Universitat Autònoma de Barcelona

Classification of Loops in Protein Structures: Applications on Loop Modeling and Protein Function

Memòria presentada per en Narcís Fernández Fuentes per optar al grau de Doctor en Ciències, Secció de Bioquímica. Treball realitzat sota la direcció del Prof. Francesc X. Avilés i Puigvert, catedràtic del Departament de Bioquímica i Biologia Molecular de la Universitat Autònoma de Barcelona i el Dr. Baldomero Oliva i Miguel, professor del Departament de Ciències Experimentals i de la Salut de la Universitat Pompeu Fabra de Barcelona.

Vist i plau dels directors,

Francesc X. Avilés i Puigvert

Baldomero Oliva i Miguel

Bellaterra, Maig de 2004

A mí familia, Elisenda incluida.

A través del cristal sucio, se veía la calle oscura y solitaria. La lluvia golpeaba incesante los adoquines gastados y ennegrecidos por el paso del tiempo. Se volvió, ella estaba acostada sobre su vieja cama. Su rostro reflejaba tranquilidad y serenidad. Su melena negra caía sobre la espalda. Mientras la observaba, le resultaba sorprendente admitir que se sentía tan magnéticamente atraído hacia ella, sin apenas conocerse. En ese momento, Ana se despertó. Le miró y se quedaron sin decir nada. Ella se levantó y se dirigió hacia él. Su mano tocó la de él, sin temblar, con suavidad. Le besó.

Ana y Hugo (Inacabada)

Table of Contents

List of Tables

List of Figures

Summary

Classification of Loops in Protein Structures: Applications on loop modeling and Protein Function

Narcís Fernandez-Fuentes

May, 2004 Bellaterra

This thesis is structured into five chapters. In chapter I, protein loops - the topics of this thesis work - are introduced. Also, a short description of biological databases and current protocols in sequence comparison are given. Chapters II to IV explore a major role that loop segments play in protein structures by using a structural bio-informatics approach: (i) the structural classification (ii) the relationship between the structure and function and (iii) the structure prediction of loops. The conclusive chapter V is devoted to several considerations that complement the conclusions given in previous chapters. Extensions of this thesis work are also suggested.

The research project on structural classification of loops, which was carried out by Dr. Oliva (Oliva et al. 1997), has been the starting point for all the other subsequent projects. In chapter II, a fully automated process for the structural classification of loops of kinases is presented. Several methodological improvements were made on the basis of Oliva's original work: (i) a newly introduced re-clustering process allows to avoid overlaps in classified loop clusters, (ii) a new web server was established to provide access and/or to query data through the internet, and (iii) cross referencing links were introduced with other biological databases. Chapter III focuses on two questions: the conservation of loop structures and functions and the extent of conservation of loop structures during evolution. An extensive analysis of a structural loop database of protein kinases was carried out. There are two main reasons why kinases were selected for the subject of this study: first, their critical biological relevance, and second the vast amount of functional information available in the literature and biological databases. Finally, in chapter IV, we apply ArchDB(Espadaler et al. 2004) for loop structure prediction. A Jack-knife test is performed to assess the usefulness of sequence information, which is included in the form of profiles in our structural clusters.

Chapter I - Introduction

Proteins are the machinery of life. As enzymes, they catalyze almost any biochemical reaction in the cell with very high accuracy and velocity, leading myriads of different products as simple as methane or as complex as alkaloids. They play the key role in light harvesting during photosynthesis, converting electromagnetic energy to a potential gradient. As fibrous proteins, they play a key role in the cellular cytoskeleton and body scaffold.

Each protein is a macromolecule built by linking together amino acid into contiguous chains. Proteins are produced in the cell according to a pattern specified by organism's DNA (DeoxyriboNucleic Acid), with RNA (RiboNucleic Acid) serving as an intermediate specification. The *primary structure* of a protein refers to its amino acid sequence. Regions of amino acids within the protein chain tend to arrange themselves into regular formation. These patterns are referred to as *secondary structure*, and can be recognized by the angles and hydrogen bonds patterns between the backbone atoms. Secondary structure is usually divided into three classes: α-helix, β-sheet and loops. The spatial arrangements of secondary structure elements lead to the *tertiary structure* or *fold*; occasionally protein structures contain compact modules called *domains*. A domain can be thought of as an independent-folding section of the protein. Finally, the *quaternary structure* refers to the assembling of individual protein chains into a supramolecular complex.

At the beginning of the structural biology loops were named as "random coils" and a secondary role was assigned. However, as it will be exposed in the following pages, loops play an important role both in structure and protein function.

1.1 Loops - Historical background

In 1958 Kendrew et al.(Kendrew et al. 1958) solved the first three dimensional structure of a protein, the sperm whale myoglobin. In 1965 Blake et al.(Blake et al. 1965) described the structure of hen egg-white lysozyme. Theses two structures confirmed the existence of dominant conformations of α-helices and β-sheets, just as proposed earlier by Linus Pauling, Robert Corey and Herman Brason(see review (Eisenberg 2003)). These regions have repetitive conformation of ϕ and φ dihedral angles and conserved hydrogen bond patterns in protein structures. There was however a second type of conformation in the structures with non-repetitive patterns connecting α–helices or β-strands: the loop regions.

Originally, loops were considered as an "irregular conformation" and sometimes misnamed as "random coils". Due to their flexibility and non-periodic nature, loops long escaped structural classifications. It was in 1968 when Venkatachalam (Venkatachalam 1968) introduced three categories of four residue β-turns. As more structures were solved new structurally conserved patterns were found. Venkatachalam's classification was subsequently extended(Richardson 1981). These concepts were updated and refined in a classification scheme by Thornton and co-workers (Wilmot and Thornton 1988; Hutchinson and Thornton 1994). Three residue chain reversals, termed γ-turns first identified in 1972 by Matthews(Matthews 1972), were also classified(Rose et al. 1985; Milner-White 1987).

Several studies focused on the conformation of loops linking specific secondary. Analysis of ββ loops connecting two adjacent anti-parallel strands (i.e. of the β-hairpin) identified specific preferences for certain β-turn families and found novel, commonly occurring substructures (Milner-White and Poet 1986; Sibanda et al. 1989). Subsequent analyses identified commonly occurring structural families, often accompanied by sequence patterns, for αα, αβ, βα, and ββ arches(Edwards et al. 1987; Thornton et al. 1988; Rice et al. 1990; Colloch and Cohen 1991; Efimov 1991; Efimov 1993). More general classification of short to medium loop lengths have been done (Donate et al. 1996; Oliva et al. 1997; Oliva et al. 1998; Wojcik et al. 1999; Burke et al. 2000) and recently we presented ArchDB, which classified loops up to 13 residues long length(Espadaler et al. 2004) according to bounding secondary structures (see figure 1.1).

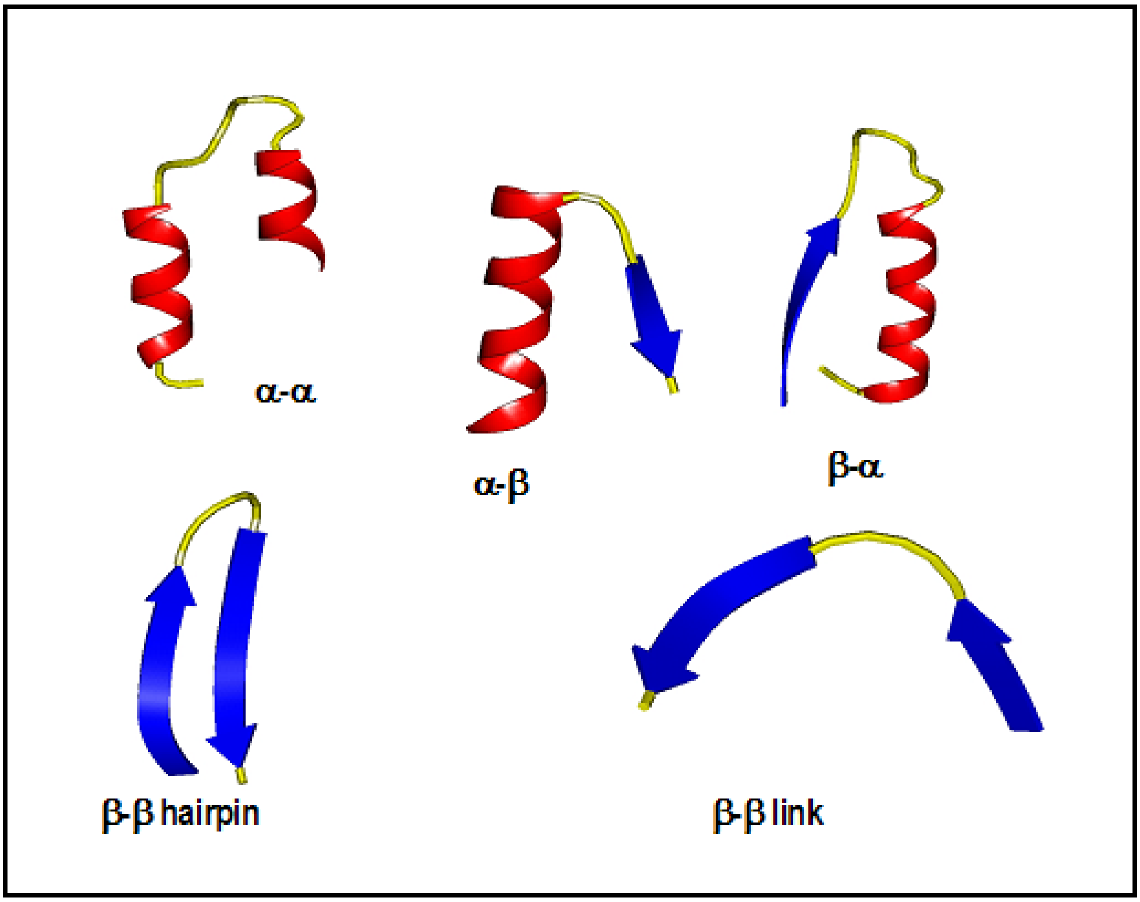

Figure 1.1 Type of loops according to connected secondary structures.

1.2 Importance of the loops in protein

Loops play important roles in protein function, stability and folding(Fetrow 1995). They are found mainly at the surface of globular proteins(Leszczynski and Rose 1986). Loops were long considered as random coils, but now are recognized as an additional structural class. Loops represent an important part of the protein structure. In 1986 Leszczynski & Rose(Leszczynski and Rose 1986) carried out an analysis over 67 known proteins structures. They found that the 26% of residues were located in α-helices, 19% in β-sheets, and 47% in non-repetitive regions (26% in turns –short loops- and 21% in loops). In a recent work, it was found that over a non-redundant set of proteins, 57% of catalytic residues were located in loops (Dr. P. Aloy, personal communication). On the other hand, loops sometimes correspond to highly flexible regions and therefore difficult to be described either by X-ray crystallography or NMR spectroscopy. To provide structural prediction for loop segments is of great interest and the classification of loops may address this problem.

1.2.1 Loops and protein function

There are many examples in the scientific literature that relate loops with protein function. They can play a wide repertoire of roles related to protein function: (i) recognition sites (CDRs(Kim et al. 1999), (ii) protein-protein interactions: signaling cascades (Zomot and Kanner 2003; Bernstein et al. 2004) , dimerization(Fritz-Wolf et al. 1996) , PDZ-motifs(Feng et al. 2003) , (iii) ligand binding (p-loop(Saraste et al. 1990) EF-hands(Kawasaki and Kretsinger 1995), NAD(P) binding loops(Wierenga et al. 1986), glycin-rich-loop(Schenk and Snaar-Jagalska 1999)), (iv) DNA-binding (helix-turn-helix motifs(Tainer et al. 1995), M13 phage(Coleman et al. 1986)); (v) forming enzyme active sites (e.g. Ser-Thr kinases(Johnson et al. 1998) or serine proteases(Wlodawer et al. 1989)), see figure 1.2.

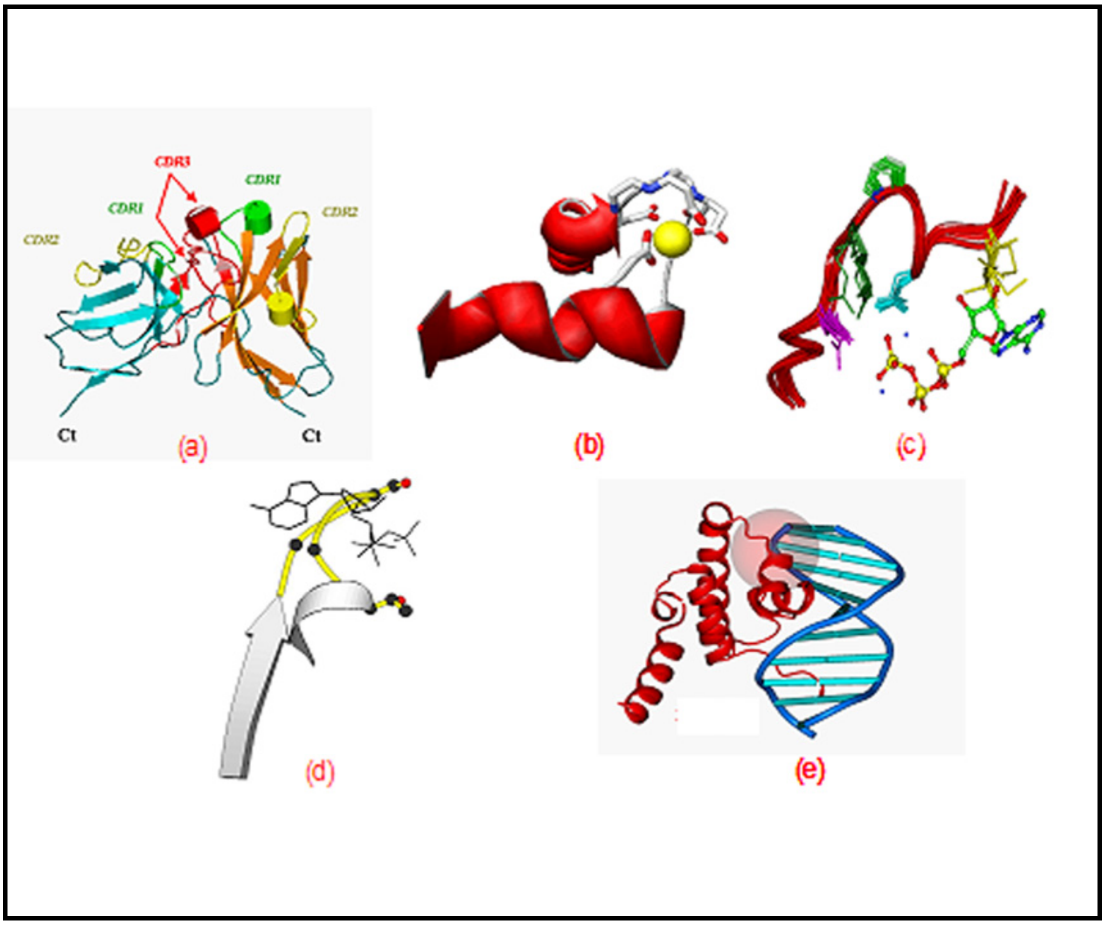

Figure 1.2. Some examples of functional loops are shown: (a) Cdrs of immunoglobulins; (b) EF-hand; (c) Catalytic loop of Ser-Thr kinases; (d) p-loop; (e) helix-turn-helix DNA interaction motif.

Functional differences between the members of the same protein family are usually a consequence of structural differences on the protein surface. In a given fold, structural variability is a result of substitutions, insertions and deletions of residues between members of the family. Such changes frequently correspond to exposed loop regions that connect elements of secondary structure in the protein fold(Blouin et al. 2004). Thus, loops often determine the functional specificity of a given protein framework, contributing to active and binding sites(Fetrow 1995).

Enzyme function often involves conformational changes in protein structures. Loop flexibility plays a vital role in correctly positioning catalytically important residues. Motions range form a simple bending and stretching of bonds to subunit rotations and translations. There are a few examples of documented loops that are involved in protein motion related with function. Gunasekaran et al.(Gunasekaran et al. 2003) have identified "triggering" loops whose conformational change is required for the catalytic process of β1,4-Galactosyltransferase. Zgiby et al. have described that the flexibility of a loop was important for the correct functioning of the class II FBPA (Fructose 1,6-bisphosphate aldolase)(Zgiby et al. 2002). It is well known the conformational change in the LID region of NMPKs (Nucleotide monophosphate kinases) linked with substrate binding and catalysis(Sinev et al. 1996), in triosephosphate isomerases(Joseph et al. 1990) and also the conformational change in the activation loop of the protein kinases(Johnson et al. 1996; Adams 2003).

1.2.2 Loops and protein structure

A large body of experimental and theoretical evidence suggests that local structural determinants are frequently encoded in short segments of protein sequence. Local sequences-sequence-structure relationships derived from local structure/sequence analyses could significantly enhance the capacities of protein structure prediction methods(Yang and Wang 2003). The reports of Salem et al.(Salem et al. 1999), Wood & Pearson (Wood and Pearson 1999) and Lupas et al. (Lupas et al. 2001) suggested that folds are mainly made up of a number of simple local units of super-secondary structures (motifs), formed by few secondary structures connected by loops (see figure 1.3). The smallest motif is two secondary structures connected by a loop.

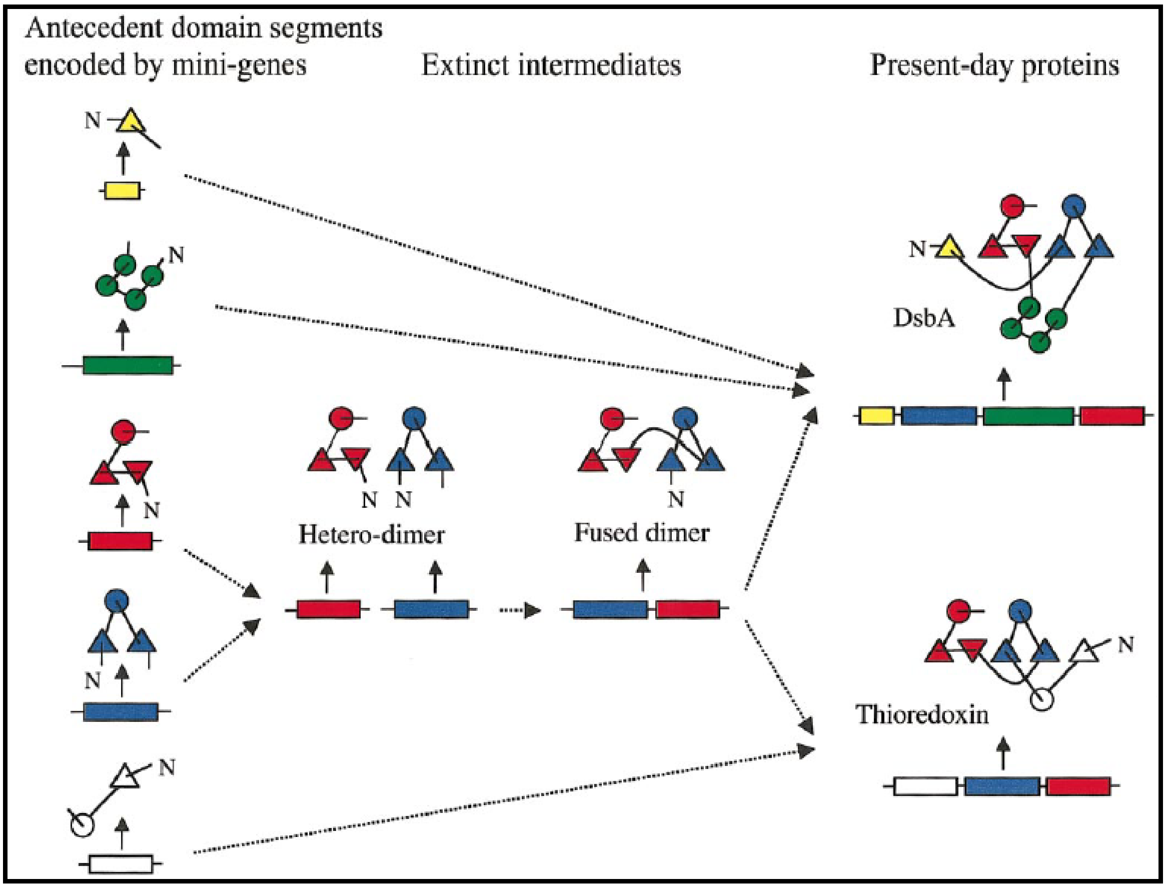

Figure 1.3. Schematic showing the hypothetical evolutionary scenario that might have led to the evolution of thioredoxin and DsbA from a series of different antecedent domain segments (ADSs). α-Helices are represented by circles and β-strands by triangles (taken from Lupas et al 2000).

Protein loops are also important in the folding process of proteins. Single substitutions in a loop residue could destabilize the whole protein structure(Hoedemaeker et al. 1993). A stabilizing role of surface loops in eightfold beta alpha proteins has been shown (Ulfer and Kirschner 1992). Experimental studies show that for some proteins the formation of a single loop is the rate-determining step, whereas for others, a loop can misfold to serve as the hinge loop region for domain-swapped species(Linhananta et al. 2002). Batori et al. have demonstrated the relationship between loop elongation and stability in a fibronectin type III domain(Batori et al. 2002). Fersht suggested in 2000 the importance of loop-loop long-range interaction in the folding process of proteins(Fersht 2000). Recently, Krishna et al. have proven the active role of loops in the folding/unfolding process in cytochrome c.(Krishna et al. 2003) and Collinet et al. in phosphoglycerate kinase(Collinet et al. 2001).

Furthermore, loops are involved in the thermo stability of proteins. The stability of proteins has been studied from two points of view: (i) to study and compare protein structures in folded state, and (ii) to study the chain entropy and its reduction to favor the folded state. For both approaches, knowledge of loops is of great interest. In the first case, the stability of thermophilic proteins has been attributed to greater hydrophobicity, better atom packing, smaller and fewer cavities, increased amount of buried surface area upon oligomerization, residue substitution within and outside the secondary structures, decrease of thermolabiles residues, increased helical content, increased polar surface area, better hydrogen bonding, and better salt bridge; but also, the shortening of loops and an increased occurrence of proline residues in loops may also contribute to thermo stability (see review (Kumar and Nussinov 2001)). In connection with the second issue, Zhou(Zhou 2004) reviewed the importance of loops in entropy-based strategies, because loop and their lengths are important elements that affect the total value of the chain entropy.

1.3 Prediction of loop structure

Despite recent improvements in the techniques of structure determination at atomic level, in X-ray diffraction and Nuclear Magnetic Resonance (NMR) spectroscopy, there is a large difference between known protein sequences (~ 1.5 million)(Boeckmann et al. 2003) and protein structures (~ 24 000)(Berman et al. 2002). In the absence of an experimentally determined structure, *ab initio* and threading methods or comparative modeling methods can sometimes provide a useful 3D model.

In general, these methods tend to predict correctly the protein core when the structure of a close homologue of the target protein is available, but not the loop regions. Errors in loops are the dominant problem in comparative modeling above 35% sequence identity. In this range of overall identity, loops among the homologues vary while the core regions are still relatively conserved and aligned accurately(Fiser et al. 2002). The recent improvements of the performance of fold prediction and homology modeling methods in successive CASP experiments (Critical Assessment of Techniques for Protein Structure Prediction) (Venclovas et al. 2001) have not proved to be as successful in loop model building.

Loop prediction can be seen as a mini protein-folding problem. The correct conformation of a given segment of a polypeptide chain has to be calculated from the sequence of the segment influenced by the core limb regions that span the loop and by the structure of the rest of the protein that cradles the loop. Many loop-modeling procedures have been described. Similarly to the prediction of whole protein structures, there are two basic approaches in loop structure prediction: (i) database search methods (also named knowledge-based), and (ii) *ab initio* (also named conformational search).

1.3.1 Database search methods

The knowledge-based methods, which were initiated by Jones(Jones and Thirup 1986), consist of finding a segment of main-chain that fits two stem regions of the loop. The stems are defined as the main-chain atoms that precede and follow the loop, but are not part of it. They span the loop and are part of the core of the spanned secondary structures. The search is performed through a database of many known protein structures, not only homologs of the modeled protein. Usually, many different alternative segments that fit the stem residues are obtained, and possibly sorted according to geometric criteria or sequence similarity between the template and target loop sequences. The selected segments are then superposed and annealed on the stem regions.

In the context of homology modeling, the most difficult unsolved problem after sequence alignment is the prediction of loop structures(Mosimann et al. 1995; Sali 1995) . Classification of loop conformations is of major benefit in the prediction of protein structure by comparative (homology) modeling following a knowledge-based approach. Blundell and coworkers(Donate et al. 1996; Burke and Deane 2001; Deane and Blundell 2001) have employed their loop database, Sloop, in loop structure prediction with encouraging results. A work of Wojcik et al. (Wojcik et al. 1999) showed that they were able to predict 94% of query loops with Root Mean Square Deviation (RMSD) 3.8 A for 8-residues loops. However in certain cases, the database search methods become obvious when canonical loop conformations exist. This has been consistently exposed in the case of Complentary Determining Region (CDR) loop predictions(Chothia and Lesk 1987; Martin and Thornton 1996; Oliva et al. 1998).

The database methods are limited by the exponential increase in the number of possible conformations in agreement with the ring closure as a function of loop length. Only segments of less than 7 residues had most of their conceivable conformations present in the database of known protein structures (Fidelis et al. 1994) (Lessel and Schomburg 1999). However, a recent work published by Du et al (Du et al. 2003) argued that there exists sufficient coverage to model even a novel fold using fragments from the Protein Data Bank, as the current database of known structures has increased enormously in the last few years (see figure 1.4).

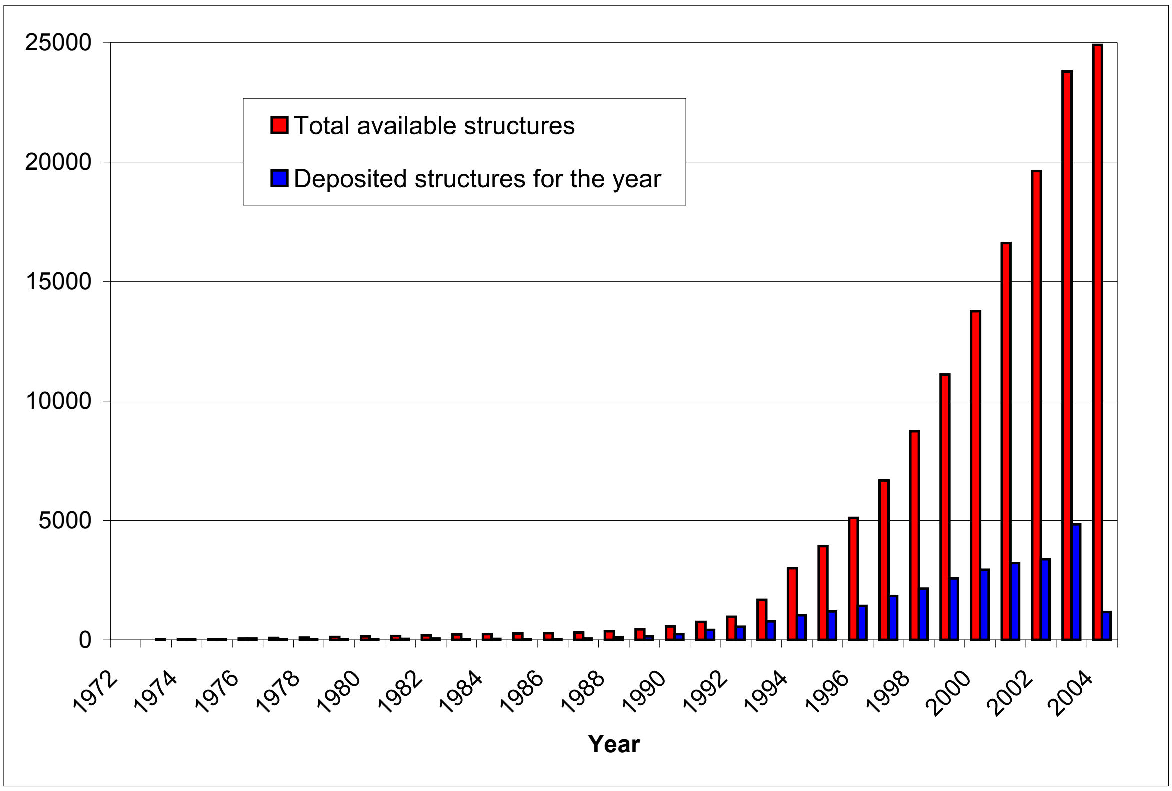

Figure 1.4. Protein Data Bank content growth (taken from http://www.rcsb.org 01-Apr-2004)

1.3.2 *Ab initio* methods

The *ab initio* methods are based on a conformation search or enumeration of conformation, or decoys, in a given environment, guided by a scoring or energy function. There are many such methods, exploiting different protein representations, energy function terms and optimization or enumeration algorithm (Moult and James 1986; Fiser et al. 2000; Xiang et al. 2002).

Candidate loops (up to 12 residues) with conformations close to the native can always be found if the number of loops generated is large enough(Rapp and Friesner 1999). Usually native conformation does not mean lowest energy(Smith and Honig 1994; Pellequer and Chen 1997). Thus, limitations of this approach are two: (i) the accuracy of the applied scoring function that often are not accurate enough to properly rank the many alternative conformations, especially in case of longer loops; and (ii) how to sort all these loop candidates so that the best one has the lowest energy. Despite of recent advances in *ab initio* loop modeling approaches, modeling loops beyond 8-10 residues remains uncertain.

1.3.3 Combined methods

Combined methods use both database search and *ab initio* methods. The underlying idea is the use of database search methods to find candidate loops to a given target loop and subsequently evaluated and optimized in the target protein.

An example of a combined algorithm is that of Martin et al. (1989), in which antibody hypervariable loops were predicted using a database search followed by reconstruction of sections of the predicted loops *ab initio* and addition of side chains using the CONGEN conformational searching algorithm(Bruccoleri and Karplus 1987). This idea has been also applied by van Vlijmen & Karplus(van Vlijmen and Karplus 1997) selecting loops from a fragment databank, followed by the optimization and ranking of the possible fragments using the CHARMM energy function(Brooks et al. 1983). The method has been tested for loop of different lengths (4 to 16 residues) showing usefulness for up to nine residues loops. Recently, Deane et al.(Deane and Blundell 2001) presented the CODA a combination of two algorithms: FREAD (a knowledge-based method) and PETRA(Deane and Blundell 2000) (an *ab initio* method).

1.4 Bioinformatic Tools

The enormous quantities of biological sequence and structural data produced represent a wealth of information. However, this information has to be classified in order to be understanding and useful. Some of the most important databases (some of them cited within this work) are briefly described. Also, a short description of the main sequence comparison protocols used in bioinformatic is exposed.

1.4.1 Protein Structure Databases

1.4.1.1 PDB (http://www.rcsb.org)

The protein Data Bank(Berman et al. 2002) is the single worldwide archive of structural data of biological macromolecules. It was established at Brookhaven National Laboratories in 1971 as an archive for biological macromolecules crystal structures. In the 1980s the number of deposited structures began to increase dramatically due to the improved technology for all aspects of crystallographic process, the addition of structures determined by NMR methods, and changes in the community views about data sharing. The current numbers of deposited structures is around 25000.

1.4.1.2 SCOP (http://scop.mrc-lmb.cam.ac.uk/scop)

Structural Classification of Proteins (SCOP) database(Lo Conte et al. 2002) provides a detailed and comprehensive description of the relationships of known structures. The classification is on hierarchical levels: the first two levels, family and superfamily, describe near and distant evolutionary relationships; the third, fold, describes topological relationships.

On family level protein are clustered based of one of two criteria that imply their have a common evolutionary origin: first, all the proteins have 30% residue identity or greater; second, proteins with lower sequence identities but whose functions and structures are very similar. On superfamily level, proteins share low sequence identities but both structure or functional features suggest a common evolutionary origin. Finally, superfamilies and families are defined as having common fold if their proteins have the same major secondary structures in the same arrangement and with the same topological connections.

1.4.1.3 Dali / FSSP (http://www.ebi.ac.uk/dali)

Dali/FSSP(Holm and Sander 1997) is a fully automatic classification method of all known 3D protein structures. The classification is derived using and automatic structure alignment program (Dali) for all-against-all comparison of structure in the Protein Data Bank. From the resulting enumeration of structural neighbors a discrete fold classification is derived in three steps: (i) sequence-related families are covered by a representative set of protein chains; (ii) proteins chains are decomposed into structural domains based on the recurrence of structural motifs; (iii) folds are defined as tight clusters of domains in fold space. The FSSP database has one entry per representative, reporting the structural alignments with the representative's sequence homologues.

1.4.2 Protein Sequence Databases

1.4.2.1 SWISS-PROT (http://www.ebi.ac.uk/swissprot/)

The SWISS-PROT database(Bairoch and Apweiler 2000) is currently seen as one of the best annotated general protein sequence databases. SWISS-PROT is a curated protein sequence database that provided a high level of annotation, such as description of function, domains structure, post-translational modifications, variants, etc. A minimal level of redundancy and a high level of integration with other databases is another key aim. The current release of SWISS-PROT (version 43) contains 149913 entries.

TrEMBL is a computational derived supplement to SWISS-PROT that contains translations of EMBL nucleotide sequence entries. Annotations for TrEMBL entries are automatically derived and are generally not of the same high quality as SWISS-PROT entries. Both databases can be accessed and searched through the Sequence Retrieval System (SRS) (Etzold et al. 1996).

1.4.2.2 PIR (http://pir.georgetown.edu)

The Protein Information Resource (PIR)(Wu et al. 2002) is a division of the United Stated National Biomedical Research Foundation (NBRF). The PIR database strives to be comprehensive, accurate, consistently annotated and well organized. The last release of PIR (2.52) contains 1,226,875 entries. However, from a curation and annotation point of view, SWISS-PROT database is probably more desirable.

By the time of the writing of this thesis it has been published the merging between SWISS-PROT and PIR databases, to establish the United Protein Database (UniProt), a central resource of protein sequence and function.

1.4.3 Other Databases

1.4.3.1 Organism Specific Databases

For many model organisms, there exist specialized genome databases that have been expertly annotated by researchers with extensive knowledge of the biology of specific organisms. These databases exist for a variety of organisms, but three of these databases (SGD, FlyBase and Ensembl) are exceptional, both in terms of methodology and annotation quality.

The FlyBase(Ashburner and Drysdale 1994) (http://flybase.bio.indiana.edu/) resource was one of the earliest model organism databases. This database endeavors to capture large amounts of biological information regarding the complete genome sequence of *Drosophila melanogaster*. Detailed information has been gradually added to the database which includes: genes, proteins, genetic elements, chromosomal maps, aberrations, literature references and images.

The *Saccharomyces* Genome Database (SGD)(Christie et al. 2004) (http://genome-www.stanford.edu/Saccharomyces/) is a similar resource concerning the complete genome of *Saccharomyces cerivisiae* and related yeast strains

Finally, the Ensembl resource(Birney et al. 2004) (http://www.ensembl.org) is a joint initiative of the European Bioinformatic Institute and the Wellcome Trust Sanger Institute. The project is a complete pipeline of genome assembly, gene prediction, large-scale annotation and analysis of the draft human genome sequence. Ensembl also contains information pertaining to predicted genes and proteins, cytological markers, single nucleotide polymorphisms, protein families, and other information. Current distribution of Ensembl (build 34b) encloses 23,531 predicted genes and 31,609 gene transcripts.

1.4.3.2 Protein Domain and Family Database: Pfam and InterPro

Pfam(Bateman et al. 2004) (http://www.sanger.ac.uk/Software/Pfam/) is a database of multiple alignments of protein domains constructed using profiles hidden Markov models (HMM) (see below). The database is comprehensive well curated and very useful for determining the presence of domains or motifs within protein sequences. Domain families are constructed by building a seed alignment of related sequences. Great care is taken in the building of seed alignment. A profile HMM is constructed form a finished seed alignment using the HMMER package(Eddy 1998). This profile is used to search over SWISS-PROT database and a multiple sequence alignment for the full family is generated. Theses finished full alignments are stored in the PfamA database. In order to enrich the database further and to cover sequences in SWISS-PROT that are not already part of the PfamA database, another automatic protocol is used. The Domainer algorithm(Sonnhammer and Kahn 1994) is applied to SWISS-PROT BLAST similarity data, and used to automatically detect conserved domains and motifs within proteins. Profiles are once again constructed from these automatically generated alignments, and are used to detect more family members that are built into a final alignment. These results are stored in the PfamB division. Currently, Pfam contains 7426 families (version 13.0 April 2004).

Interpro initiative (Apweiler et al. 2001) (http://www.ebi.ac.uk/interpro/) has linked the major family databases into a single resource. This database contains families from the Pfam, Prosite, Prints, ProDom, SMART and TIGR-Fams resources. InterPro is an exceptionally useful resource as it allows the strengths of each separate family database to be combined into a single comprehensive resource for information regarding protein sequence motifs and domains. Current InterPro release 7.2 contains 10,709 entries, representing 2,411 domains, 8,035 families, 197 repeats, 20 binding sites and 20 post-translational modification sites.

1.4.4 Algorithms for sequence analysis

1.4.4.1 Pair-wise sequence comparison methods

A simplistic yet systematic approach to sequence homology detection was first developed by Fitch in 1966(Fitch 1966). This method used a substitution matrix based on the number of nucleotide mutations required to translate between amino acids. This matrix was used to calculated similarity scores for all possible ungapped alignment of a pair of sequences, and hence the significance of the homology. The first systematic sequence comparison procedure which allowed for gaps and insertions in the sequences was published by Needleman and Wunsch in 1970(Needleman and Wunsch 1970). This work was a major breakthrough at the time, offering the capability to automatically detect homology between much more divergent sequences than was previously possible. Through the use of a dynamic programming algorithm the method was also very efficient, and could be used to compare large numbers of sequences. In 1981 major improvements were made to the substitution matrix by Dayhoff and Schwartz(Dayhoff and Schwartz 1981).

The next major advance was the result of the work by Smith and Waterman in 1981(Smith and Waterman 1981), which introduced local scoring and included affined gap penalties. An extension to the dynamic programming algorithm developed by Needleman & Wunsch allowed the detection of local similarities. This allows for a match between individual domains (subsequences) belonging to larger multi-domain sequences.

The dynamic programming algorithm used by Smith and Waterman guarantees an optimal local alignment, and hence the maximum score, for a given pair of sequences and a substitution matrix. It is however possible to increase the efficiency of the algorithm by adding heuristics which consider a window of residues along the sequence. The optimal alignment is no longer guaranteed but something very close is usually achieved. In 1988 Pearson and Lipman introduced a heuristic method called FASTA(Pearson and Lipman 1988) for Fast Alignment Search Tool, and in 1990 another by Altschul et al. called BLAST(Altschul et al. 1990) for Basic Local Alignment Sequencing Tool. Further improvements were made available introducing substitution matrices, such BLOcks Substitution Matrix (BLOSUM) by Henikoff and Henikoff in 1992 (Henikoff and Henikoff 1992) or Percent Accepted Mutations (PAM) matrices(Dayhoff et al. 1978). Substitution matrices give a score for any given pair of aligned residues according to their substitution likelihood in a protein family at a certain evolutionary distance. The BLAST and FASTA programs are still in common use, although some continuing work has been done to improve the BLAST algorithm, and the loss in sensitivity due to the heuristic effect is now far less.

1.4.4.2 Profile sequence comparison methods

The problem with pair-wise sequence comparison methods is that each position is treated with equal importance and they use the same substitution matrix for every position in the sequence. In reality some positions are far more important than others. For instance, residues of the hydrophobic core of a protein are less variable than site in loop regions, and so a difference between a pair of proteins in core residues should be more heavily penalized than at a site on the surface. Furthermore, some positions are conserved for different reason than other (i.e. catalytic residues, recognition sites, …), so certain types of substitution should be more penalized than the same type of substitution at a different site.

By considering the 3D structure of a protein, some properties of the residues at each site can be characterized. Information can also be gained by creating a multiple sequence alignment of related sequences, and observing the frequency of the occurrence of different residues at each position. These two sources of information can be used together or separately to characterize the features of the sequences at different positions. A Position Specific Scoring Matrix (PSSM) representing these characteristics can be constructed. This is called also a profile, and can be used to match and align sequences of far more distantly related proteins than simple pair-wise sequence homology methods.

An early attempt at using a profile to identify protein sequence homology was made by Taylor in 1986(Taylor 1986). He demonstrated that the use of a profile-based procedure using what he called “templates fingerprints” can be used to identify conserved features in immunoglobulin sequences, and attempted to define a general algorithm for the recognition of sequence patterns in globular proteins. Further template generation and matching was subsequently carried out on the globins by Bashford et al in 1987(Bashford et al. 1987), but the focus of this work was to identify the sequence features that determine a protein fold, rather than developing a general procedure for sequence alignment and searching. As the conserved features were found to be common to all globins, they were able to distinguish globins from any other sequences.

Growing out of this work, Gribskov et al. presented PROFMAKE(Gribskov et al. 1987) a new method with specific position gap penalties and a dynamic algorithm for pair-wise comparison.

1.4.4.3 Iterative procedures

The most recent developments in the field of profile-based sequence searching that have made a big difference are the use of iterative procedures. In 1997 Altchul et al. presented Position Specific Iterated – BLAST, PSI-BLAST(Altschul et al. 1997), which is probably the most commonly used profile-based method. PSI-BLAST is an improvement, and more importantly, an extension of the original BLAST algorithm. The procedure requires a large database of known protein sequences that it uses to iteratively search for homologues to build a profile. The initial query sequence is searched against the large database for close homologues, which are then aligned. Position specific score matrices are constructed form the alignment and used to search the database for further homologues not detectable by the initial query sequence search using BLAST. The procedure repeats itself a number of times improving the profile by adding more homologous sequences from the large database to the alignment. The resulting automatically generated profile can be used to search for sequences related to the initial query sequence.

1.4.4.4 HMMs

Hidden Markov Models (HMMs) introduced by David Haussler and co-workers(Krogh et al. 1994) belong to the cited *profile sequence comparison* methods. HMM put the profile methods on firmer mathematical ground by casting statistical the problem within a tight probabilistic framework. This fact resolves many of the problems associated with profile analysis, improved performance and by 1996 were two independent (freely available) software packages being applied to larger-scale problems(Eddy 1996).

Without the probabilistic framework provided by the hidden Markov model theory, previous methods had been assigning the position specific scores and gap penalties in a relative *ad hoc* manner. A HMM is a type of dynamic statistical profile, constructed by analyzing the distribution of amino acids and other features in a training set of proteins. The HMM concept is more complex than a simple profiles and can be visualized as a *finite state machine*. They use position-specific scores for amino acid (or nucleotides) and position specific scores for opening and extending and insertion or deletion. This property of profiles captures important information about the degree of conservation at various positions in the multiple alignments, and the varying degree to which gaps and insertions are permitted. HMMs have been used to detect distant relationship between proteins, prediction of trans-membrane segments(Krogh et al. 2001), CpC islands, gene prediction and signal peptides(Kall et al. 2004).

1.5 References

Adams, J. A. (2003). "*Activation loop phosphorylation and catalysis in protein kinases: is there functional evidence for the autoinhibitor model?*" Biochemistry **42**(3): 601-7.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "*Basic local alignment search tool*." J Mol Biol **215**(3): 403-10.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "*Gapped BLAST and PSI-BLAST a new generation of protein database search programs.*" Nucleic Acids Res. **25**: 3389-3402.

Apweiler, R., T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni and F. Servant (2001). "*The InterPro database, an integrated documentation resource for protein families, domains and functional sites*." Nucleic Acids Res **29**(1): 37-40.

Ashburner, M. and R. Drysdale (1994). "*FlyBase--the Drosophila genetic database*." Development **120**(7): 2077-9.

Bairoch, A. and R. Apweiler (2000). "*The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000*." Nucleic Acids Res **28**(1): 45-8.

Bashford, D., C. Chothia and A. M. Lesk (1987). "*Determinants of a protein fold. Unique features of the globin amino acid sequences*." J Mol Biol **196**(1): 199-216.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats and S. R. Eddy (2004). "*The Pfam protein families database*." Nucleic Acids Res **32 Database issue**: D138-41.

Batori, V., A. Koide and S. Koide (2002). "*Exploring the potential of the monobody scaffold: effects of loop elongation on the stability of a fibronectin type III domain*." Protein Eng **15**(12): 1015-20.

Berman, H. M., T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain and e. al. (2002). "*The Protein Data Bank.*" Acta Crystallogr. D. Biol. Crystallogr. **58**: 899-907.

Bernstein, L. S., S. Ramineni, C. Hague, W. Cladman, P. Chidiac, A. I. Levey and J. R. Hepler (2004). "*RGS2 binds directly and selectively to the M1 muscarinic acetylcholine receptor third intracellular loop to modulate Gq/11alpha signaling*." J Biol Chem.

Birney, E., T. D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, T. Cutts, T. Down, E. Eyras, X. M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. R. Hotz, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, K. C. Woodwark, G. Cameron, R. Durbin, A. Cox, T. Hubbard and M. Clamp (2004). "*An overview of ensembl*." Genome Res **14**(5): 925-8.

Blake, C. C., D. F. Koenig, G. A. Mair, A. C. North, D. C. Phillips and V. R. Sarma (1965). "*Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution*." Nature **206**(986): 757-61.

Blouin, C., D. Butt and A. J. Roger (2004). "*Rapid evolution in conformational space: A study of loop regions in a ubiquitous GTP binding domain*." Protein Sci **13**(3): 608-16.

Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout and M. Schneider (2003). "*The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003*." Nucleic Acids Res **31**(1): 365-70.

Brooks, B., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus (1983). " *A program for macromolecular energy minimisation and dynamics calculations*." J. Comp. Chem. **4**: 87-217.

Bruccoleri, R. E. and M. Karplus (1987). "*Prediction of the folding of short polypeptide segments by uniform conformational sampling.*" Biopolymers **26**: 137-168.

Burke, D., C. Deane and T. Blundell (2000). "*Browsing the Sloop database of structurally classified loops connecting elements of protein secondary structure.*" Bioinformatics **16**: 513-516.

Burke, D. F. and C. M. Deane (2001). "*Improved protein loop prediction from sequence alone*." Protein Eng **14**(7): 473-8.

Chothia, C. and A. M. Lesk (1987). "*Canonical structures for the hypervariable regions of immunoglobulins.*" J. Mol. Biol. **196**: 901-917.

Christie, K. R., S. Weng, R. Balakrishnan, M. C. Costanzo, K. Dolinski, S. S. Dwight, S. R. Engel, B. Feierbach, D. G. Fisk, J. E. Hirschman, E. L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C. L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, D. Botstein and J. M. Cherry (2004). "*Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms*." Nucleic Acids Res **32 Database issue**: D311-4.

Coleman, J. E., K. R. Williams, G. C. King, R. V. Prigodich, Y. Shamoo and W. H. Konigsberg (1986). "*Protein chemistry-nuclear magnetic resonance approach to mapping functional domains in single-stranded DNA binding proteins*." J Cell Biochem **32**(4): 305-26.

Collinet, B., P. Garcia, P. Minard and M. Desmadril (2001). "*Role of loops in the folding and stability of yeast phosphoglycerate kinase*." Eur J Biochem **268**(19): 5107-18.

Colloch, N. and F. E. Cohen (1991). "*Beta-breakers: an aperiodic secondary structure.*" J Mol Biol **221**: 603-613.

Dayhoff, M. O. and R. M. Schwartz (1981). "*Evidence on the origin of eukaryotic mitochondria from protein and nucleic acid sequences*." Ann N Y Acad Sci **361**: 92-104.

Dayhoff, M. O., R. M. Schwartz and B. C. Orcutt (1978). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure. M. O. Dayhoff. Washington D.C., National Biomedical Research Foundation. **5, supplement 3:** 356-352.

Deane, C. M. and T. L. Blundell (2000). "*A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins*." Proteins **40**(1): 135-44.

Deane, C. M. and T. L. Blundell (2001). "*CODA: a combined algorithm for predicting the structurally variable regions of protein models*." Protein Sci **10**(3): 599-612.

Donate, L. E., S. D. Rufino, L. H. Canard and T. L. Blundell (1996). "*Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction*." Protein Sci **5**(12): 2600-16.

Du, P., M. Andrec and R. M. Levy (2003). "*Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update*." Protein Eng **16**(6): 407-14.

Eddy, S. (1998). "*Profile Hidden Markov Models.*" Bioinformatics **14**: 755-763.

Eddy, S. R. (1996). "*Hidden Markov models*." Curr Opin Struct Biol **6**(3): 361-5.

Edwards, M., M. J. E. Sternberg and J. Thornton (1987). "*Structural and sequence patterns in the loops of beta alpha beta units.*" Prot. Eng. **1**: 173-181.

Efimov, A. V. (1991). "*Structure of coiled beta-beta-hairpins and beta-beta-corners*." FEBS Lett **284**(2): 288-92.

Efimov, A. V. (1993). "*Patterns of loops regions in proteins*." Curr Opin Struct Biol **3**: 379-384.

Eisenberg, D. (2003). "*The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins*." Proc Natl Acad Sci U S A **100**(20): 11207-10.

Espadaler, J., N. Fernandez-Fuentes, A. Hermoso, E. Querol, F. X. Aviles, M. J. Sternberg and B. Oliva (2004). "*ArchDB: Automated protein loop classification as a tool for Structural Genomics.*" Nucleic Acids Res **32**: D185-D188.

Etzold, T., A. Ulyanov and P. Argos (1996). "*SRS: information retrieval system for molecular biology data banks*." Methods Enzymol **266**: 114-28.

Feng, W., Y. Shi, M. Li and M. Zhang (2003). "*Tandem PDZ repeats in glutamate receptor-interacting proteins have a novel mode of PDZ domain-mediated target binding*." Nat Struct Biol **10**(11): 972-8.

Fersht, A. R. (2000). "*Transition-state structure as unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism*." PNAS **15**: 1525-1529.

Fetrow, J. S. (1995). "*Omega loops: nonregular secondary structure significant in protein function and stability.*" FASEB **9**: 708-717.

Fidelis, K., P. S. Stern, D. Bacon and J. Moult (1994). "*Comparison of systematic search and database methods for constructing segments of protein structure.*" Protein Eng. **7**: 953-960.

Fiser, A., R. K. Do and A. Sali (2000). "*Modeling of loops in protein structures.*" Protein Sci **9**: 1753-1773.

Fiser, A., M. Feig, C. L. Brooks, 3rd and A. Sali (2002). "*Evolution and physics in comparative protein structure modeling*." Acc Chem Res **35**(6): 413-21.

Fitch, W. M. (1966). "*An improved method of testing for evolutionary homology*." J Mol Biol **16**(1): 9-16.

Fritz-Wolf, K., T. Schnyder, T. Wallimann and W. Kabsch (1996). "*Structure of mitochondrial creatine kinase*." Nature **381**(6580): 341-5.

Gribskov, M., A. D. McLachlan and D. Eisenberg (1987). "*Profile analysis: detection of distantly related proteins*." Proc Natl Acad Sci U S A **84**(13): 4355-8.

Gunasekaran, K., B. Ma and R. Nussinov (2003). "*Triggering loops and enzyme function: identification of loops that trigger and modulate movements*." J Mol Biol **332**(1): 143-59.

Henikoff, S. and J. G. Henikoff (1992). "*Amino acid substitution matrices from protein blocks*." Proc Natl Acad Sci U S A **89**(22): 10915-9.

Hoedemaeker, F. J., R. R. van Eijsden, C. L. Diaz, B. S. de Pater and J. W. Kijne (1993). "*Destabilization of pea lectin by substitution of a single amino acid in a surface loop*." Plant Mol Biol **22**(6): 1039-1046.

Holm, L. and C. Sander (1997). "*Dali/FSSP classification of three-dimensional protein folds*." Nucleic Acids Res **25**(1): 231-4.

Hutchinson, E. G. and J. M. Thornton (1994). "*A revised set of potentials for beta-turn formation in proteins*." Protein Sci **3**(12): 2207-16.

Johnson, L. N., E. D. Lowe, M. E. Noble and D. J. Owen (1998). "*The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases*." FEBS Lett **430**(1-2): 1-11.

Johnson, L. N., M. E. Noble and D. J. Owen (1996). "*Active and inactive protein kinases: structural basis for regulation*." Cell **85**(2): 149-58.

Jones, T. A. and S. Thirup (1986). "*Using known substructures in protein model building and crystallography.*" EMBO J **5**: 819-822.

Joseph, D., G. A. Petsko and M. Karplus (1990). "*Anatomy of a conformational change: hinged "lid" motion of the triosephosphate isomerase loop*." Science **249**(4975): 1425-1428.

Kall, L., A. Krogh and E. L. Sonnhammer (2004). "*A combined transmembrane topology and signal Peptide prediction method*." J Mol Biol **338**(5): 1027-36.

Kawasaki, H. and R. H. Kretsinger (1995). "*Calcium-binding proteins 1: EF-hands*." Protein Profile **2**(4): 297-490.

Kendrew, J. C., G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff and D. C. Phillips (1958). "*A three-dimensional model of the myoglobin molecule obtained by x-ray analysis*." Nature **181**(4610): 662-6.

Kim, S. T., H. Shirai, N. Nakajima, J. Higo and H. Nakamura (1999). "*Enhanced conformational diversity search of CDR-H3 in antibodies: role of the first CDR-H3 residue*." Proteins **37**(4): 683-96.

Krishna, M. M. G., Y. Lin, J. N. Rumbley and S. Walter Englander (2003). "*Cooperative Omega Loops in Cytochrome c: Role in Folding and Function*." Journal of Molecular Biology **331**(1): 29-36.

Krogh, A., M. Brown, I. S. Mian, K. Sjolander and D. Haussler (1994). "*Hidden Markov models in computational biology. Applications to protein modeling*." J Mol Biol **235**(5): 1501-31.

Krogh, A., B. Larsson, G. von Heijne and E. L. Sonnhammer (2001). "*Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes*." J Mol Biol **305**(3): 567-80.

Kumar, S. and R. Nussinov (2001). "*How do thermophilic proteins deal with heat?*" Cell Mol Life Sci **58**: 1216-1233.

Lessel, U. and D. Schomburg (1999). "*Importance of anchor group positioning in protein loop prediction.*" Proteins **37**: 56-64.

Leszczynski, J. F. and G. D. Rose (1986). "*Loops in globular proteins: a novel category of secondary structure*." Science **234**(4778): 849-55.

Linhananta, A., H. Zhou and Y. Zhou (2002). "*The dual role of a loop with loop contact distance in folding and domain swapping*." Protein Sci **11**(7): 1695-1701.

Lo Conte, L., S. E. Brenner, T. J. Hubbard, C. Chothia and A. G. Murzin (2002). "*SCOP database in 2002: refinements accommodate structural genomics*." Nucleic Acids Res **30**(1): 264-7.

Lupas, A. N., C. P. Ponting and R. B. Russell (2001). "*On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?*" J Struct Biol **134**(2-3): 191-203.

Martin, A. C. and J. M. Thornton (1996). "*Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies*." J Mol Biol **263**(5): 800-15.

Matthews, B. W. (1972). "*The gamma turn. Evidence for a new folded conformation in proteins.*" Macromolecules **5**: 818-819.

Milner-White, E. J. (1987). "*Beta-bulges within loops as recurring features of protein structure*." Biochim Biophys Acta **911**(2): 261-5.

Milner-White, E. J. and R. Poet (1986). "*Four classes of beta-hairpins in proteins*." Biochem J **240**(1): 289-92.

Mosimann, S., R. Meleshko and M. N. G. James (1995). "*A critical assessment of comparative molecular modeling of tertiary structures of proteins*." Proteins: Structure, Function and Genetics **23**(3): 301-317.

Moult, J. and M. N. James (1986). "*An algorithm for determining the conformation of polypeptide segments in proteins by systematic search.*" Proteins **1**: 146-163.

Needleman, S. B. and C. D. Wunsch (1970). "*A general method applicable to the search for similarities in the amino acid sequence of two proteins*." J Mol Biol **48**(3): 443-53.

Oliva, B., P. A. Bates, E. Querol, F. X. Aviles and M. J. Sternberg (1997). "*An automated classification of the structure of protein loops*." J Mol Biol **266**(4): 814-30.

Oliva, B., P. A. Bates, E. Querol, F. X. Aviles and M. J. Sternberg (1998). "*Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction*." J Mol Biol **279**(5): 1193-210.

Pearson, W. R. and D. J. Lipman (1988). "*Improved tools for biological sequence comparison*." Proc Natl Acad Sci U S A **85**(8): 2444-8.

Pellequer, J. L. and S. W. Chen (1997). "*Does conformational free energy distingish loop conformations in proteins?*" Biophys J **73**(5): 2359-2375.

Rapp, C. S. and R. A. Friesner (1999). "*Prediction of loop geometries using a generalized born model of solvation effects*." Proteins **35**(2): 173-83.

Rice, P. A., A. Goldman and T. A. Steitz (1990). "*A helix-turn-strand structural motif common in alpha-beta proteins.*" PROTEINS: Structure, Function and Genetics **8**: 343-340.

Richardson, J. S. (1981). "*The anatomy and taxonomy of protein structure*." Adv Protein Chem **34**: 167-339.

Rose, G. D., L. M. Gierasch and J. A. Smith (1985). "*Turns in peptides and proteins*." Adv Protein Chem **37**: 1-109.

Salem, G. M., E. G. Hutchinson, C. A. Orengo and J. M. Thornton (1999). "*Correlation of observed fold frequency with the occurrence of local structural motifs*." J Mol Biol **287**(5): 969-81.

Sali, A. (1995). "*Modelling mutations and homologous proteins*." Current Opinion in Biotechnology **6**(4): 437-451.

Saraste, M., P. R. Sibbald and A. Wittinghofer (1990). "*The P-loop--a common motif in ATP- and GTP-binding proteins*." Trends Biochem Sci **15**(11): 430-4.

Schenk, P. W. and B. E. Snaar-Jagalska (1999). "*Signal perception and transduction: the role of protein kinases*." Biochim Biophys Acta **1449**(1): 1-24.

Sibanda, B. L., T. L. Blundell and J. M. Thornton (1989). "*Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering*." J Mol Biol **206**(4): 759-77.

Sinev, M. A., E. V. Sineva, V. Ittah and E. Haas (1996). "*Domain closure in adenylate kinase*." Biochemistry **35**(20): 6425-37.

Smith, K. C. and B. Honig (1994). "*Evaluation of the conformation free energies of loops in proteins*." Proteins **18**(2): 119-132.

Smith, T. F. and M. S. Waterman (1981). "*Identification of common molecular subsequences*." J Mol Biol **147**(1): 195-7.

Sonnhammer, E. L. and D. Kahn (1994). "*Modular arrangement of proteins as inferred from analysis of homology*." Protein Sci **3**(3): 482-92.

Tainer, J. A., M. M. Thayer and R. P. Cunningham (1995). "*DNA repair proteins*." Curr Opin Struct Biol **5**(1): 20-6.

Taylor, W. R. (1986). "*Identification of protein sequence homology by consensus template alignment*." J Mol Biol **188**(2): 233-58.

Thornton, J. M., B. L. Sibanda, M. S. Edwards and D. J. Barlow (1988). "*Analysis, design and modification of loop regions in proteins*." Bioessays **8**(2): 63-9.

Ulfer, R. and K. Kirschner (1992). "*The importance of surface loops for stabilizing an eightfold beta alpha protein*." Protein Sci **1**(1): 31-45.

van Vlijmen, H. W. and M. Karplus (1997). "*PDB-based protein loop prediction: parameters for selection and methods for optimization*." J Mol Biol **267**(4): 975-1001.

Venclovas, C., A. Zemla, K. Fidelis and J. Moult (2001). "*Comparison of Performance in Successive CASP Experiments*." PROTEINS: Structure, Function and Genetics **Suppl 5**: 163-170.

Venkatachalam, C. M. (1968). "*Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units*." Biopolymers **6**(10): 1425-36.

Wierenga, R. K., P. Terpstra and W. G. Hol (1986). "*Prediction of the occurrence of the ADP-binding beta alpha beta-fold in proteins, using an amino acid sequence fingerprint*." J Mol Biol **187**(1): 101-7.

Wilmot, C. M. and J. Thornton (1988). "*Analysis and prediction of the different types of beta-turns in proteins.*" J Mol Biol **203**: 221-232.

Wlodawer, A., M. Miller, M. Jaskolski, B. K. Sathyanarayana, E. Baldwin, I. T. Weber, L. M. Selk, L. Clawson, J. Schneider and a. Kent et (1989). "*Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease*." Science **245**(4918): 616-621.

Wojcik, J., J. P. Mornon and J. Chomilier (1999). "*New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification*." J. Mol. Biol. **255**: 235-253.

Wood, T. and W. Pearson (1999). "*Evolution of protein sequences and structures.*" J Mol Biol **291**: 977-995.

Wu, C. H., H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Z. Hu, R. S. Ledley, K. C. Lewis, H. W. Mewes, B. C. Orcutt, B. E. Suzek, A. Tsugita, C. R. Vinayaka, L. S. Yeh, J. Zhang and W. C. Barker (2002). "*The Protein Information Resource: an integrated public resource of functional annotation of proteins*." Nucleic Acids Res **30**(1): 35-7.

Xiang, Z., C. S. Soto and B. Honig (2002). "*Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction*." Proc Natl Acad Sci U S A **99**(11): 7432-7.

Yang, A. S. and L. Y. Wang (2003). "*Local structure prediction with local structure-based sequence profiles*." Bioinformatics **19**(10): 1267-74.

Zgiby, S., A. R. Plater, M. A. Bates, G. J. Thomson and A. Berry (2002). "*A functional role for a flexible loop containing Glu182 in the class II fructose-1,6-biphosphate aldolase from Escherichia coli*." J Mol Biol **315**(2): 131-140.

Zhou, H. X. (2004). "*Loops, Linkage, Rings, Catenanes, Cages, and Crowders: Entropy-Based strategies for stabilizing proteins*." Acc Chem Res **37**: 123-130.

Zomot, E. and B. I. Kanner (2003). "*The interaction of the gamma-aminobutyric acid transporter GAT-1 with the neurotransmitter is selectively impaired by sulfhydryl modification of a conformationally sensitive cysteine residue engineered into extracellular loop IV*." J Biol Chem **278**(44): 42950-8.