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 Experiments 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 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 conocercse. 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

PREFACE .................................................................................................................................................................................. ii

TABLE OF CONTENTS.......................................................................................................................................................... 4

LIST OF TABLES ...................................................................................................................................................................... 8

LIST OF FIGURES..................................................................................................................................................................... 9

SUMMARY............................................................................................................................................................................... 11

CHAPTER I - INTRODUCTION .............................................................................................................................................. 13

1.1 Loops - Historical background .......................................................................................................................................... 15

1.2 Importance of the loops in protein ..................................................................................................................................... 17

1.2.1 Loops and protein function ................................................................................................................................................ 17

1.2.2 Loops and protein structure ............................................................................................................................................ 19

1.3 Prediction of loop structure................................................................................................................................................ 21

1.3.1 Database search methods .................................................................................................................................. 22

1.3.2 Ab initio methods ................................................................................................................................................. 24

1.3.3 Combined methods ............................................................................................................................................. 25

1.4 Bioinformatic Tools............................................................................................................................................................ 26

1.4.1 Protein Structure Databases ............................................................................................. ........................................ 26

1.4.1.1 PDB (http://www.rcsb.org) ....................................................................................................................... 26

1.4.1.2 SCOP (http://scop.mrc-lmb.cam.ac.uk/scop) ............................................................................................ 26

1.4.1.3 Dali / FSSP (http://www.ebi.ac.uk/dali) ..................................................................................................... 27

1.4.2 Protein Sequence Databases ................................................................................................................................... 28

1.4.2.1 SWISS-PROT (http://www.ebi.ac.uk/swissprot/)....................................................................................... 28

1.4.2.2 PIR (http://pir.georgetown.edu)................................................................................................................ 28

1.4.3 Other Databases .................................................................................................................................................. 29

1.4.3.1 Organism Specific Databases .................................................................................................................. 29

1.4.3.2 Protein Domain and Family Database: Pfam and InterPro ....................................................................... 30

1.4.4 Algorithms for sequence analysis ............................................................................................................................. 31

1.4.4.1 Pair-wise sequence comparison methods ................................................................................................. 31

1.4.4.2 Profile sequence comparison methods ..................................................................................................... 32
# Table of Contents

1.4.4.3 Iterative procedures .................................................................................................................................... 34  
1.4.4.4 HMMs .......................................................................................................................................................... 34  
1.5 References ...................................................................................................................................................................... 36  

## Chapter II - Building a Functional Annotated Structural Loop Database of Kinases: ARCHKI... 45

2.1 Abstract ............................................................................................................................................................................ 46  
2.2 Introduction ...................................................................................................................................................................... 47  
2.3 Material and Methods ..................................................................................................................................................... 48  
2.3.1 Set of analyzed kinases ..................................................................................................................................... 48  
2.3.2 Clustering of motifs ............................................................................................................................................ 48  
2.3.3 Assignation of function to loop sub-classes ....................................................................................................... 50  
2.3.4 Web Server ....................................................................................................................................................... 52  
2.4 Results ............................................................................................................................................................................ 54  
2.4.1 General Overview ................................................................................................................................................... 54  
2.4.2 Web Database ......................................................................................................................................................... 56  
2.5 Discussion and Conclusions ........................................................................................................................................... 58  
2.6 References ...................................................................................................................................................................... 59  

## Chapter III - Mining ARCHKI: On the Common and Characteristic Motifs with Functional Information. ....................................................................................................................................................................... 62

3.1 Abstract ........................................................................................................................................................................... 63  
3.2 Introduction ..................................................................................................................................................................... 64  
3.3 Results ............................................................................................................................................................................ 66  
3.3.1 General Overview ..................................................................................................................................................... 66  
3.3.2 Data Collection of protein kinase folds ..................................................................................................................... 71  
3.3.2.1 Protein kinase like ..................................................................................................................................... 71  
3.3.3.2 P-loop containing nucleotide triphosphate hydrolases ............................................................................... 73  
3.3.3.3 Ribokinase like .......................................................................................................................................... 74  
3.3.3.4 Phosphoglycerate kinase .......................................................................................................................... 75  
3.3.3.5 Ribonuclease-H-motif ............................................................................................................................... 76  
3.3.3.6 β/α Tim barrel SCOP ................................................................................................................................ 76  
3.3.3.7 Glutamine synthase/guanido kinase catalytic domain .............................................................................. 77  
3.3.3.8 PRTase-like .............................................................................................................................................. 77
CHAPTER IV - KNOWLEDGE-BASED LOOP MODELING: APPLICATION OF ARCHDB IN LOOP MODELING .......... 103

4.1 Abstract ........................................................................................................................................................................ 104

4.2 Introduction .................................................................................................................................................................... 105

4.3 Material and Methods ....................................................................................................................................................... 109

4.3.1 Loop database .......................................................................................................................................................... 109

4.3.2 Sequence profiles calculation ................................................................................................................................ 109

4.3.3 Jack-knife test .......................................................................................................................................................... 110

4.3.4 Discrimination between ββ_hairpins and ββ_links ..................................................................................................... 112

4.4 Results ............................................................................................................................................................................. 114

4.4.1 Overview of results ................................................................................................................................................... 115

4.4.2 Class prediction ........................................................................................................................................................ 116

4.4.3 Sub-class prediction .................................................................................................................................................. 120

4.4.4 Discrimination between ββ_links and ββ_hairpins .................................................................................................... 122

4.4.5 Prediction of supersecondary structures .................................................................................................................. 123

4.4.6 Examples of predicted classes and sub-classes of loops ......................................................................................... 125

4.5 Discussion and Conclusions ......................................................................................................................................... 129
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5.1 Applications</td>
<td>129</td>
</tr>
<tr>
<td>4.5.2 Discrimination between $\beta$-hairpins and $\beta$-links</td>
<td>130</td>
</tr>
<tr>
<td>4.5.3 Comparison with other previously published works</td>
<td>131</td>
</tr>
<tr>
<td>4.6 References</td>
<td>132</td>
</tr>
<tr>
<td>CHAPTER V - CONCLUDING REMARKS AND FUTURE WORK</td>
<td>136</td>
</tr>
<tr>
<td>5.1 Concluding remarks</td>
<td>137</td>
</tr>
<tr>
<td>5.2 Future Work</td>
<td>138</td>
</tr>
<tr>
<td>APPENDIX A</td>
<td>140</td>
</tr>
<tr>
<td>A.1 ArchType Program: Algorithm details</td>
<td>140</td>
</tr>
<tr>
<td>A.1.1 Type of structural motifs</td>
<td>140</td>
</tr>
<tr>
<td>A.1.2 Definition of the loop conformation</td>
<td>141</td>
</tr>
<tr>
<td>A.1.3 Definition of the loop geometry</td>
<td>142</td>
</tr>
<tr>
<td>A.1.4 Clustering of loop and classification</td>
<td>143</td>
</tr>
<tr>
<td>APPENDIX B</td>
<td>144</td>
</tr>
<tr>
<td>B.1 Internet resources and software used</td>
<td>144</td>
</tr>
<tr>
<td>APPENDIX C</td>
<td>145</td>
</tr>
<tr>
<td>C.1 Manuscripts written during this work</td>
<td>145</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>146</td>
</tr>
</tbody>
</table>
List of Tables

Table 2.1. Summary of clustered loops, classes and sub-classes by motif type ................................................................. 54
Table 3.1. Summary of clustered loops, classes and sub-classes. .......................................................................................... 66
Table 3.2. Sub-classes of loops with functional information. .................................................................................................... 68
Table 3.3. Sub-classes with functional residues involved in specific functions and shared by different folds are indicated.... 70
Table 3.4 Test for prediction of sub-class assignment .................................................................................................................. 93
Table 4.1. Percentages of accuracy of class and sub-class prediction using the four types of sequences profiles, FP, P, FH and H ........................................................................................................................................ 116
Table 4.2 Percentages of $\beta$-strands or $\beta$-hairpins type on the first hit, the top five hits and top twenty hits ....................... 123
Table 4.3 Types of predicted supersecondary structures ........................................................................................................ 124
Table B.1: URLs for internet resources cited or used within this work. .................................................................................. 144
Table B.2: Software used within this work. ............................................................................................................................ 144
List of Figures

Figure 1.1 Type of loops according to connected secondary structures. .......................................................... 16

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. ................................................................. 18

Figure 1.3. Schematic showing the hypothetical evolutionary scenario that might have led to the evolution of thioredoxin and sbA 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). ................................................................................. 20

Figure 1.4. Protein Data Bank content growth (taken from http://www.rcsb.org 01-Apr-2004) .......................... 24

Figure 2.1. Flowchart for functional annotation of loop sub-class. Functional residues were identified by literature, “PDB sites” or closeness to ligands and were mapped on the profile of aligned loops. The degree of conservation of functional residues and the percentage of loops from the same super-family of SCOP (more than 50% of loops of the sub-class) was used to assign a likely function for the sub-class. ........................................................................... 50

Figure 2.2. Distribution of motifs in relation to loop length. The loop length is the number of residues between the ends of the secondary structures as defined by Kabsch & Sander(Kabsch and Sander 1983). ......................................................................................... 55

Figure 2.3. Averaged RMSD between loops in subclasses. The averaged RMSD of the sets of loop structures on each subclass was calculated with the main-chain atoms of the residues in the loop plus two bracing residues at each side. Additional extensions of the bars show the standard deviations of the averages. The number of sub-classes is shown at the top. ................................................................................................................................. 56

Figure 2.4. Screenshot of ArchKI information HTML pages. Classification browser with subclass information, multiple alignments of sequence and conformation, the profile pattern, and the image of structurally superimposed motifs as viewed with Rasmol(Sayle and Milner-White 1995). ............................................................... 57

Figure 3.1. Composition to show a general overview of sub-class αα5.1.1 and sub-class βα5.1.1. Side-chain residues involved on substrate specificity are indicated in red (for adenine kinases) and in yellow (for uridilate kinase). Residues specific for substrate binding are highlighted in red (plus #) in the multiple sequence alignment, while positions implicated on specificity are marked with an asterisk. Substrate molecules of AMP (extracted from the structure with PDB code 2ak3) and C5P (taken from the structure with PDB code 1qf9) were included on the superimposition of loops together with the side-chains of the loops involved in substrate binding. A detailed view is presented in the right top corner (A). All the Figures were generated with PREPI v.0.9 S.A. Islam & M.J.E. Sternberg (available upon request on: http://www.sbg.bio.ic.ac.uk/prepi/index.html). ............................................................................ 64

Figure 3.2. Coiled trace of the loops of sub-class βα16.1.1 and βα17.1.1. Side chains of conserved and functional residues are shown. The structure with PDB code 1bdg of a hexokinase in complex with glucose was used to superimpose the glucose molecule with the rest of loops from this sub-class. ........................................................................ 85

Figure 3.3. Coiled coil representation of the P-loop motifs of sub-class βα6.1.1. The structure of a protein in complex with ATP (PDB code 1dvr) was used to superimpose the ATP molecule with the set of loops of this subclass. The side-chain of Lys in position A1 is included to show the binding of ATP. ................................................................. 86

Figure 3.4. Coil representation of the main-chain of loops from sub-class βα11.2.2. Only residues with functional annotations are shown (Ser position -B8 and Arg position -B4 and -B6). The structure of a protein in complex with ADP (PDB code 1bg0) was taken to illustrate the pattern of interaction between the conserved residues and the ADP molecule. ........... 88

Figure 3.5. Trace plate of the sub-class β-pairpin2.2.2 showing the Gly-rich-loop motif found in kinases of protein kinases like fold. The molecule of ATP, taken from the structure with PDB code 1b38, was superimposed with the rest of loops of the sub-class to show the pattern of interactions between the adenosine ring and the hydrophobic residue (Ile or Leu) on the Nt β-strand and a Val in the Ct β-strand. .......................................................... 89
Figure 3.6. A) Coil trace of the sub-class ββlinks5.1.1 showing the DFG triplet motif. Side-chains of the Asp of the triplet and conserved residues located in –B1 are included in the picture. B) Coiled coil representation of the loops of sub-class ββlinks8.1.1 with the characteristic set of catalytic loops of the protein-kinase like fold. Side-chains for conserved residues are depicted (in red). The ATP of the structure with PDB code 1phk was superimposed with the rest of loops of the sub-class. C) Coiled plate with the combination of sub-classes ββlinks8.1.1 and ββlinks5.1.1 showing the main interactions with ATP. .................................................................................................................. 91

Figure 3.7. Representation of the loops of sub-class αβ10.1.1. Only side-chains with functional annotation are shown (Arg in position L6 and Trp in position L8). ........................................................................................................................................................................... 92

Figure 4.1. Flowchart of the methodology for prediction and analysis. ........................................................................................................................................................................................................................................................................................................ 113

Figure 4.2(a,b). (a) Number of classes vs. loop length in ArchDB, classes are shown according to different secondary structure types. (b) Number of subclasses vs. loop length in ArchDB. Number of subclasses is shown according to different secondary structure types. ........................................................................................................................................................................................................................................................................................................ 114

Figure 4.3. Distribution of predicted loops vs. length. Lengths of query loops are defined as the number of coil residues between flanket secondary structures predicted by PSI-PRED. ........................................................................................................................................................................................................................................................................................................ 115

Figure 4.4A to 4.4F. Loops class prediction. Variation of the accuracy of the prediction and ln(p_value) vs loop length and split according to the type (A for αα motifs, B for αβ motifs, C for βα motifs, D for ββ motifs, E for ββhairpins motifs and F for ββlinks motifs). Accuracy of prediction for first hit (solid square (●) and continuous line) and for ln(p_value) (solid square (●) and broken line). Accuracy of prediction for the top five hits (solid triangle (▲) and continuous line) and for ln(p_value) (solid triangle (▲) and broken line). Accuracy of prediction for the top twenty hits (solid diamond (♦) and continuous line) and for ln(p_value) (solid diamond (♦) and broken line). ββ motifs include both ββhairpins and ββlinks. The values of accuracy have been obtained with FH profiles. The values of accuracy have been obtained with FH profiles and the values of significance have been obtained comparing each motif prediction against 100 independent randomly selected loop subclass on the searching space (see material and methods). The red line marks the limit of significance (p = 0.05 ). ........................................................................................................................................................................................................................................................................................................ 118

Figure 4.5A to 4.5F. Loops subclass prediction. Variation of the accuracy of the prediction and ln(p_value) vs loop length and split according to the type (A for αα motifs, B for αβ motifs, C for βα motifs, D for ββ motifs, E for ββhairpins motifs and F for ββlinks motifs). Accuracy of prediction for first hit (solid square (●) and continuous line) and for ln(p_value) (solid square (●) and broken line). Accuracy of prediction for the top five hits (solid triangle (▲) and continuous line) and for ln(p_value) (solid triangle (▲) and broken line). Accuracy of prediction for the top twenty hits (solid diamond (♦) and continuous line) and for ln(p_value) (solid diamond (♦) and broken line). ββ motifs include both ββhairpins and ββlinks. The values of accuracy have been obtained with FH profiles. The values of accuracy have been obtained with FH profiles and the values of significance have been obtained comparing each motif prediction against 100 independent randomly selected loop subclass on the searching space (see material and methods). The red line marks the limit of significance (p = 0.05 ). ........................................................................................................................................................................................................................................................................................................ 120

Figure 4.6 Coiled coil representation of αβ4.1.1 subclass. Conserved Pro and conserved hydrogen bonds are represented in the picture. Residue positions within clusters are denoted by the secondary structure: B for β-sheets, A for α-helix, and L for coil, and counting backwards from the C-terminal residue of the first secondary structure and forwards from the N-terminal residue of the loop. Figures 4.6 to 4.8 were generated using PREPI v.0.9 (SA Islam and MJ Sternberg; available upon request at: http://www.sbb.bio.ic.ac.uk/prepi/index.html). ........................................................................................................................................................................................................................................................................................................ 126

Figure 4.7. Trace plate of the sub-class βα3.1.1 subclass. Conserved gly residues and conserved hydrogen bonds are shown. ........................................................................................................................................................................................................................................................................................................ 127

Figure 4.8. Coiled coil representation of the ββhairpin5.1.1 subclass. Conserved gly residues and conserved hydrogen bonds are described........................................................................................................................................................................................................................................................................................................ 128

Figure A.1. Ramachandran loop conformation, 9 x 9 matrix for the partition of the Ramachandran angles showing the most accessible regions indicated by a, b, p, b/p, g, I/g and e. ........................................................................................................................................................................................................................................................................................................ 141

Figure A.2. Loop-bracing geometry, definition for I plane and the representation of the vector M1 and M2 and the distance D. Geometry angles are described........................................................................................................................................................................................................................................................................................................ 142
Summary
Classification of Loops in Protein Structures: Applications on loop modeling and Protein Function

Narcís Fernandez-Fuentes

May, 2004

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. These two structures confirmed the existence of dominant conformations of $\alpha$-helices and $\beta$-sheets, just as proposed earlier by Linus Pauling, Robert Corey and Herman Brason (see review (Eisenberg 2003)). These regions have repetitive conformation of $\phi$ and $\psi$ 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 $\alpha$-helices or $\beta$-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 $\beta$-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 $\gamma$-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 $\alpha$-helices, 19% in $\beta$-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), glycine-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](image)

Figure 1.2. Some examples of functional loops are shown: (a) Cdr5 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 from 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.
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 Å 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 Complementarity 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).
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 proteins are clustered based on 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 an 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/](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) ([http://genome-www.stanford.edu/Saccharomyces/](http://genome-www.stanford.edu/Saccharomyces/)) is a similar resource concerning the complete genome of *Saccharomyces cerevisiae* and related yeast strains.

Finally, the Ensembl resource ([http://www.ensembl.org](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/](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/](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 calculate 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 Altschul 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 from 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


Chapter I. INTRODUCTION


