, **49**, 2417-28 (1994). Scaling Behavior of Some Molecular Shape Descriptors of Polymer Chains and Protein Backbones.
154 G. A. Arteca, *Phys. Rev. E*, **51**, 2600-2610 (1995). Scaling Regimes Self-Entanglements in Very Compact Proteins.
155 G. A. Arteca and O. Tapia, *J. Chem. Info. Comp. Sci.*, **39**, 642-649 (1999). Characterization of Fold Diversity among Proteins with the Same Number of Amino Acid Residues.
156 C. T. Reimann, G. A. Arteca and O. Tapia, *Phys. Chem. Chem. Phys.*, **4**, 4058-64 (2002). Proteins in Vacuo. A Connection between Mean Overcrossing Number and Orientationally-Averaged Collision Cross Section.
157 P. Rogen and B. Fain, *Proc. Natl. Acad. Sci. (USA)*, **100**, 119-124 (2003). Automatic Classification of Protein Structure by Using Gauss Integrals.
158 J. H. White, *Am. J. Math.*, **91**, 693-& (1969). Self-Linking and Gauss-Integral in Higher Dimensions.
159 O. Gonzalez and J. H. Maddocks, *Proc. Natl. Acad. Sci. (USA)*, **96**, 4769-4773 (1999). Global Curvature, Thickness, and the Ideal Shapes of Knots.
160 J. R. Banavar, A. Maritan, C. Micheletti and A. Trovato, *Proteins: Struct. Func. Genet.*, **47**, 315-322 (2002). Geometry and Physics of Proteins.
161 J. R. Banavar, A. Maritan and F. Seno, *Proteins: Struct. Func. Genet.*, **49**, 246-254 (2002). Anisotropic Effective Interactions in a Coarse-Grained Tube Picture of Proteins.
162 J. R. Banavar and A. Maritan, *Rev. Modern Phys.*, **75**, 23-34 (2003). Colloquium: Geometrical Approach to Protein Folding: A Tube Picture.
163 J. R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan and A. Trovato, *J. Phys: Cond. Matter*, **15**, S1787-96 (2003). Tubes near the Edge of Compactness and Folded Protein Structures.
164 A. Maritan, C. Micheletti, A. Trovato and J. R. Banavar, *Nature (London)*, **406**, 287-90 (2000). Optimal Shapes of Compact Strings.
165 T. X. Hoang, A. Trovato, F. Seno, J. R. Benavar and A. Maritan, *Proc. Nat. Acad. Sci. (USA)*, **101**, 7960-7964 (2004). Geometry and Symmetry Presculpt the Free-Energy Landscape of Proteins.
166 F. Zu-Kang and M. J. Sippl, *Folding & Design*, **1**, 123-132 (1996). Optimum Superimposition of Protein Structures: Ambiguities and Implications.
167 C. A. Ouzounis, R. M. R. Coulson, A. J. Enright, V. Kunin and J. B. Pereira-Leal, *Nature Rev. Genet.*, **4**, 508-519 (2003). Classification Schemes for Protein Structure and Function.
168 A. D. Michie, C. A. Orengo and J. M. Thornton, *J. Mol. Biol.*, **262**, 168-185 (1996). Analysis of Domain Structural Class Using an Automated Class Assignment Protocol.
169 S. Jones, M. Stewart, A. Michie, M. B. Swindells, C. Orengo and J. M. Thornton, *Protein Sci.*, **7**, 233-242 (1998). Domain Assignment for Protein Structures Using a Consensus Approach: Characterization and Analysis.
170 L. Holm and C. Sander, *Science*, **273**, 595-602 (1996). Mapping the Protein Universe.
171 S. Dietmann and L. Holm, *Nature Struct. Biol.*, **8**, 953-957 (2001). Identification of Homology in Protein Structure Classification.
172 L. Holm and C. Sander, *Nucl. Acids. Res.*, **26**, 316-319 (1998). Touring Protein Fold Space with Dali/FSSP.
173 L. Holm and C. Sander, *Nuc. Acids. Res.*, **22**, 3600-3609 (1994). The FSSP Database of Structurally Aligned Protein Fold Families.
174 C. Hadley and D. T. Jones, *Structure*, **7**, 1099-1112 (1999). A Systematic Comparison of Protein Structure

Classifications: Scop, Cath and FSSP.
175 R. Day, D. A. C. Beck, R. S. Armen and V. Daggett, *Protein Sci.*, **12**, 2150-2160 (2003). A Consensus View of Fold Space: Combining Scop, CATH, and the Dali Domain Dictionary.
176 A. Harrison, F. Pearl, R. Mott, J. Thornton and C. Orengo, *J. Mol. Biol.*, **323**, 909-926 (2002). Quantifying the Similarities within Fold Space.
177 G. Apic, J. Gough and S. Teichmann, *J. Mol. Biol.*, **310**, 311-325 (2001). Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes.
178 S. Teichmann, A. G. Murzin and C. Chothia, *Curr. Opin. Struct. Biol.*, **11**, 354-363 (2001). Determination of Protein Function, Evolution and Interactions by Structural Genomics.
179 B. Rost, *Curr. Opin. Struct. Biol.*, **12**, 409-416 (2002). Did Evolution Leap to Create the Protein Universe?

Appendix: Hierarchical clustering.

The aim of clustering is to group a collection of objects (or observations) into subsets of "clusters", such that those within each cluster are more similar to one another than objects assigned to different clusters. There are two main elements in any clustering technique: the definition of similarity, or dissimilarity between objects, and the algorithm used to partition the data into clusters. Here I assume 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. I focus on the latter.
In hierarchical clustering, the data are regrouped into clusters through a series of partition, which 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, which proceed by series of fusions of the objects into groups, and *divisive* methods, which separate the objects successively into finer groupings. Again, I only focus on agglomerative methods, as these are the ones used for generating protein structure classifications.

An agglomerative hierarchical clustering technique proceeds through 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 clusters that are closest together. Differences between methods arise because of different ways of defining the distance between clusters. The four main agglomerative hierarchical clustering techniques are:

- ***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:

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

- ***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:

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

- ***Average linkage clustering***: the distance between the two clusters A and B is defined as the average of distances between all pairs of objects, where each pair is made up 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 the 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.

$$D(A,B) = Average\{d(i,j), \quad (i,j) \in (A \cup B)^2\} \qquad [A.4]$$

There is unfortunately no answer to the question on which of these techniques performs best. Clustering is an exploratory data analysis procedure. The choice of the actual technique used for clustering often comes from a very good understanding of the objects to be clustered.