Chapter V - Concluding Remarks and Future Work.

5.1 Concluding remarks

In **Chapter II** we described a process to classify an important part of protein structures, the loops. The process of classification is fully automated and easily updatable. We built a web server, ArchKI, which provides a quick and easy access to the complete classification of kinases loops. Links to related databases and information about functional residues and/or regions in protein kinases were also given. All this classified information shapes a powerful tool with applications in different areas of biological sciences and bioinformatics.

In **Chapter III** we focused on the analysis of ArchKI. We automatically identified, clustered and classified functional motifs described in the literature (i.e. P-loop, gly-rich-loop among others). We demonstrated the importance of the conservation of loop structure related with protein function in protein kinases. We examined the usefulness of loop classification on the study of the catalytic mechanism of kinases by the identification of conserved positions of the loops along the structurally aligned sequences. We showed that the conservation of loop structure can be linked with the evolutionary history. In addition, we proved the application of ArchKI in loop structure prediction of kinase loops.

Finally, in **Chapter IV** we applied ArchDB to approach the problem of loop structure prediction. We described a method that relies only on sequence information and standard protocols in bioinformatics (i.e. HMM models and PSSM) for the prediction of loop structure. The values of accuracy and significance obtained with our procedures validate its employ in loop modeling with high reliability. The dramatic increase of classified loops provide an huge amount of information about loop conformations thus applicable on loop modeling. Also, we demonstrated

the prediction of supersecondary motifs by means of prediction of consecutive motifs whith a clear application on *fold recognition*. Furthermore, we proved the usefulness of ββ profiles in the discrimination of two geometrically different ββ arches: the $\beta\beta_{hairpins}$ and $\beta\beta_{links}$. It should be remembered that loops are the most mobile parts of protein structure, as shown by NMR solutions and X-ray temperature factors, and the most prone to error during structure refinement. All attempts at loop prediction must be viewed in this context.

5.2 Future Work

As a complement to this work the following perspective could be investigated:

(i) Interesting extensions to the work exposed in **Chapter II** include protein family specific loop classifications. Specific loop classifications could be generated for the families with many known structures, such as lipase, the aspartic and serine proteinases, globins and other well-represented groups in PDB. These family specific classes should be useful for the modeling of structures belonging to these populated protein families. This point was demonstrated in our studies of protein kinases. In addition, the study of the structural conserved motifs may help the description of important regions of the proteins.

(ii) From the work exposed in **Chapter III**, we showed that the conservation of loop structure is related to their function in protein kinases. Is it possible to infer function through the analysis of structural conserved loops?. When sequence or structure

comparisons fail to suggest a function, insights can come from discovery of functionally important local structural patterns. A subclass is a set of conserved local structural patterns. This leads us to a new question: There are subclasses related with function? To answer this question several analyses have to be made. For instance, we should further explore if the structural conservation of loops is correlated with functional descriptors such as Swiss Prot. Keywords, SCOP codes, EC codes or contacts with ligands among loops of the same structural sub-class.

(iii) Two main proposals arise from the work described in **Chapter IV**. First, the building of a web server for automated modeling of loops in protein structures, a valuable tool in protein structure prediction field. Second, the assembling of short fragments from known structures has been a widely used approach to construct protein structures. Recently, Kolodny&Levitt (2002) and Du et al.(2003) have employed short protein fragments to build protein structures. Our work offers a method for local structure prediction, a loop between two regular secondary structures. Thus, the combination of local structure predictions leads to the prediction of the whole protein structure. Nevertheless, the assembling of local fragment is not a trivial problem: definition of geometric and topological constraints, structure compactness or objective evaluation of the models are among of the main difficulties.

Appendices

Appendix A

A.1 ArchType Program: Algorithm details

Archtype is a computer program written in C and published in 1997 by Dr. Oliva(Oliva et al. 1997). ArchType is and automatic procedure for protein loop classification based on loop conformation and the geometry of the motif.

A.1.1 Type of structural motifs.

The protein loop considered by ArchType, are the irregular regions between α-helices and β-strands as defined with the program DSSP(Kabsch and Sander 1983). The chain segment formed from the loop and the two bracing regular secondary structures is referred to as a motifs. ArchType considers five types of motifs: $\alpha-\alpha$ between two α-helices; $\alpha-\beta$ between a α-helix and a β-strand: $\beta-\alpha$ between a β-strand and a α-helix; and $\beta-\beta$, between two β-strands. The $\beta-\beta$ class is further split into $\beta\beta_{hairpins}$, which are those loops between two b-strands with at least one hydrogen bond between both (outside the loop), and $\beta\beta_{links}$, which are the complementary set on $\beta-\beta$ (see figure 1.1 in Chapter I). The identification of antiparallel hydrogen bonding for $\beta\beta_{hairpins}$ is imposed by consideration of the relative geometry of the component secondary structures. Hydrogen bonds were taken from DSSP with an energy less than –0.5 kcal/mol. Loops involving a connection with a 3.10 helix or a polyproline helix are not considered.

A.1.2 Definition of the loop conformation.

The possible conformation of the residues within the loops were defined by assigning the most accessible regions in [ϕ,ψ] space. The regions are α, α_l, γ, β, β_p and ε (coded by ArchType as a, l, g, b,p and e, see figure A.1). Cis proline residues are also considered by assignation of a different region on the Ramachandran space denoted w. For a pair of loops, a conformational similarity score is obtained as the percentage of the total number of residues that can be equivalenced with identical conformational codes. Two special regions denoted l/g and b/p are considered as a transition regions between the l and g conformations and between the b and p conformations. Any residue in l/g conformation gives an optimal score when aligned with a residue in l, or g or l/g conformation. The region b/p is treated analogously with respect to b and p regions.

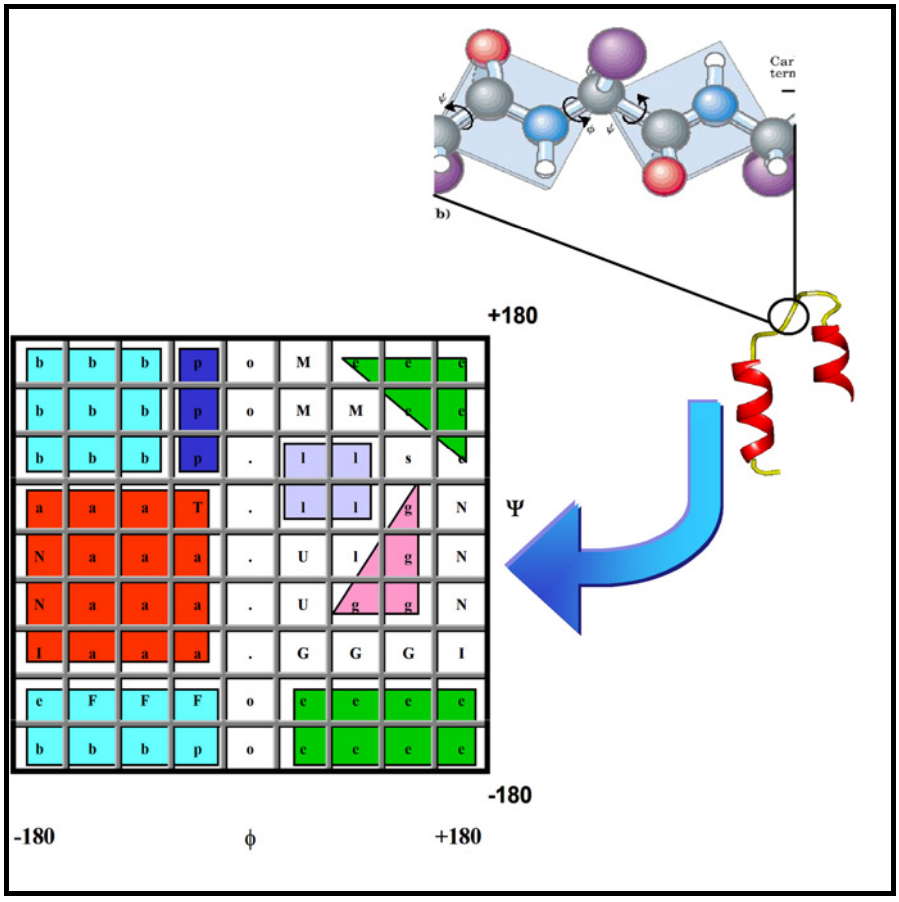

Figure A.1. Ramachandran loop conformation, 9 x 9 matrix for the partition of the Ramachandran angles showing the most accessible regions indicated by a, b, p, b/p, g, l/g and e.

A.1.3 Definition of the loop geometry.

An axis for a α or β secondary structure is defined based on the shortest of the principal moments of inertia of that structure. The geometry of each motif is defined by four internal co-ordinates (see figure A.2.). P1 and P2 are the start and end points of the loop and the absolute value of the vector that joins P1 and P2 is D (D=|L|). M1 and M2 are the axis vectors of Nt and Ct secondary structure respectively. Loop geometry is quantified as: (1) D, the distance between P1 and P2; (2) δ, the hoist angle, the angle between M1 and L; (3) θ, the packing angle, the angle between M1 and M2; and (4) ρ, the meridian angle, the angle between M2 and the plane that contains the vector M1.

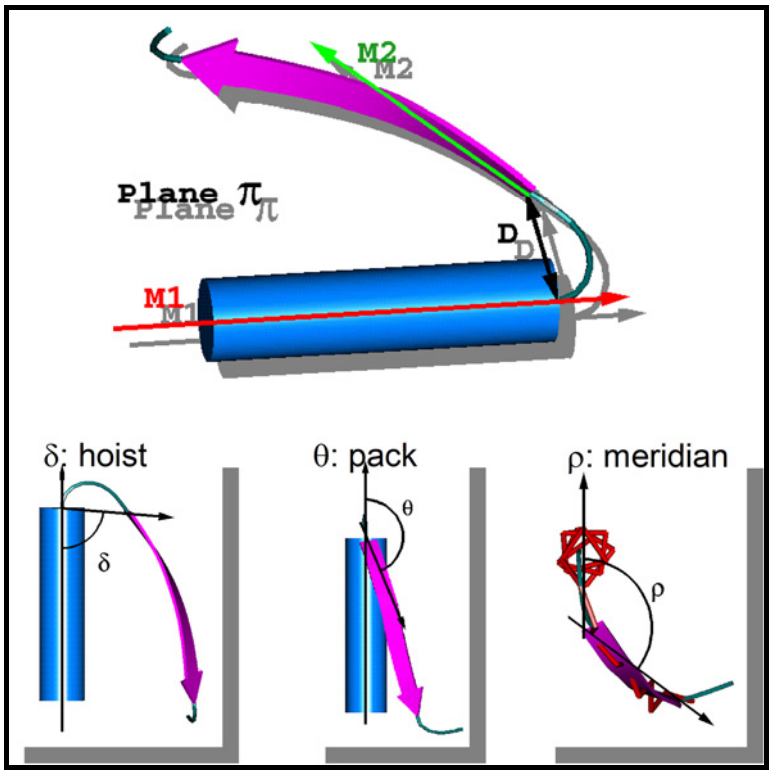

Figure A.2. Loop-bracing geometry, definition for Π plane and the representation of the vector M1 and M2 and the distance D. Geometry angles are described.

For D the total interval considered spans 0 to 40 Angstroms and is binned into 2 Angstroms interval, represented by integers form the mid-point of each interval (1,3,5,…,etc). For δ and θ the total interval spans form 0 to 180 degrees and for ρ from 0 to 360 degrees. These angles are binned into intervals of 45 degrees. Two motifs share the same geometry if $\Delta(D,\delta,\theta,\rho)$ belongs to the four-dimensional semi-open interval I=[(0,0,0,0),(2,45,45,45)].

A.1.4 Clustering of loop and classification.

Motifs of the same geometry are clustered based on the conformational similarity score (stage 1 clustering). The clustering method is based upon the density or mode-seeking technique (searching for regions containing a relatively dense concentration of loops), a version of single-linkage analysis(Everitt 1974). Because of the difficulty in defining the termini of the secondary structures, the loops compared have the same number of residues plus a potential extension of +/- 1 residue.

Clusters with a common Ramachandran pattern are grouped into classes (stage 2 clustering). Finally, each of these classes is split into subclasses, each containing at least three loops and differing only by their brace geometry. Subclasses are analyzed by the superimposition of their loops and the compactness of the cluster measured by the averaged RMSD of the superposition.

A consensus sequence is extracted from the multiply aligned clustered loops. If there is at least 75% agreement within the multiple alignment, a code is used to represent the chemical properties of the consensus at that position: (1) the one letter amino acid code; (2) p for polar residues [D,E,H,K,N,Q,R,S,T,Y]; and (3) h for non-polar residues [A,C,F,G,I,L,M,P,V,W,Y]. No sequence consensus is denoted by X. A consensus Ramachandran pattern was obtained analogously.

Appendix B

B.1 Internet resources and software used

Table B.1: URLs for internet resources cited or used within this work.

URL	Description
http://astral.stanford.edu	Protein sequences for SCOP domains
http://scop.mrc-lmb.cam.ac.uk/scop/	Structural Classification Of Proteins
http://www.biochem.ucl.ac.uk/bsm/cath	CATH structural classification of proteins
http://www.rcsb.org	Protein Data Bank
http://www.sbg.bio.ic.ac.uk/~3dpssm	3D-PSSM
http://www.cmbi.kun.nl/swift/pdbfinder/overview.html	PDB-FINDER
http://www.ebi.ac.uk/swissprot/	Swiss Prot Database
http://www.geneontology.org/	Gene Ontology Consortium
http://www.chem.qmw.ac.uk/iubmb/enzyme/	Enzyme Nomenclature
http://www.ebi.ac.uk/dali/	Dali Server
http://www.google.com	Google search tools

Table B.2: Software used within this work.

PROGRAM	Brief description
ArchType	Automatic method for protein loop classification
DSSP	Secondary structure assignment for protein structures
HMMER	Suite for sequence analysis using HMM profiles
Xam	RMS/RMSD calculation program
PSI-PRED	Secondary structure prediction method
AL2CO	Calculation of positional conservation in protein sequence alignment
RasMol	Molecular Visualization program
Prepi	Molecular graphics program

Appendix C

C.1 Manuscripts written during this work

Espadaler J., **Fernandez-Fuentes N.**, Hermoso A., Querol E., Aviles FX., Sternberg MJE., Oliva B. ArchDB: Automated protein loop classification as a tool for Structural Genomic. *Nucleic Acids Research*, **32**:1:D185-D188 (2004).

Fernandez-Fuentes N., Hermoso A., Espadaler J., Querol E., Aviles F.X., and Oliva B. Classification of common functional loops of super-families of kinases. *Proteins: Structure, Function, and Bioinformatics. In press*.

Fernandez-Fuentes N., Oliva B., Querol E., Aviles FX. and Sternberg MJE. Prediction of the conformation and geometry of loops in globular proteins: Implications for fold recognition and *ab initio* modeling. Submitted to *Journal of Molecular Biology*.

Fernandez-Fuentes N., Espadaler J., Hermoso A., Querol E., Aviles FX., Sternberg MJE., Oliva B. ArchDB and ArchKI: Expanding the Database of Classified Structural Motifs. Submitted to *Nucleic Acids Research*.