Exploiting Protein Fragments in Protein Modelling and Function Prediction

Jaume Bonet Martínez

DOCTORAL THESIS UPF / 2014

THESIS DIRECTOR:

Dr. Baldomero Oliva Miguel

Structural Bioinformatics Lab (SBI)

Research Program on Biomedical Informatics (GRIB)

Department of Experimental and Health Sciences (CEXS)

upf. Universitat
Pompeu Fabra
Barcelona

A mi madre.

Acknowledgements

A mi familia. A mis tíos, mis primos y primas (biológicos y políticos) y a sus respectivas progenies. Y, sobre todo, a mis padres Nila y Santi. Por su cariño, su esfuerzo y su sacrificio. No estaría aquí si no fuera por ellos.

A la nissaga Rimbau, per sempre fer-nos sentir com a part de la família.

A Víctor, por ser el único español que trabaja y, por tanto, estar manteniendo él sólo a todo el país. I al Carles, per sempre trobar moments per a compartir les històries. A tots dos, per molts anys d'amistat compartida.

Al Carlos i al Rubèn, per les tardes de futbol, les festes karaoke, els festivals de Sitges, els *"kikus"* i el grandíssim viatge a Japó... lluny de Japó.

Als meus companys de biologia. Per la seva inavaluable companyia i suport. I per els llits, matalassos i *"plegatins"* quan la nit s'ha allargat més del compte. A l'Ana, l'Anna, la Dora, l'Esther M., l'Esther P., el Gerard, el Jesús, el Manolo, el Melqui, la Mireya, la Neus, la Laura, el Xavi i la Vero. Y a Patricia, por los miles de quilómetros recorridos.

Thanks to the stackoverflow community. Thanks to every guy who had the same doubt I had before me and asked, and to every guy who knew the answer and replied. You all made my work easier.

A la gent del PRBB, sense els quals hagués costat molt més mantenir l'aparença de seny i per ajudar-me a trencar amb la rutina. En especial a l'Amadís, l'Eneritz, l'Inma, la Núria i l'Steve.

A la gent del GRIB, en especial a la Carina per ajudar-me a manegar-me amb el *"papeleo"* cada vegada que perdia els papers, i als I.T.: a l'Alfons, la Judith i el Miguel, per cada una de les vegades que m'han arreglat la màquina, canviat els discos, recuperat les dades... Y al Dr. Jordi Villà i Freixa, que em va enganxar a això de la bioinformàtica.

This journey would not have been the same without the traveling companions (new and old, transient and permanent) at the SBI lab. Angeliki, Attila, Ale, Bernat, Billur, Daniel, Dani, David, Elin, Elisenda, Emre, Jascha, Javi, Joan, Manu, Narcís, Ramón, Olga and Oriol. It has been so much fun sharing these years with all of you.

Per acabar, gràcies al meu director de tesis, el Dr. Baldo Oliva. Gràcies per trobar sempre temps per a nosaltres, per estar sempre obert a discussió i, en general, per haver creat un lloc on dona gust venir-hi a treballar. Em sento molt afortunat d'haver format part de l'SBI.

This work has been supported by the Spanish Ministry of Science and Innovation (MICINN) [FEDER BIO2008-0205]

Abstract

Proteins are the basic functional and structural blocks of life. Knowledge of their structure and function is key to understand how it works. Lately, the gap between new discovered proteins and systematically studied ones has increased due to the advance of sequencing techniques. As experimental methods are not able to keep up with the growing data, computational approximations to fill that gap are required. This thesis presents the study of protein loops (aperiodic regions in their three dimensional conformation) and their computational applications. It describes the latest developments regarding their classification (ArchDB), their use in protein structure modelling (Frag'r'Us) and in protein function prediction (Archer). It takes a special focus in protein redesign, which can be exploited to produce catalyst for non-biological processes or to design new synthetic biology pathways.

Resum

Les proteïnes son els blocs funcionals i estructurals de la vida. Conèixer la seva estructura i funció és vital per entendre-la. Últimament, la diferència entre noves proteïnes descrites i proteïnes estudiades sistemàticament ha augmentat explosivament degut a les noves tècniques de seqüenciació. Donat que el mètodes experimentals no poden fer-ho, és responsabilitat dels mètodes computacionals minimitzar aquesta diferència. Aquesta tesis presenta un estudi sobre els llaços de proteïnes (les regions aperiòdiques en la seva conformació tridimensional) i les seves aplicacions computacionals. Descriu els últims avanços referents a la seva classificació (ArchDB) i la seva utilitat en el modelatge de proteïnes (Frag'r'Us) i en la predicció de la seva funció (Archer). Fa un especial èmfasis en el seu redisseny, el qual pot ser explotat per catalitzar processos no-biològics o per el disseny sintètic de funcions biològiques.

Preface

I never expected to become a computational biologist when I started biology. In my defence, I didn't even know such a thing existed.

But things happened, decisions were made and, suddenly, I was doing a summer internship at Dr. Villà i Freixa's lab. My first impression was a little bit puzzling. I did like computers, and I was used to work around MS-DOS and even changing the position of jumpers in the motherboard (mainly because most of the games I bought for my 486 tended not to work otherwise) but I had no idea how UNIX worked or what was Perl. His answer was to give me a week to learn how to manage around them. It was awesome. I was studying biology, doing research (which was what I was sure I wanted to do)… and playing with computers. On the process, and amongst other questionable achievements, I managed to crash the CESCA, the first of many super-computers that have fallen against the might of my job-submitting (in)abilities.

Two years latter I started this journey with Dr. Oliva. At the SBI I have learned a lot. I have learned about the sea of data we are faced with, sprinkled with some known places, but mostly full of barely known regions, uncharted territories and some *"here there must be dragons"* locations. I have learn how gratifying is the feeling of achievement when things go according to plan, and how to manage (more or less) the frustration when it takes a bit too much time to see that they do not.

And, albeit I do like the idea of going *"big picture"* and try to guess why things work the way they do and try to apply that knowledge to find other things that will work the same way, I have relished in the opportunity to collaborate with multiple experimentalist that has been given to me. The possibility to study and focus in particular biological issues and deepen into different fields has been incredibly interesting and, luckily, quite productive. I guess that, despite the fact that the computational engineer is taking the wheel, the biologist is still riding shotgun.

All in all, I am happy with the work I have done and glad to be given the opportunity of reporting it in this thesis. I hope you kind it interesting.

Table of Contents

List of Figures

List of Tables

List of Publications

Publications in this section are listed in reverse chronological order.

Articles 2, 5 and 6 conform the main body of this thesis. Other articles are mentioned, when relevant, though the different sections.

Articles

(*) Indicates that the authors have contributed equally to a given article.

1. Resa-Infante, P., Paterson, D., **Bonet, J.**, Otte, A., Oliva, B., Fodor, E., & Gabriel, G. **Targeting importin-α7 as an antiviral approach against pandemic H1N1 influenza viruses.** (*Manuscript to be submitted*)
2. **Bonet, J.***, Garcia-Garcia, J.*, Planas-Iglesias, J., Fernandez-Fuentes, N., & Oliva, B. **Archer: Predicting protein function using local structural features. A helpful tool for protein redesign.** (*Manuscript to be submitted*)
3. Wright, R. H. G., LeDily, F., Soronellas, D., Pohl, A., **Bonet, J.**, Nacht, A. S., Vicent, G. P., Wierer, M., Oliva, B., & Beato, M. **ADP-ribose derived Nuclear ATP is Required for Chromatin Remodeling and Hormonal Gene Regulation.** *Nature*. (*Manuscript in revision*)
4. **Bonet, J.**, Fiser, A., Oliva, B., & Fernandez-Fuentes, N. **Smotifs as structural local descriptors of super-secondary elements: classification, completeness, and applications.** *BAMS*. (*In press*)
5. **Bonet, J.**, Segura, J., Planas-Iglesias, J., Oliva, B., & Fernandez-Fuentes, N. (2014). **Frag"r"Us: knowledge-based sampling of protein backbone conformations for de novo structure-based protein design.** *Bioinformatics (Oxford, England)*.
6. **Bonet, J.**, Planas-Iglesias, J., Garcia-Garcia, J., Marín-López, M. A., Fernandez-Fuentes, N., & Oliva, B. (2014). **ArchDB 2014: structural classification of loops in proteins.** *Nucleic Acids Research*, *42*(1), D315–9.
7. Fornes, O., Garcia-Garcia, J., **Bonet, J.**, & Oliva, B. (2014). **On the Use of Knowledge-Based Potentials for the Evaluation of Models of Protein-Protein, Protein-DNA, and Protein-RNA Interactions.** *Advances in Protein Chemistry and Structural Biology*, *94*, 77–120.
8. Rivera-Hernández, G., Marin-Argany, M., Blasco-Moreno, B., **Bonet, J.**, Oliva, B., & Villegas, S. (2013). **Elongation of the C-terminal domain of an anti-amyloid β single-chain variable fragment increases its thermodynamic stability and decreases its aggregation tendency.** *mAbs*, *5*(5), 678–689.

9. Planas-Iglesias, J.*, Marín-López, M. A.*, **Bonet, J.***, Garcia-Garcia, J., & Oliva, B. (2013). **iLoops: a protein-protein interaction prediction server based on structural features.** *Bioinformatics (Oxford, England).*

10. Planas-Iglesias, J., **Bonet, J.**, Garcia-Garcia, J., Marín-López, M. A., Feliu, E., & Oliva, B. (2013). **Understanding Protein-Protein Interactions Using Local Structural Features.** *Journal of Molecular Biology.*

11. Blanco-Toribio, A., Sainz-Pastor, N., Alvarez-Cienfuegos, A., Merino, N., Cuesta, A. M., Sánchez-Martín, D., **Bonet, J.**, Santos-Valles, P., Sanz, L., Oliva, B., Blanco, F. J., & Alvarez-Vallina, L. (2013). **Generation and characterization of monospecific and bispecific hexavalent trimerbodies.** *mAbs*, *5*(1), 70–79.

12. Wright, R. H. G., Castellano, G., **Bonet, J.**, Le Dily, F., Font-Mateu, J., Ballaré, C., et al. (2012). **CDK2-dependent activation of PARP-1 is required for hormonal gene regulation in breast cancer cells.** *Genes & Development*, *26*(17), 1972–1983.

13. Planas-Iglesias, J., **Bonet, J.**, Marín-López, M., Feliu, E., Gursoy, A., & Oliva, B. (2012). **Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence.** In *Protein-Protein Interactions - Computational and Experimental Tools*. InTech.

14. Garcia-Garcia, J., **Bonet, J.**, Guney, E., Fornes, O., Planas-Iglesias, J., & Oliva, B. (2012). **Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details.** *Molecular Informatics*, *31*(5), 342–362.

15. Cerdà-Costa, N.*, **Bonet, J.***, Fernández, M. R., Avilés, F. X., Oliva, B., & Villegas, S. (2011). **Prediction of a new class of RNA recognition motif.** *Journal of Molecular Modeling*, *17*(8), 1863–1875.

16. Cuesta, A. M., Sainz-Pastor, N., **Bonet, J.**, Oliva, B., & Alvarez-Vallina, L. (2010). **Multivalent antibodies: when design surpasses evolution.** *Trends in Biotechnology*, *28*(7), 355–362.

17. Cuesta, A. M., Sánchez-Martín, D., Sanz, L., **Bonet, J.**, Compte, M., Kremer, L., Blanco, F. J., Oliva, B., & Alvarez-Vallina, L. (2009). **In vivo tumor targeting and imaging with engineered trivalent antibody fragments containing collagen-derived sequences.** *PloS One*, *4*(4), e5381.

18. Aragues, R., Sali, A., **Bonet, J.**, Marti-Renom, M. A., & Oliva, B. (2007). **Characterization of protein hubs by inferring interacting motifs from protein interactions.** *PLoS Computational Biology*, *3*(9), 1761–1771.

19. **Bonet, J.**, Caltabiano, G., Khan, A. K., Johnston, M. A., Corbí, C., Gómez, A., Rovira, X., Teyra, J., & Villà-Freixa, J. (2006). **The role of residue stability in transient protein-protein interactions involved in enzymatic phosphate hydrolysis. A computational study.** *Proteins*, 63(1), 65–77.

Posters

1. Manuel A. Marin-Lopez, Joan Planas-Iglesias, **Jaume Bonet**, and Baldo Oliva. **Understanding Protein Recognition Using Structural Features.** *XIII Spanish Symposium on Bioinformatics (JBI2014)*, Sevilla, Spain, 2014

2. Joan Planas-Iglesias, **Jaume Bonet**, Javier Garcia-Garcia, Manuel A. Marin-López, Elisenda Feliu, Baldo Oliva. **Understanding protein-protein interactions using local structural features.** *12th CRG Symposium . BCN^2: Biological Control Networks in Barcelona*, Barcelona, Spain, 2013

3. Manuel A. Marin-Lopez, **Jaume Bonet**, Joan Planas-Iglesias and Baldo Oliva. **iLoops Server: A Protein-Protein Interaction Prediction Utility Based On Local Structural Features.** *XI Spanish Symposium on Bioinformatics (JBI2012)*, Barcelona, Spain, 2012

4. **Jaume Bonet**, Pascal Braun, Marc Vidal and Baldo Oliva. **Predicting Interacting Motifs from Protein-Protein Interaction Networks.** *17th Annual International Conference on Intelligent Systems for Molecular Biology and 8th European Conference on Computational Biology, ISMB/ECCB09*, Stockholm, Sweden, 2009

5. David Alarcon, Aggeliki Kosmopoulou, **Jaume Bonet**, Oriol Fornes, Roberto Mosca, José Manuel Mas, Patrick Aloy and Baldo Oliva. **Exploring type-2 Diabetes Protein Interactions Network by modeling their complex structures.** *17th Annual International Conference on Intelligent Systems for Molecular Biology and 8th European Conference on Computational Biology, ISMB/ECCB09*, Stockholm, Sweden, 2009

6. Oriol Fornes, Ramon Aragues, Jordi Espadaler, **Jaume Bonet**, Marc A. Marti-Renom, Andrej Sali and Baldo Oliva. **ModLink+: improving fold recognition by using protein-protein interactions.** *3DSIG 2009: The 5th Structural Bioinformatics and Computational Biophysics, ISMB satellite meeting*, Stockholm, Sweden, 2009

7. David Alarcon, Aggeliki Kosmopoulou, **Jaume Bonet**, Alejandro Panjkovich, Jose Manuel Mas, Patrick Aloy and Baldo Oliva. **Assessing Alzheimer's Protein-Protein Interaction Network.** *7th European Conference on Computational Biology, ECCB08*, Cagliri, Sardinia, Italy, 2008-09-22/26

8. **Jaume Bonet** and Baldomero Oliva. **On the role of non-regular secondary structure regions in protein-protein interactions.** *7th Spanish Symposium on Bioinformatics and Computational Biology*, Zaragoza, Spain, 2006-11-20/22

Summary

Chapter 1: Introduction**.** The thesis starts with an overview of the importance of proteins in the field of biological research and highlights the necessity to progress in the understanding of their structure and functions. It mentions the different techniques to approach both structure and function definition and prediction, covering both experimental and computational methods. Finally, it provides a brief introduction to the study of protein loops. Specific introductions regarding clustering, structure prediction and functional annotation protein loops are included in their respective chapters.

Chapter 2: Objectives**.** It lists the expected contributions of this thesis and briefly describes which ones are tackled in the different sections.

Chapter 3: Classifying Structural Fragments**.** This chapter includes a brief review of some other approaches to protein loops classification and a historical overview of ArchDB. As its main content, it includes the new version of the database. Finally, it focuses in the contributions of this classification against other fragment clustering approximations, highlighting its significance over previous ones. It contains the following publication.

Bonet, J., Planas-Iglesias, J., Garcia-Garcia, J.,
Marín-López, M. A., Fernandez-Fuentes, N., & Oliva, B. (2014).
ArchDB 2014: structural classification of loops in proteins.
Nucleic Acids Research, 42(1), D315–9.
http://dx.doi.org/10.1093/nar/gkt1189

Chapter 4: Modelling with Structural Fragments**.** This chapter revisits some of the methods used to model protein loops based on structural fragments. Then, it explains Frag'r'Us, a web application that not only can bridge undetermined structural regions in protein structures but also is specifically focused in the proposal of alternative backbone conformations for loop regions while maintaining the scaffold of the bracing secondary structures. This chapter contains the following publication.

Bonet, J., Segura, J., Planas-Iglesias, J., Oliva, B.,
& Fernandez-Fuentes, N. (2014).
Frag"r"Us: knowledge-based sampling of protein backbone conformations for de novo structure-based protein design.
Bioinformatics (Oxford, England).
http://dx.doi.org/10.1093/bioinformatics/btu129

Chapter 5: Predicting Protein Function with Structural Fragments**.** This chapter mentions some correlation studies between protein fragments and function prediction. Finally it centres in Archer, an application for function transfer annotation capable of highlighting the regions more likely to be related to the predicted function and propose mutations for protein design. It contains the article describing the annotation transfer process.

Bonet, J., Garcia-Garcia, J., Planas-Iglesias, J., Fernandez-Fuentes, N., & Oliva, B.
Archer: Predicting protein function using local structural features. A helpful tool for protein redesign.
(Manuscript to be Submitted)

Chapter 6: Discussion**.** This section represents a final summary of the work presented in this thesis. It extends upon the contributions of this work, some observations about the intricacies of computational research to the field of basic biological research and lists some of the possible works that can continue the path initiated in this thesis.

Chapter 7: Conclusions**.** Lists the final conclusions of this work.

Finally, the Appendix lists all the articles published during the thesis period that do not directly belong to its content.

1. INTRODUCTION

Proteins are the main gears and chassis of biological machines. They are involved in virtually every process carried out by the cell, from catalysis to photosynthesis, including signal transduction, molecular recognition and transport, storage, and structural support (Appendix 8.8)[1]. Their presence is so ubiquitous that they represent more than 50% of the dry weight of the cell [2].

As a testimony to their importance, the name protein derives from the Greek word *"proteios"*, which means "in the lead", and was coined in 1938 by J. J. Berzelius and G. J. Mulder [3].

The central dogma of molecular biology states that the inheritable information encoded in the deoxyribonucleic acid (DNA) is transcribed to ribonucleic acid (RNA) and this is, in its turn, translated to protein [4]. Thus, this principle states that, despite the fact that the information transferred from generation to generation is encoded in the DNA, proteins are the ones who are expected to perform the majority of the organism's functions. Amongst much other implications, this process highlights the fact that, by being able to read the genome (the complete genetic material) of an organism, we should be able to identify its proteins through the application of the universal genetic code. As of today, this simple principle has revolutionized the field of biological research [5].

We now know that there are tweaks and exceptions to this central dogma [6]. From interfering RNAs to alternative splicing or gene silencing, there are multiple factors that hamper our ability to translate a genome into its proteome (the entire set of proteins of an organism). And this is without even mentioning the difficulties to identify a gene's start codon and its reading frame. Furthermore, the genetic code has proven not to be as universal as it was once thought [7].

Despite all these drawbacks, the genomic projects have moved along [8] and have provided us with the drafts of the proteomes of multiple model organisms [9]. Of course, the direct implication of that is the generation of huge amounts of new data; especially of new proteins for which we still have a very limited amount of information. Databases such as Uniprot [10] are trying to manage with all that new and incomplete data [Figure 1.1].

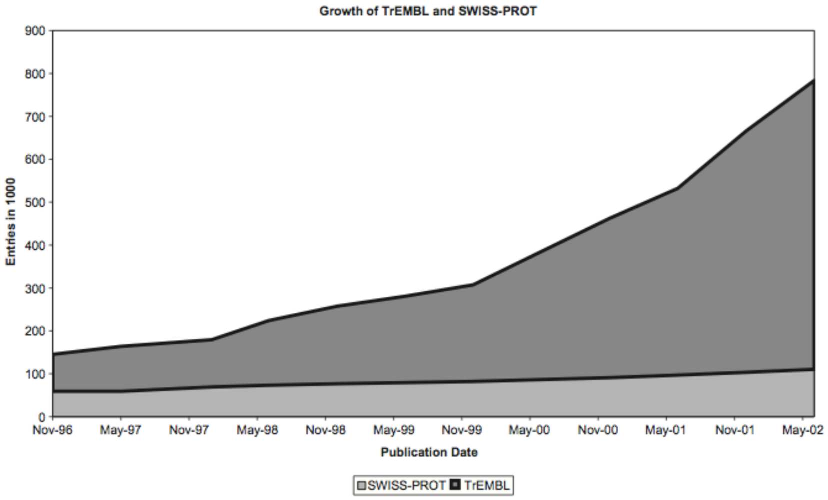


Figure 1.1 Growth of Uniprot. The figure represents the growth (in thousands) of protein sequences of the two divisions of Uniprot: Swiss-Prot, that contains the well described and manually curated proteins, and TrEMBL, which gathers all the rest, from unknown to theoretical sequences. Image extracted from [11].

Obviously, the amino acid sequence itself is the most fundamental piece of information about the protein. Before any kind of study or annotation, it is imperative to assess the reliability of any newly described protein. Multiple problems can arise from the high-throughput identification of proteins: from erroneous or missing protein segments to the description of non-existent proteins due to the identification of mRNA that do not actually translate to protein. Fortunately, there are several databases that curate all the new data making it available to the scientific community [Table 1.1]. Unfortunately, due to the specificities of each database, a new need has arisen to develop technologies able to reliably cross-reference the information of the different databases [12].

Name and Description	URL
Uniprot/Swiss-Prot [10] General protein sequence annotation.	`uniprot.org`
RefSeq [13] Reviewed and not reviewed protein sequences.	`ncbi.nlm.nih.gov/refseq`
CCDS [14] Compiles a *core* set of reliable human and mouse protein sequences.	`ncbi.nlm.nih.gov/CCDS`
HPRD [15] The Human Protein Reference Databank.	`hprd.org`
Mouse Genome Informatics [16] Curated information in mouse genes and phenotypes.	`informatics.jax.org`
Flybase [17] Curated data in *Drosophila* genes.	`flybase.bio.indiana.edu`
Wormbase [18] Curated data in nematode genes.	`wormbase.org`
Ensembl [19] Complete eukaryotic genomes.	`ensembl.org`
Disprot [20] Database of experimentally identified disordered regions.	`disprot.org`

Table 1.1 Protein Sequence Databases.

But the mere description of a protein's sequence is not enough to understand its function. On the one hand, their three-dimensional (3D) conformation, the shape that they take once synthesized, is key for the protein's function. On the other hand, proteins rarely act alone. They mostly carry out their purpose through their relation with other molecules [21].

Thus, we require (a) the 3D conformation of all proteins to comprehend their molecular function and (b) to know how they relate to each other and to other biological components in order to understand their participation in biological processes. Both elements are essential if we expect to fully understand how a cell or an organism works.

1.1. Protein Structure

Proteins are linear polymers that can range from tens to several thousand residues [22]. Specifically, they are polypeptides, that is, chains of amino acids linked by peptide bounds. The spatial configuration that those residues acquire in order to obtain a stable conformation between themselves and with their surroundings defines four levels of protein structure [Figure 1.2].

The **primary structure** refers to the sequence of residues that form the protein. There are 20 amino acids described as standard proteinogenic residues that are directly encoded in the genome. Although they bind to one another through the main chain, forming the backbone of the protein, it is the combined effect of their side chains that determine the final structure of the protein [23].

The **secondary structure** of a protein is defined by highly regular local sub-structures. Those regions determine specific geometries through the hydrogen bonds between the main-chain peptide groups. The two main secondary structures are helices and β-strands [24]. There are different types of helices, depending on the relative torsion between each pair of residues, being the α-helix the most commonly found. The flexible, non-regular regions between two secondary structures are known as loops [25]. Several consecutive secondary structures conform what is known as a **supersecondary structure** or structural motif.

The **tertiary structure** is defined as the native configuration of a single protein molecule in the three-dimensional space. It defines the spatial relationship between all the secondary structures of a protein creating a globular structure. Contrary to the secondary structure, the tertiary structure is stabilized by non-local interactions, that is, interactions between residues situated far apart in the primary structure. The tertiary structure is usually described by **domains**: independently stable units of 3D structure [26]. Usually, a domain has a degree of functionality on its own and similar domains can appear in seemingly unrelated proteins [27]. The tertiary structure of a protein can be described by one or multiple domains. The study of domains has become so relevant that several databases have been created just to catalogue them. Amongst the most noteworthy are SCOP [28], PFAM [29] and CATH [30].

Finally, the **quaternary structure** is the conformation formed by multiple protein molecules. In this context, a quaternary structure is known as a protein complex, and each molecule is called a protein subunit. Each subunit is linked to the others by non-covalent protein-protein interactions (PPIs), which results in different levels of stability over time. The proximity obtained through these complexes improves the speed and efficiency of the myriad of biological processes in which different complexes intervene.

It is worth mentioning that some proteins do not reach a secondary, nor a tertiary structure. Those proteins are said to have a **random coil** structure, as the lateral chains of the amino acids are assumed to be oriented randomly and are know as intrinsically unstructured proteins (IUPs) [31]. IUPs can still achieve a quaternary structure by interacting with others; sometimes even acquiring a tertiary structure in the process [32]. Similarly, other proteins are unable to achieve a native conformation on their own and require of a special kind of proteins called chaperones to fold themselves [33]. Finally, the plasticity of protein structures needs to be considered. Structures are not static, but they remain in permanent motion. Mostly, their flexibility results in small displacements, but, upon contact with other molecules, it can result in more significant conformational changes [34].

Practically all the know structures are collected in the Protein Data Bank (PDB) [35], a repository created in the 1970s to stockpile all the 3D structures and unify their format. From it, multiple databases with different focus have been developed [Table 1.2].

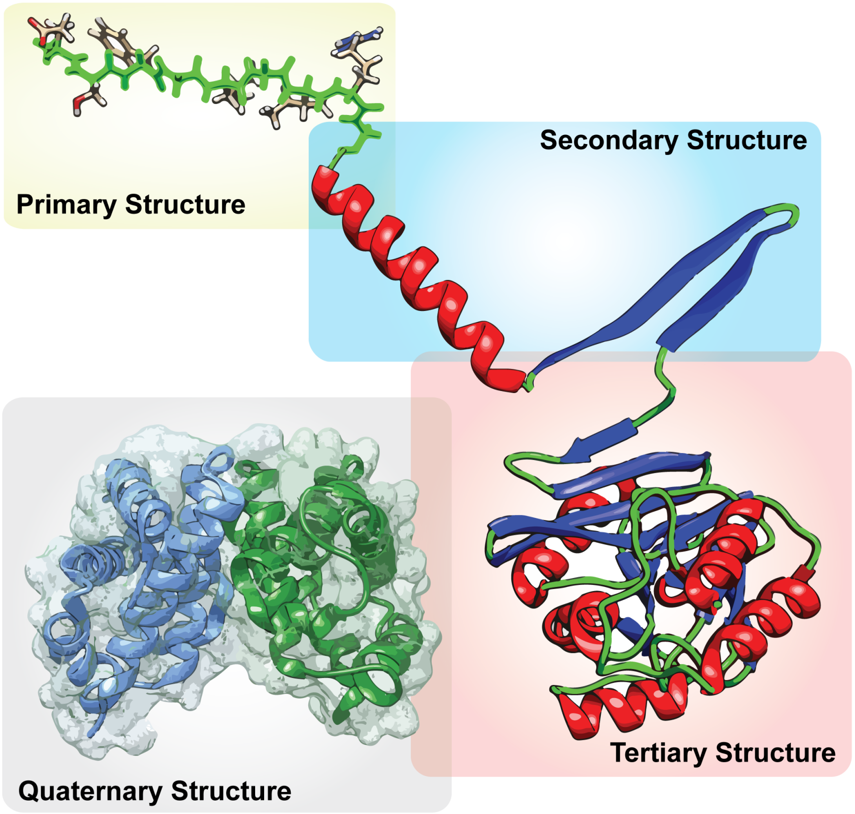


Figure 1.2 The different levels of protein structure.

Name and Description	URL
PDB [35] The main protein structure resource.	`pdb.org`
SCOP [28] The structural classification of proteins.	`scop.mrc-lmb.cam.ac.uk/scop`
PFAM [29] Protein families. Not all rely on structure.	`pfam.xfam.org`
CATH [30] Classification of protein structures.	`cathdb.info`
Dali [36] All against all structure comparison.	`ekhidna.biocenter.helsinki.fi/dali`
3DID [37] Structural domain interactions.	`3did.irbbarcelona.org`
SNAPPI-DB [38] Curated data in *Drosophila* genes.	`www.compbio.dundee.ac.uk/SNAPPI`

Table 1.2 Protein Structure Databases.

a) Protein Structure Determination

Experimental methods to determine the 3D structure of a protein are normally costly and time consuming. Despite multimillionaire international projects such as the Protein Structure Initiative [39], up to now, no reliable high-throughput methods have been developed.

According to the PDB, there are 11 main experimental techniques to determine the structure of a protein. Amongst them, X-ray diffraction [40] and Nuclear Magnetic Resonance spectroscopy (NMR) [41] cover 99% of the structures deposited in the PDB [Figure 1.3].

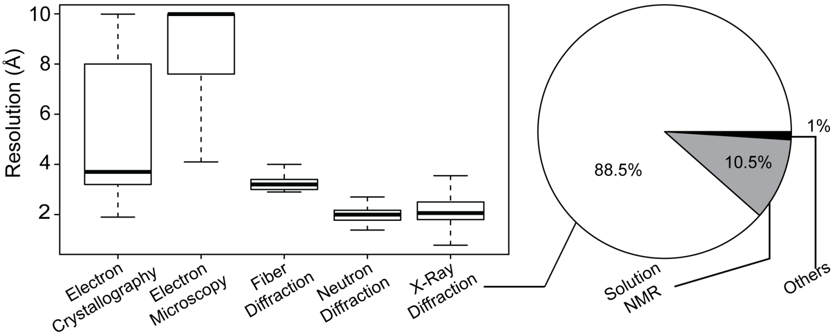


Figure 1.3 Structural determinant methods of the PDB. The image shows the percentage of crystals in the PDB according to the method used to determine them and the resolution range between the different methods in which it can be measured.

X-Ray Crystallography

X-ray crystallography is based on the diffraction suffered by X-rays when they interact with the electrons of a molecule. By measuring the intensities and angles of diffraction, a picture of the electron density map of that molecule can be produced. By knowing the composition of the protein or macromolecule that is being studied, the 3D structure of it can be fitted, with different levels of resolution, in the electron density map.

As the signal of a single molecule is too weak, multiple molecules are required to be positioned in a lattice forming a highly regular pattern: a crystal. Creating a crystal usually requires high concentrations of the purified protein or macromolecule and the appropriate experimental conditions. Thus, the crystallization of the protein itself becomes one of the main limitations to scale the method, as different condition of solvent and protein concentration are required in each particular case [42]. Those limitations are especially noteworthy in the case of transmembrane proteins.

This requirement of regularity and repetition also limits the applicability of the technique. As stated before, proteins have an inherent flexible nature, even when they have acquired a globular conformation. Therefore, sometimes the diffraction patterns will not be suitable to identify atomic details of the protein. In some instances, that will allow to identify the backbone and the fold of the protein but not the position of the lateral chain of the residues. In top of that, highly flexible regions such as long loops in a globular structure will leave gaps or missing coordinates in the 3D structure of the protein [43]. For this same reason, X-ray crystallography is not suited for the identification of IUPs, which, due to their lack of a well-organized structure, will place themselves differently in each cell of the lattice.

Regardless of all those drawbacks, X-ray crystallography is still one of the most successful, and most used, methods to determine the 3D structure of a protein or protein complex, not being limited by the size of it.

NMR spectroscopy

As previously stated, crystallization implies certain difficulties and can be both time and resource consuming. NMR, on the contrary, is applied to molecules in solution. This, of course, makes this method specially suited to identify the structure of IUPs and long loops.

The technique is based on the excitation of certain atoms through electromagnetic energy and the capture of the emitted radiation when the atoms returns to their equilibrium state. The atoms most exploited are ^{1}H, ^{13}C and ^{15}N, as they possess a magnetic moment that can rise to different energy levels when excited by specific radio frequencies. While ^{1}H is easily present in a regular medium, the last two need to be added by growing the cells that express the proteins in an enriched medium in such isotopes.

Different radio frequencies provide different data about the sample. Two of the most frequently used types are COSY (correlation spectroscopy) and NOESY (Nuclear Overhauser Enhancement Spectroscopy). COSY allows for the detection of covalent links, that is, enables the identification of adjacent residues. NOESY reveals residues that are close in space regardless of their relative position in the protein sequence. Once those distances are measured, and by knowing the sequence of the protein, a 3D structure can be generated by solving a distance geometry problem [44].

The main drawback of the NMR is that its analysis is, generally, limited to proteins smaller than 45kDa.

Non-Atomistic Techniques

There are other methods that are not gathered in the main distribution of the PDB. Amongst them, the most remarkable examples are cryo-electron tomography [45] and small angle X-ray scattering (SAXS) [46]. The former can be used to visualize large macro complexes, whilst the later is an X-ray approach that does not require a crystal and can overcome the molecular mass limitation of NMR, at the cost of resolution.

b) Protein Structure Prediction

The rise of high-throughput sequencing methods has resulted in the identification of a tremendous amount of new proteins [47]. In contrast, as we have seen in the previous section, the identification of the structure of a protein is still a slow and mostly manual labour. Thus, in the last year the difference between the amount of known proteins and the number of proteins with know structure has been increasing; and it was already big to begin with. As a rule of thumb, it is considered than less than 2% of the known sequences have a representative structure in the PDB [48]. This difference is often referred to as the sequence-structure gap [49]. Furthermore, due mostly to the limitations of the experimental protocols, there is a high rate of homology between the structures of the PDB.

We can check this last affirmation by studying the number of protein chains in the PDB in relation to the number of clusters of protein homologs at 90% and 60% of homology by means of CD-HIT [50] [Figure 1.4]. It is quite remarkable that, as of today, less than 20% of the proteins in the PDB represent different sequences at a 90% of sequence homology, and how fast that percentage drops at lower homologies.

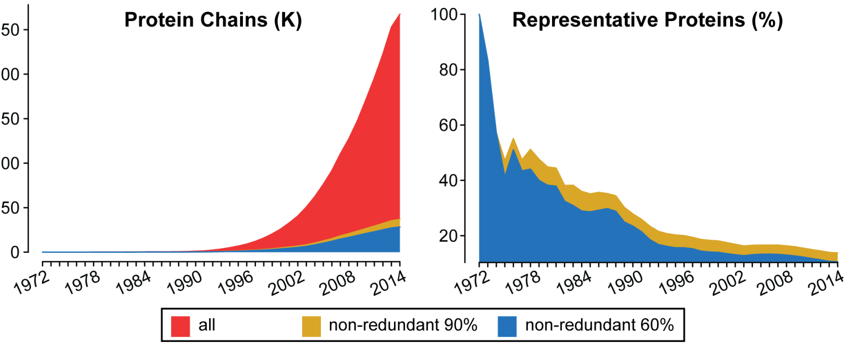


Figure 1.4 Evolution of the PDB protein content. The left image represents the raw count of protein chains in the PDB (in thousands) along with the number of clusters that can be obtained at 90% and 60% homology. The right image displays the amount of actual different sequences on the PDB depending on the homology threshold.

It is clear then, that with the methods available nowadays, the experimental determination of protein structures is not going to be able to gain on the number of known proteins any time sooner. Fortunately, theoretical methods can be used to bridge this gap by inferring the secondary and tertiary structure of a protein from its sequence [Table 1.3]. This methods are based in the paradigm that structure is more conserved than sequence [51]. Basically, this means that proteins with a similar enough sequence (homologs) will acquire a similar conformation. Regarding the prediction of the 3D conformation, three approaches deserve to be highlighted (Appendix 8.7) [52].

Name	Method
SECONDARY STRUCTURE PREDICTION	
Sable [53]	Neural network
Porter [54]	Neural network
PSIPRED [55]	Neural network
DISSpred [56]	Support vector machine
UNSTRUCTURED PROTEINS PREDICTION	
IUPred [57]	Pairwise energy
GlobPlot [58]	Russell/Linding scale of disorder
AB INITIO TERTIARY STRUCTURE PREDICTION	
I-TASSER [59]	*ab initio* folding
ROSETTA [60]	Fragment assembly
EVfold [61]	Correlated mutations
QUARK [62]	Monte Carlo fragment assembly
THREADING TERTIARY STRUCTURE PREDICTION	
GenTHREADER [63]	Sequence profile
Phyre [64]	Remote template detection
HHpred [65]	Remote template detection
MODELLING TERTIARY STRUCTURE PREDICTON	
MODELLER [66]	Satisfaction of spatial restraints
SWISS-MODEL [67]	Local similarity and fragment assembly
YASARA [68]	Detection of templates
Prime [69]	Physics-based energy function

Table 1.3 Protein Structure Prediction Software.

Ab Initio methods

De novo protein structure prediction tries to infer the tertiary structure of the protein directly from its sequence. This method is applied when no homologs with known tertiary structure (templates) are found for a given query protein. The idea behind it is that, by applying statistical tendencies gathered from known structures, it should be possible to obtain the 3D conformation of a protein. Those methods explore the known conformational space, generating multiple structural candidates (decoys). Thus, scoring functions, either knowledge based or physics based, are necessary to rank and identify native-like conformations. Afterwards, high-resolution refining can be applied to improve the final prediction of the tertiary structure. Despite the lack of information that motivates the use of these techniques, some programs, such as I-TASSER [59] or ROSETTA [60] have been able to reach a remarcable level of success.

Threading

Similarly to *ab initio*, threading is especially useful when no suitable template can be found. The method is based on the knowledge that there is a limited number of folds in nature and, thus, the query protein must fit into one of them (assuming it has tertiary structure). There are four main steps in the threading pipeline:

Template database construction: It requires the selection of non-homologous set of representatives from the PDB. By picking non-homologous sequences, the redundancy of the final database is removed, hence ensuring that single representatives of each fold are selected. If correctly performed, this step guarantees that the predictions over the query protein are not biased due to the over-representation of certain folds in the database.

Generation of the scoring function: Creating the optimal function to score the suitability of a given sequence when mapped over a certain fold is key. The final accuracy of the prediction will be directly related to the reliability of the scoring function. Typically, the function will integrate as much information as possible, including environment fitness potentials and secondary structure compatibilities amongst others.

Threading alignment: This is, by far, the most computationally costly part of all the process. At this point, the query sequence is aligned to each possible structure of the template database.

Threading prediction: Finally, by using the scoring function, each alignment is analysed and the best possible template is selected. Then, the model is built by arranging the atoms of the query protein around the backbone of the template.

Services such as GenTHREADER [63] or Phyre [64] allow the user to streamline al these steps.

Homology Modelling

This is, without a doubt, the most dependable method to predict the conformation in the space of a protein lacking a known 3D structure. In contrast to the methods previously described, homology modelling relies in the existence of a template or protein homolog to the query sequence with know 3D structure. Thus, by applying the structure conservation paradigm [51], we can transfer the structure from the template to the query.

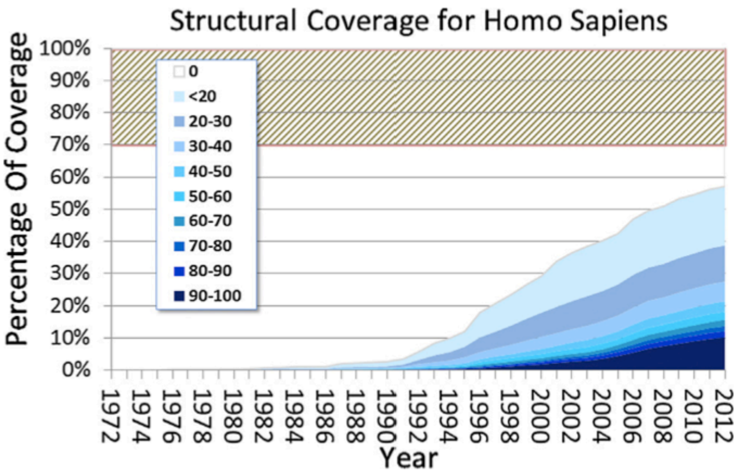


Figure 1.5 Coverage of modelled structures over the human proteome. Obtained from [70]. The colours represent the different levels of sequence-template homology.

The modelling protocol has provided a means to bridge the gap between sequence as structure; up to a certain point. Figure 1.5 shows the percentage of the human proteome whose structure can be predicted by comparative modelling. The red shaded section represents the 30% IUPs estimated to be in the human proteome [71], thus setting the theoretical limit of known protein structures at a 70% of all the proteome. By being very tolerant in the template selection, we are able to almost cover all the human proteome. But, as we will see later, this does not ensure reliable structure predictions.

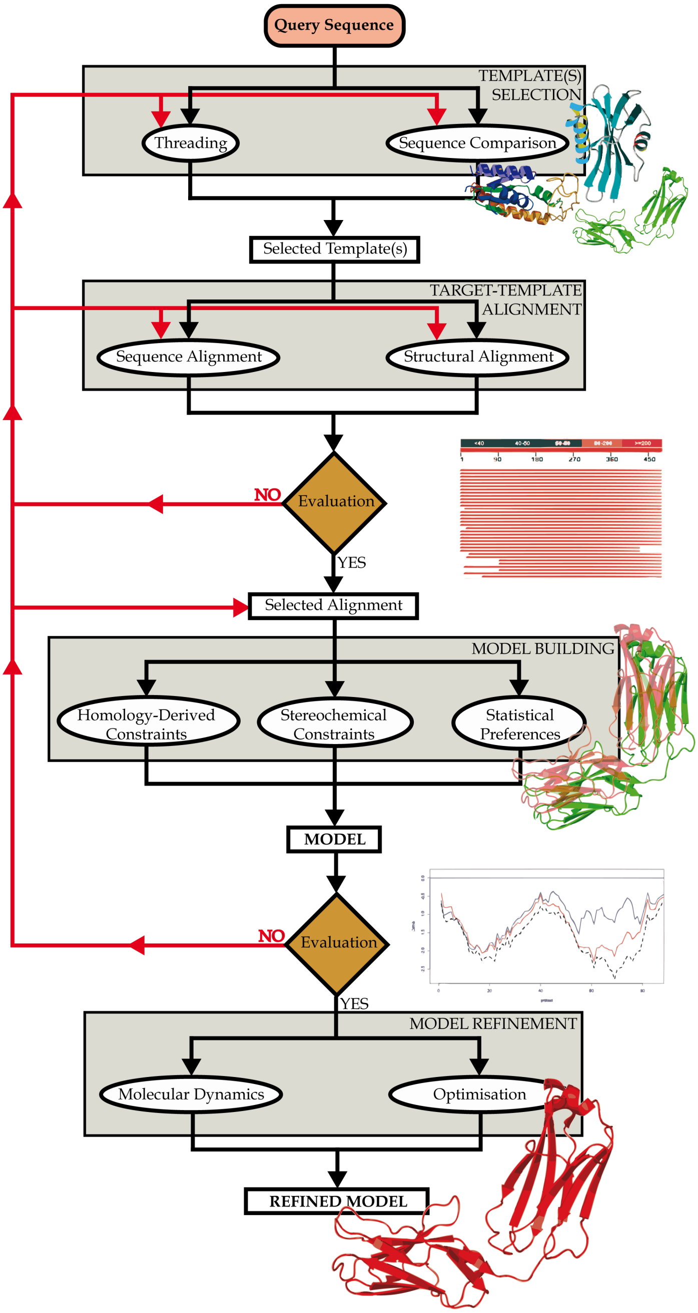


Figure 1.6 Flowchart for protein modelling. Schema of the methods used for modelling, comprising template(s) selection, template-target alignment, model building, model evaluation, and model refinement steps. Figure extracted from (Appendix 8.7) [52].

Regardless, reaching a 50% of sequence homology we can cover around a 20% of the human proteome. That is 10-fold improvement over the actual coverage. It is relevant to mention that, as a rule, prokaryotic proteomes present a higher structural coverage than eukaryotic ones [72].

The modelling pipeline is composed of four main steps [Figure 1.6]:

Template identification: This is the key step in homology modelling. The purpose of this step is to identify, amongst all the structures in the PDB, those whose sequence is closest (more similar) to the query protein. The search for homologs is, normally, performed through local sequence alignment tools such as BLAST or PSI-BLAST [73] or through Hidden Markov Models (HMM) domain profiles with HMMER [74]. As a rule of thumb, templates with a high homology will tend to produce better results than those with lower homology [70]. Specifically, if a template is found through local homology with a sequence identity around 40%, it falls in what it is known as the twilight zone. On this fuzzy region, there is a certain degree of uncertainty whether or not the template is similar enough to represent the correct fold of the protein of interest. Some empirical rules have been devised to overcome this decision [51], and some methods even try to filter evolutionary divergence between the query and the template y means of interologs (homologue pairs of interactors) [75].

In 1999, Rost empirically defined a method to assess the assignation of protein templates [51], with an special focus in solving the problem of low homology template assignation. This method represents an improvement over the HSSP-curve [76]. He aligned of 792 non-redundant (<25% sequence identity) proteins with known structure over the PDB. Each alignment was distributed across a two dimensional plane defined by the length of the aligned sequence and the number of identities [Figure 1.7 A, B] or positives [Figure 1.7 C, D]. By knowing the structure of the query proteins, he could distinguish good from bad template assignments, hence, defining the optimal curves to separate both sets [Eq 1.1 and Eq 1.2].

$$p^{i}(n) = n + 480 * L^{-0.32\left(1+e^{-L/1000}\right)}$$ **Eq 1.1**

$$p^{s}(n) = n + 420 * L^{-0.335\left(1+e^{-L/2000}\right)}$$ **Eq 1.2**

where **L** is the number of aligned residues, $\boldsymbol{p^{i}(n)}$ the probability threshold according to the number of identity positions and $\boldsymbol{p^{s}(n)}$ the probability threshold according to the number of positives. The parameter **n**, allows toggling the precision/record balance of the curves by translating them through the y axis [Figure 1.7].

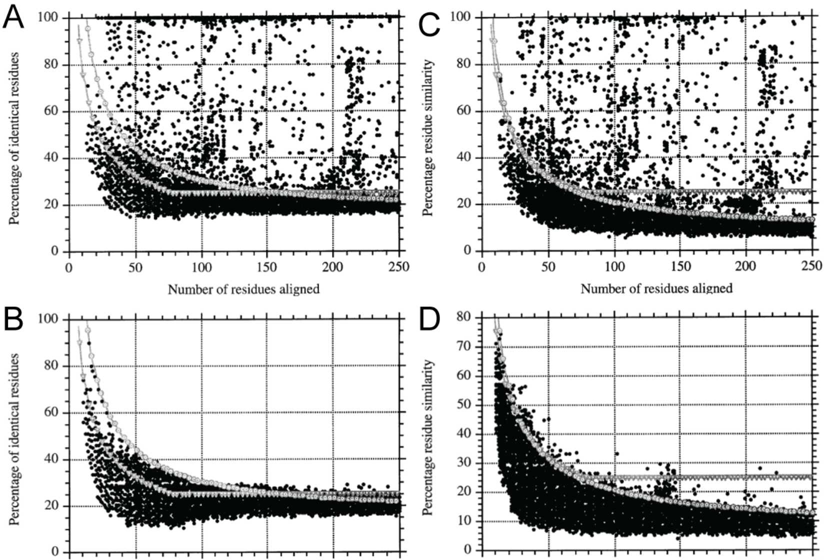


Figure 1.7 Twilight zone curves. Curves for homology detection through identity (A, B) and similarity (C, D). The curves separate true homologs (A, C) from false homology assignations (B, D). Scatter plots from [51].

As commented above, **L** represents the number of aligned residues, not the length of the sequence alignment. That is, gaps in the alignment have no more impact than shortening **L**. Similarly; the percentage of the query sequence covered by the alignment is not taken into account. These two factors imply that, while optimal to evaluate template assignment, the twilight zone curves cannot be used to assess global homology between two proteins.

Sequence-template alignment: When working with highly homologous templates, the alignment obtained through the template search should probably be good enough to proceed to the next step of the modelling protocol. But when that's not the case, or depending on certain requirements, alignments can be redone with applications such as CLUSTALW [77], T-COFFEE [78] or Matcher [79].

Model building: This is the step that actually creates the new 3D coordinates. At this point, the conformational information from the template is applied, guided by the sequence alignment, to the query protein. Among the multiple applications capable of to perform this process, MODELLER [66] is one of the most used. One of it advantages is that, provided a correct sequence alignment, the process is almost automatic. The created models should satisfy several spatial constraints; namely, (1) homology-derived constraints, (2) stereochemical constraints, and (3) statistical preferences for dihedral angles and non-bonded interatomic distances. The fact that it does not only depend on the homology-derived constraints allows the generation of multiple models for a given sequence-template alignment, each of the models fulfilling the other two constraints in different manner. This grants the possibility to analyse statistical fluctuations in the final predicted fold.

Evaluation: Assessing the global and local quality of a model is crucial to get a measure of its usefulness and to discern the confidence of the information that can be implied from it. Logically, modelling tools have their own energy evaluation methods, as is the case of MODELLER and the DOPE and GA341 energies [66]. But the use of independent evaluation tools is useful to ensure the quality of the model. Stereochemical restrictions such as clashes (occupation of the same space) and impossible or improbable orientations between consecutive amino acids (as defined in the Ramachandran diagram [80]) can be evaluated with PROCHECK [81]. Furthermore, knowledge-based or statistical potentials can be used to assess other possible problems of the model on a more atomistic scope. PROSA [82] is one of the most frequently used applications to perform that evaluation. There are, though, multiple criteria that can be used to generate different conformation statistical potentials [83].

Some methods have been developed to iterate between template alignment, model building and evaluation to alter iteratively alter the alignment and perform all the process in order to obtain the best scored and, thus, the most reliable model [84].

Refining: Once the optimal model is found, the model can be optimized to minimize its energetic landscape. Programs such as GROMACS [85] perform this task through molecular dynamic simulation.

1.2. Protein Function

The ultimate objective of the study of proteins is to understand their function and, as a result, how they work in a biological system. Genome-sequencing technology has not only widened the gap between proteins with known and unknown structure, but it has also reduced the percentage of encoded proteins with defined functional significance.

Reportedly, Swiss-Prot, the section of Uniprot dealing with manually annotated and reviewed proteins, represents around a 1% of the full sequence content of the database [10]. Furthermore, around 64% of the contents in Uniprot belong to proteins with inferred annotations. This set has a noticeable degree of overlap with Swiss-Prot. Lastly, a full third of the proteins in the database (35%) are directly categorized as “putative”, “hypothetical”, “with unknown function” or similar terms that provide little to no information or can be directly misleading [10]. But this numbers seem quite optimistic when looking at actual functional annotations in Swiss-Prot [Figure 1.8].

It is undeniable, then, that the development of high-throughput methods to bridge the gap between the continuously increasing number of sequenced proteins and the amount of those with known function is one of the main challenges of biological research [86].

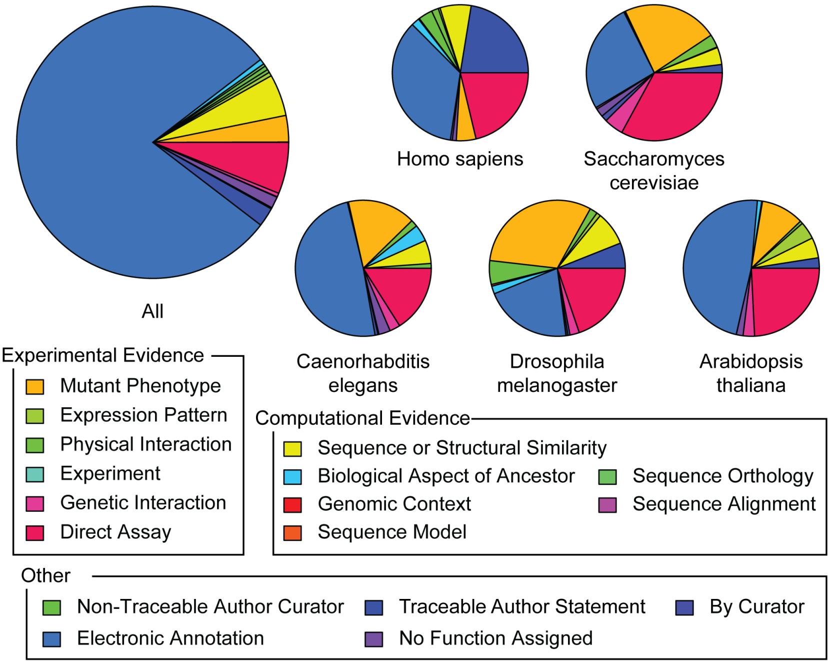


Figure 1.8 GO molecular functions mapped on Swiss-Prot proteins. Represented for the entire database and some selected model organisms. It can be seen that, with the notable exception of *Drosophila melanogaster*, most annotations are electronically assigned. The drawback of electronic annotations is that they must be revised yearly or they are deleted, which can compromise the development of predictive methods. As a rule, experimental evidences cover a small part of all known functions.

The experimental determination of the function of a protein relies on the gathering of vast amounts of information through experiments. This includes the identification of cofactors and post-transcriptional modifications, *in vitro* analysis of enzymatic activity, and even the evaluation of phenotypic effects in knockout models [87]. Despite being the more reliable procedure to identify a protein's function, experimental approaches are time consuming and expensive, making them difficult to scale.

In consequence, computational methods are required in order to reconcile the protein sequence-function gap. This, at the same time, yields some related problems whose solutions will determine the ability of the different bioinformatic strategies to predict the function of a protein. Amongst them, we can highlight: (1) the definition of the functional landscape, and (2) the development of similarity metrics.

a) Classifying Protein Functions

The functional landscape of the proteome, that is, the collection of all the possible functions that can be performed by any protein, needs to be defined if it is to be used for the prediction of protein functions. The definition, characterization and classification of functions can be considered, in itself, as a complete area of study. As any computational method devoted to protein function prediction is based on information transfer, the definition of the functional landscape will limit the type of functional annotations that can be transferred [88]. Similarly, the pattern used to define and categorize the different functions will define the scope to which the predictions will be able to deepen. It is for this very reason that multiple schemes for protein function classification have been developed through the years.

One of the most straightforward schemes divides the functional classes according to their biological involvement: energy, information, and communication and regulation [89]. These categories define quite general activities, and are limited to the biological process of the proteins. This means that proteins with identical molecular activity (phosphorylation, for example) could be classified in completely different clusters, being unable to assess the functional similarity between them.

Other classifications have been developed. Some are focused on systematize the labelling of specific sets of functions or species while others strive to create a unified dictionary for all known functions [Table 1.4]. The Enzyme (EC) database [90] and the Gene Ontology (GO) [91] are amongst the most used.

Name and Description	URL
EC [90] Hierarchical classification of enzymatic functions.	`enzyme.expasy.org`
GO [91] Ontological classification of functional terms.	`geneontology.org`
MEROPS [92] Classification of peptidases and their inhibitors.	`merops.sanger.ac.uk`
KEGG [93] Classification of high-level functions.	`www.genome.jp/kegg`
BRENDA [94] Literature based enzyme classification.	`brenda-enzymes.org`
PANTHER [95] Functionally related protein superfamilies.	`pantherdb.org`
EcoCyc [96] Literature based data on *Escherichia coli* K-12.	`ecocyc.org`

Table 1.4 Protein Function Databases.

EC database

The Enzyme (EC) database [90] is, at its name indicates, restricted to the annotation of enzymatic functions. Its first version dates from 1955, from a collaboration between the International Union of Biochemistry (IUB) and the International Union of Pure and Applied Chemistry (IUPAC).

EC tries to create a hierarchical classification of enzymes, giving each protein a four-field code. Each field corresponds to a number from 1 to n, and each consecutive field represents a higher degree of information with respect to the enzymatic function of the protein and the process by which it is performed.

Thus, the first level of the EC code defines the six main divisions of enzymes: (1) oxidoreductases, (2) transferases, (3) hydrolases, (4) lyases, (5) isomerases and (6) ligases. The information contained in the second and third fields depends on the first [Table 1.5]. The last field specifies on the concrete enzymatic reaction [97].

Field 1	Fields 2-3	
(1) Oxidoreductases	2	Substrate
	3	Acceptor
(2) Transferases	2	Class of item transferred
	3	Acceptor
(3) Hydrolases	2	Kind of bond cleaved
	3	Molecular context
(4) Lyases	2	Kind of bond formed
	3	Molecular context
(5) Isomerases	2	Class of reaction
	3	Specific class of reaction
(6) Ligases	2	Type of bond formed
	3	Type of molecule bonded

Table 1.5 EC levels information content.

From this classification method, it can be seen that EC does not classify enzymes, but enzymatic functions. Due to that, non-homologous isofunctional enzymes [98], that is, enzymes that perform the same function regardless of the fact that they belong to completely different protein folds, are classified together.

The strict hierarchical architecture of the EC classification makes it easy to compare the function of two proteins, thus being a straightforward similarity metric.

GO database

Created by the Gene Ontology Consortium, its main goal is the systematic classification of functions by means of a dictionary of terms and the relationships between them [91]. By creating this dictionary and rules of syntax, the expectation is for other researchers or groups to integrate this nomenclature in their own projects.

Although the syntax created by GO may yield more than one parent for a given concept (being an ontology and not a hierarchy), the first level of the ontology divides it into three different conceptual groups whose children do not relate between them; up to the point that they can be considered three different ontologies:

The **molecular function** relates to the function that the protein performs in itself. In other words, given the required components, it should be, theoretically, able to perform that same function *in vitro*.

The **biological process** refers to the activity of the protein in the living system. This is the function that the protein conducts in coalition with other biomolecules of the organism, and represents the cellular point of view.

The **cellular component** describes the location of the protein. It is important to track this, as many processes are dependent upon their cellular location.

Due to its ontological nature [99], it is difficult to compare the function of two proteins. If we take a look at Figure 1.9 we can see that the term "mitotic anaphase" can be considered as a level 4 term of "biological process" or as a level 9 term. The multiple inheritance present in the ontology makes it impossible to determine a specific level of functional definition that can be used as a measure of the functional similarity between different proteins. Some similarity measures between GO terms have been devised in order to cope with those problems [100,101].

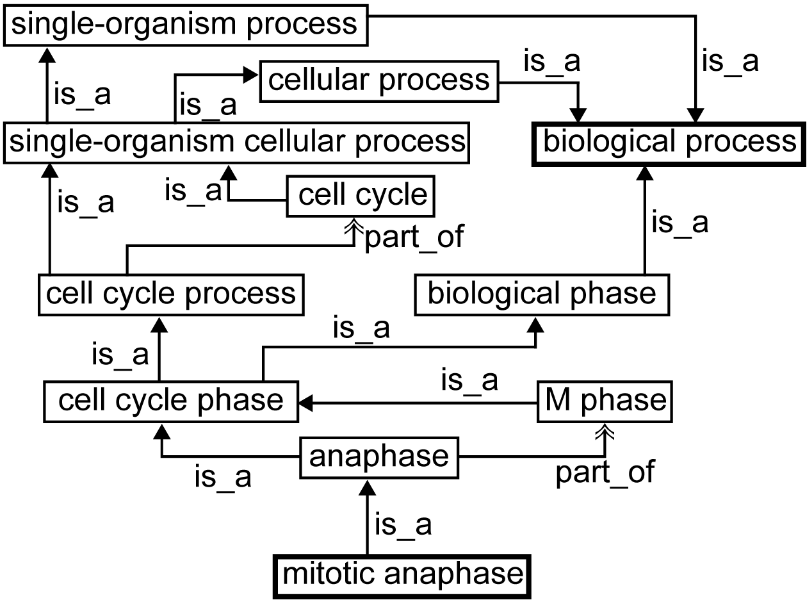


Figure 1.9 Ontology of the term GO:0000090; mitotic anaphase.

b) Protein Function Prediction

Similarly to protein structure prediction, the prediction of a protein's function depends on information transfer; that is, functional information is transferred from one protein to another relying on their similarity. The reasoning behind it being that evolutionary proximity implies a shared function [88]. Contrary to protein structure prediction, though, even when transferring enzymatic annotation between proteins with sequence identities up to 70%, around a 10% of those transfers are erroneous, being those differences quite common near 50% of identity [102].

This functional divergence is a key aspect that needs to be taken into account when performing functional annotation transfer. Two proteins related by descend from a common ancestor are homologs. Those two proteins will share a relatively high amount of sequence identity depending on their degree of homology. When these two proteins appear in different species, they are called orthologues. When they appear in the same species, they are called paralogues. Classically, it is considered that, as one protein can carry its given function, the paralog is, up to a certain degree, free of the evolutionary pressure, allowing the arising of new functionality. Therefore, functional annotation inferences based on sequence homology are more secure between orthologues than they are between paralogues [87]. The problem widens, as there is no fail proof method to assess if two homologous proteins from different organisms derive from a shared ancestral protein or from two different paralogues.

There are other factors to be taken into account regarding the complexity of functional annotation transfer. Opposite to divergence, we found functional convergence, that is, non-homologous proteins presenting a similar function [98]. Due to this convergence, transfers between non-homologous proteins cannot be disregarded, adding a new level of complexity to the process. Finally, single mutations can produce perturbations resulting in considerable fold changes (and, thus, functional variation) while variations on whole segments of sequences can have no effect whatsoever in the protein's function [102].

Despite all this limitations, or maybe because of them, multiple algorithms and methods have been devised to predict the function of a protein. Approaches based on global and local similarity try to predict the biochemical function of the protein (similar to what GO catalogues as molecular function), while genomic context and protein network methods try to elucidate the activity of the protein (in the lines of what GO defines as biological function).

Protein Similarity Methods

Annotation through global similarity is the most straightforward method for function annotation transfer. For example, the transfer of GO annotations through BLAST has been extensively used [Table 1.6]. As mentioned before, the major pitfall of the methods based on global similarity is that, under a 70% of sequence homology, their precision decreases greatly

[103]. It is worth mentioning that, due to the alignment algorithm of BLAST, the transfer does not have always to correspond to a global alignment but also to the alignment of a conserved section (like a domain). This means that sequence coverage in the alignment has to be taken into account when scoring the global homology between two proteins before the annotation transfer.

Local similarity tries to overcome the problems of functional divergence at high homologies by defining sequence patterns or motifs amongst proteins with shared function [104]. Those protein motifs can be represented as structural segments [105], HMM [65] or as sequential regular expressions [106]. As previously discussed, single residue changes can completely alter the function of a protein whilst changes on whole sections have no effect. The idea of the protein motifs is to represent those sections of the protein that are related to its function. By capturing the most conserved regions between proteins with shared function it is expected to build those motifs. In practice, not only active sites are found through this process, but also post-transcriptional modification sites [107], structural signals [108] and non-informative segments. The application of motifs in the prediction of protein functions usually works through the creation of pattern databases with associated functions and the assignation of those patterns to query proteins through pattern matching algorithms [Table 1.6]. Pfam [29] and PROSITE [106] are amongst the most used pattern databases.

Structure based methods for function annotation transfer work on the same conservation principle as sequence based methods do. But, instead of evaluating the sequence similarity, they are mostly based on the merit of their superimposition [36] or in the location of correlated 3D segments that could create and active site or a chemical pocket [109]. Structural methods, in general, are able to produce slightly better results when sequence homology drops below 40% [110]. There are several services, a part from DALI [36], that offer the possibility to predict a protein's function given its structure [Table 1.6].

Name and Description	URL
GLOBAL SEQUENCE HOMOLOGY	
HAMAP [112] Catalogue of genetic similarities and differences in human beings.	hapmap.ncbi.nlm.nih.gov
InParanoid [113] Orthologous groups with inparalogs.	inparanoid.sbc.su.se
GOEngine [114] Sequence homology and text information	geneontology.org
FRAGMENT BASED DATABASES	
Pfam [29]	pfam.xfam.org
PROSITE [106]	prosite.expasy.org
PRINTS [115] Groups of conserved motifs.	www.bioinf.manchester.ac.uk/dbbrowser/PRINTS
COG [116] Cluster of orthologous groups of proteins.	ncbi.nlm.nih.gov/COG
InterPro [117] Combine multiple sequence signature databases.	www.ebi.ac.uk/interpro
HSSP [118] Local multiple structure alignment.	swift.cmbi.ru.nl/gv/hssp
STRUCTURAL SIMILARITY	
DALI [36] All against all 3D comparison.	ekhidna.biocenter.helsinki.fi/dali
CE [119] Structural alignment.	source.rcsb.org/jfatcatserver
SSAP [120] Pairwise structure comparison.	cathdb.info/cgi-bin/SsapServer.pl
ASSOCIATION BASED METHODS	
STRING [121] Known and predicted protein interactions.	string-db.org
IntAct [122] Molecular interaction data from literature	www.ebi.ac.uk/intact
TRANSPATH [123] Mammalian signal transduction and metabolic pathways.	genexplain.com/transpath-1
KEGG [93] High-level biological functions.	www.genome.jp/kegg

Table 1.6 Protein Function Prediction Services and Databases.

Association Methods

Placing a protein in its biological context can be used to transfer functional annotation between them, especially when referring to predicting its biological function. **Gene fusion** (or Rosetta Stone) methods are based on the idea that gene fusions produced in a particular genome should indicate that the product of those genes in other genomes might cooperate in a particular function. A somehow similar idea is applied to **gene neighbourhood**, in which closely located and regulated genes are expected to perform a combined function. In a closely related manner, **protein-protein interactions** (PPI) can also be exploited, as biological functions in proteins are performed through their relations [111]. Those physical relations can be further extended through different organism creating **phylogenetic profiles** that try to ensure that functions are maintained through homologs by comparing their interaction networks.

1.3. Protein Loops

The study of protein loops, is key to understand protein structure and its function. In the previous section Protein Structure we defined a loop as the non-regular region linking two consecutive secondary structures.

To understand this principle of non-regularity, we can take a look at the Ramachandran Plot [Figure 1.10] [80]. The Ramachandran Plot is built by a scatter plot representation of all the residues in a non-redundant set of the PDB. The plot represents the backbone's dihedral angles Φ and ψ of each residue, which denote the angles of the $C_\alpha - N$ and $C_\alpha - C$ links respectively. The peptide bond's angle (ω) is normally 180° as its double bound nature keeps it planar. What can be extracted from the plot is that not all the available conformations $[\phi|\psi]$ are possible. Some are especially frequent (favoured) and some are extremely rare or simply impossible (not-allowed combinations). Sequences of **n** consecutive residues whose $[\phi|\psi]$

angles can be found inside the same favoured region are considered secondary structure. The value of **n** varies depending on the type of secondary structure. The rest of protein segments are considered coil and are mostly found in loops.

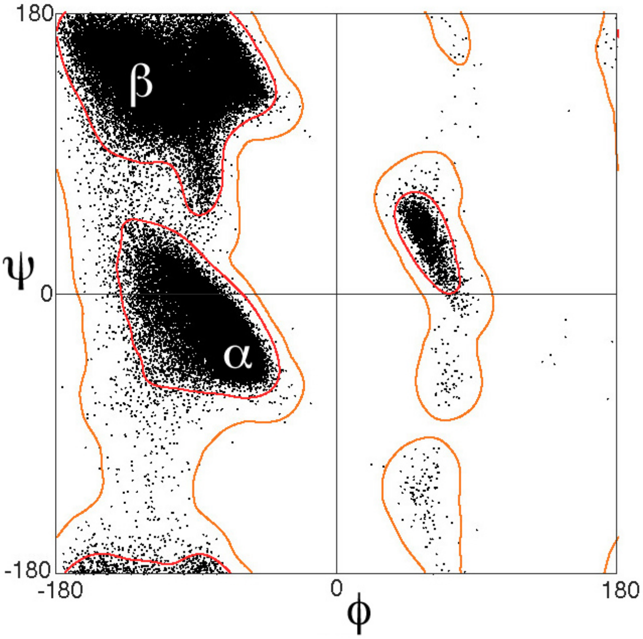


Figure 1.10 Ramachandran Plot. Updated version obtained from [124].

Loops cover a wide section of the PDB. As we commented in the section Protein Structure Determination, flexible regions, usually loops, are a problem in X-Ray crystallography. Despite the fact that NMR can overcome that particular limitation, due to the clear prevalence of X-Rays in the PDB [Figure 1.3], it is not uncommon to find gaps in determined structures of the database. The persistence of loops in the database can be seen in Figure 1.11.

At a structural level, loops play an important role in the folding and dynamics of proteins. Up to the point that, for some proteins, the correct configuration of a loop is a rate-determining step for the folding while, for others, a loop can misfold to serve as a hinge region for domain-swapped species [126]. Loops can also act as hinges facilitating the folding/unfolding process [127], given their intrinsic flexible nature. In addition, it has been shown that long-range loop-loop interactions are important in the folding of proteins [128]. Even the size of the loops has been related to the stability of proteins [129] and their thermo-stability [130]. In extreme cases, a single substitution in a loop can cause the destabilization of the entire protein [131].

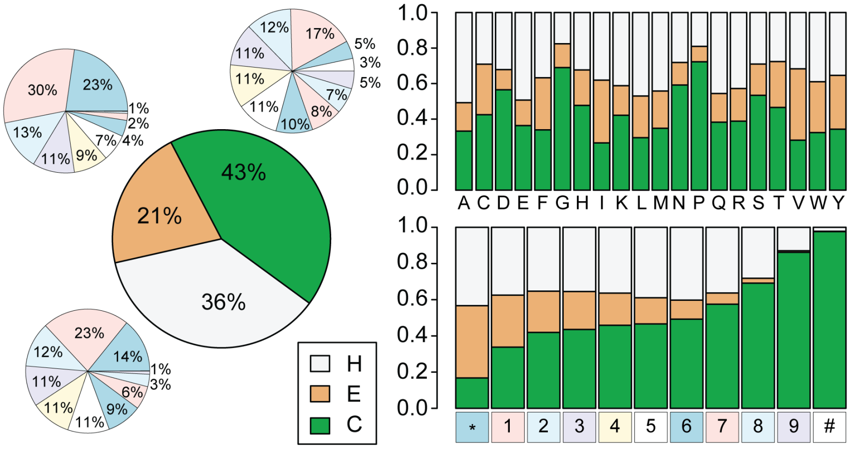


Figure 1.11 Structure and exposition distribution in PDB. Here is represented the distribution of secondary structure and exposition (as calculated with DSSP [125]) over a 90% homology non-redundant version of the PDB (representative sequences selected with CD-hit [50]). The centred pie chart displays the global structural distribution of helix, betas and loops (alternate helix conformations are all clustered together). Each surrounding pie represents the residue exposition distribution for its closest secondary structure. Exposition ranges from * to # in intervals of 10%. The top right barplot represents the incidence of each residue in a given secondary structure, while the lower right barplot represents the percentage of each structural conformation found depending on the residue exposition. Undetermined structural regions are ignored.

Loops also play a central role in the function of proteins and in their associations to other biomolecules. There are several protein families whose functional specificity is regulated by determined loops [132]. This is the case of co-factor binding regions as the P-loop [133], the EF-hands [134], catalytic sites like the serine proteases [135] or Ser/Thr kinases [136]. Given their flexible nature, loops play an important role in the conformational changes of enzymes and often are responsible for the correct positioning of catalytic residues [137], for the function activation through auto-inhibition [138], for recognizing motifs in signalling pathways [139], and even for the regulation of the function efficiency (Appendix 8.6)[140]. Finally, loops are important in protein-protein interaction [141] and recognition (Appendix 8.4)[142] and protein-nucleic acid associations [143]. Of course, and amongst all others, the complementary determining regions (CDR) of the immunoglobulins represent the ultimate example of specific target recognition and biological function [144].

Seeing the importance of loops in defining both the structure and the function of a protein, it has been suggested that they can be used to predict both the 3D conformation of the protein and its activity. Regarding tertiary structure prediction, it is worth to see the value on using know loops to complete and predict structures, especially when considering that most unresolved segments in crystalized proteins of the PDB would correspond to loopy regions due to their degree of flexibility [20]. On the field of function prediction, as long as specific functions have been correlated with certain loop conformations [145], it makes sense to consider the possibility of exploiting those loops in order to predict a protein's activity. As with most data sources, making use of the protein loops will require first the classification and analysis of the available data.

1.4. Motivation

As we have shown during this Introduction, detailed knowledge of a protein's function is a pivotal step towards the comprehension of cellular processes and life itself. A deep understanding of them is not only critical in medical research [146], but also has to allow us to devise non-biological processes for bioremediation through the engineering of proteins and their functional pathways, as are the case of the degradation of crude oil spills at sea [147] or of discarded non-steroidal anti-inflammatory drugs [148] which are, apparently, not fully degradable even by sewage treatment plants.

Fortunately, some relatively new experimental techniques such as next-gen sequencing [47] have allowed us to identify in a high-throughput manner most of the proteins comprising some model organisms proteomes [8]. These techniques are continuously advancing and improving, up to the point that, nowadays, sequencing a full human genome is supposed to cost up to $1000 for a lab with the necessary equipment [149].

Unfortunately, while those techniques allow us to identify possible new and unknown proteins, they do not grant much more information. The scientific community is continuously trying to develop new high-throughput techniques in order to assess the structure, function and interactions of all those newly identified proteins, but, despite some successes [150], there are many limitations to overcome. This is the point in which computational biology makes its appearance.

According to the NIH official definition (July, 2000), computational biology is "the development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological, behavioural, and social systems". In other words, it consists in grabbing known data, analyse it, process it and apply it to fill the gap of knowledge. Applied to our case of interest, it means using the well annotated data in the protein databases to transform next generation sequencing described proteins into fully annotated and categorized ones.

Although the aim of computational biology is helping us to understand through rules why biological entities behave the way they do, from a practical standpoint, computational methods are ultimately devoted to reduce the time and monetary cost of research and provide, with the higher possible guarantees, answers that will match those found through experimental techniques. As they do that by exploiting whatever data is already known, most computational applications should be considered knowledge-based systems.

A knowledge-based system (KBS) is a computer program that reasons and uses known data to solve complex problems. There are two basic types of sub-systems related to KBS: (a) knowledge base and (b) inference engine.

The knowledge base represents facts, often through clustering mechanism and ontology. The inference engine represents logical assertions and conditions based on the acquired and processed knowledge, which can be reduced to more or less complicated sequences of IF-THEN rules. Representing knowledge explicitly via rules had several advantages:

i. **Acquisition & Maintenance.** Pre-established rules codified in the adequate working environment allow experts in a specific field to define and maintain them.

ii. **Explanation.** By representing knowledge explicitly, it allows automatic systems to reach conclusions and keep the traceability of the data in order to understand the process followed to reach the conclusion.

iii. **Reasoning.** Allows the creation of inference engines able to reach and develop new rules outside the initial parameters of the developer. This is why the first KBS were developed by Artificial Intelligence engineers.

Thus, the exploitation of known data is basic in the development of any new computational pipeline. Even thinks as standardized in the modern computational world as are the BLOSUM/PAM matrices [151] or the Ramachandran Plot [80] need to be updated from time to time [124,152]. This is becoming more and more a requisite of most computational developments as the experimental data is being generated at a nothing short of exponential growing rate.

For instance, take a look at the ratio of update of Protein Data Bank: PDB data does not grow as much as genomic or protein sequence databases, and, still, it is updated almost daily [Figure 1.12].

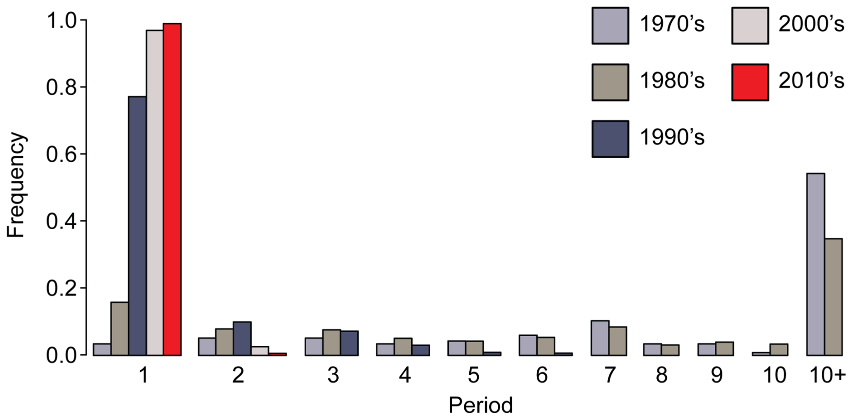


Figure 1.12 Update ratio of the PDB per decades. Frequencies of 0 are omitted. The Protein Data Bank releases weekly a statement of new, modified and deprecated entries, but changes occur at an even faster pace. Let us consider a PDB version as each unique release of the database. That is, during a given version there are no new, modified or deprecated entries. Here, the period (x axis) represents the number of days that a version lasts through the years. It can be from 1 to 10 days or more than 10 days. The change of the update ratio is evident. Actually, in 2013, and according to our working definition, there are 365 versions of the PDB, which means that changes were performed on the databases' contents every day.

This thesis is focused in the analysis of structural protein fragments extracted from the Protein Data Bank [35]. Working with the concept of Smotif (see Chapter 3), it studies new classification algorithms in order to improve the efficiency and coverage of the obtained clusters over the whole of the PDB. Furthermore, it explores new applications of the Smotifs, from modelling (see Chapter 4) to function prediction (see Chapter 5), with an emphasis on *de novo* design of protein regions. The mentioned applications try to offer alternative conformations for the Smotif on a given protein region minimizing its effects over the global scaffold of the protein.

2. OBJECTIVES

This thesis aims to fulfil the following objectives:

i. Devise a method to cluster and categorize protein structural fragments able to manage the actual volume of crystallographic data.
ii. Correlate structurally similar fragment with known protein functions integrated from different functional annotation sources.
iii. Develop a method for the prediction of loop conformation in incomplete protein structures.
iv. Apply the correlation of protein function to loop clusters to predict protein functions both from sequence and structure.
v. Develop a recommendation algorithm for protein redesign able to propose structurally conservative sequence substitutions that can affect the function of the protein.
vi. Create web interfaces to make available the data and its applications.

Points **(i)**, **(ii)** and **(vi)** are analysed in Chapter 3: Classifying Structural Fragments, in which **ArchDB** is presented. The new version of ArchDB features a novel, fast and user-friendly web-based interface, and a novel graph-based, computationally efficient, clustering algorithm. Furthermore, it statistically correlates the obtained clusters to EC [90], GO [91] and DrugBank [153]. The database can be freely accessed, browsed and downloaded at **http://sbi.imim.es/archdb**.

Objectives **(iii)**, **(v)** and **(iv)** are presented in Chapter 4: Modelling with Structural Fragments, which introduces **Frag'r'Us**. The method allows the sampling of protein loops through the geometry of their flanking secondary structure in order to fill protein regions with unknown conformations. This can be used both to fill gaps in proteins with feasible templates to model an incomplete protein as well as to offer new alternative backbone conformations for the region of interest. Frag'r'Us is available at **http://www.bioinsilico.org/FRAGRUS**.

Finally, the points **(iv)**, **(v)** and **(vi)** are discussed in Chapter 5: Predicting Protein Function with Structural Fragments. This chapter presents **Archer**. The method exploits ArchDB's hierarchy of supersecondary structures to map GO [91] and Enzyme [90] functions upon protein regions and, thus, infer the function of a protein. It relies on either the sequence or structure of the protein of interest and returns the mapping of functional subclasses extracted from ArchDB. Moreover, it computes the functional enrichment and significance of each subclass, combines the functional descriptors and predicts the function of the query-protein. Furthermore, it offers variants of the target sequence that swap the region of a supersecondary structure by another that putatively fits in the same scaffold. It is accessible at **http://sbi.imim.es/archer**.

3. CLASSIFYING STRUCTURAL FRAGMENTS

Due to their non-regular nature, loop classifications encounter complications even before starting to group their items of interest. One of those issues is the definition of the protein fragment and its properties.

This chapter revolves around the topic of protein loop classification and it is divided in three sections; the first one devoted to review some of the existing approximations to protein fragment clustering.

The second section pivots around the explanation of the Smotif loop definition and the history of the different versions of ArchDB until reaching to today's last update. This section is heavily based on:

Bonet, J., Fiser, A., Oliva, B., & Fernandez-Fuentes, N. **Smotifs as structural local descriptors of super-secondary elements: classification, completeness, and applications.** *BAMS*. (*In press*)

After getting the required perspective, the chapter focuses upon ArchDB 2014 through its published article. Supplementary Figures 1 and 2 of the article are not included in the thesis content but can be accessed online.

Bonet, J., Planas-Iglesias, J., Garcia-Garcia, J., Marín-López, M. A., Fernandez-Fuentes, N., & Oliva, B. (2014).
ArchDB 2014: structural classification of loops in proteins.
Nucleic Acids Research, 42(1), D315–9.
http://dx.doi.org/10.1093/nar/gkt1189

Finally, the chapter is closed with a summary of the advantages and limitations of the new classification over previous versions as well as over other loop approximations.

3.1. Protein Fragments Classifications

There exist multiple attempts to classify fragments of proteins. Here we will review some of those devoted to the general classification of protein fragments though structural features. This means that we will not discuss:

i. Exclusively sequence-based classification methods such as PROSITE [106].
ii. Methods devoted to specific types of supersecondary structure such as β-turns [154], β-strands [155] or α-helices [156].
iii. Protein domain classifications such as SCOP [28] or CATH [30].

a) Sequence Sliding-Window Fragments

Sequence Sliding-Window approximations to protein fragment definition and clustering (**SSW** from now on) are based on the consecutive fragmentation of the protein and are collected by global structure similarity. Although they might use secondary structure as part of they process, usually it is not used as one of the main descriptors of the fragment. As there are multiple similar approximations, we will select some representatives on general methodologies.

RRW

Rooman, Rodriguez and Wodak [157] defined one of the earlier methods of fragment clustering (**RRW** method from now on). They used 75 high-resolution proteins (<2.5Å) and divided them into fragments of 4 to 7 residues. As the method clusters only fragments of identical length, this results into 4 parallel classifications, one for each length. They constructed a hierarchical classification by pairs. The method works as follows for each classification:

First, distances between each pair of fragments are calculated as the RMS deviation (**RMSD**) between the inner distances of the C_α atoms.

$$RMS(X,Y) = \sqrt{\frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(d_{ij}^{x} - d_{ij}^{y}\right)^{2}} \qquad \textbf{Eq 3.1}$$

where **X** and **Y** are the fragments, **n** is their length, and $\boldsymbol{d_{ij}^{x}}$ and $\boldsymbol{d_{ij}^{y}}$ are the distances between the residues **i** and **j** of **X** and **Y** respectively. According to the authors, this method is computationally more efficient than the RMSD calculated over superimposed structures, which should be a requirement considering that, by sliding window partition, they ended up with more than a 10.000 fragments from those 75 proteins.

Once the first pairs are calculated, a similarity index between each cluster needs to be defined. They called it the inertia coefficient **I**.

$$I(A) = \sum_{X \in A} RMSD^{2}(X,g) \qquad \textbf{Eq 3.2}$$

with

$$d_{ij}^{g} = \frac{1}{P(A)}\sum_{X \in A} d_{ij}^{x} \qquad \textbf{Eq 3.3}$$

where **X** represents the elements of the cluster **A**, **P(A)** is the size or number of elements of the cluster and **g** is the centre of mass expressed as a virtual fragment in which each position is averaged from all the elements positions [Eq 3.3].

At each clustering step, two clusters are merged under the condition that they keep the global inertia coefficient of the system to a minimum. Thus, while the inertia coefficient of the first clusters is considered 0, a new cluster's **I** is the sum of the inertia coefficient of the two joint clusters plus the squared distance of their centres of mass.

With this methodology, they ended up with a classification tree with leaves of **I(A)=0** and a single cluster root with **I(R)=max(I)**. At this point, a given **I** is selected as threshold and the top nodes of the tree under that threshold are selected as the clusters of the classification. Those clusters are fragmented into groups of fragments with $[\phi|\psi]$ angles in the same domain of the Ramachandran Plot [80]. Different **I** thresholds can be defined depending on the experiment, which will determine a different number of clusters.

FragKB

FragKB [158] creates its set of protein fragments with a 8 residues sliding window. Each fragment is defined by its sequence and backbone C_α atoms. With the Cα atoms, tetrahedrons are constructed by joining each possible sequence of 4 consecutive atoms (tetrahedron_gap_0) and each possible sequence of 4 non-consecutive atoms (tetrahedron_gap_1); all together, 6 tetrahedrons are generated from the fragment. Each tetrahedron is represented by a set of 9 Geometrical Invariants (**GIs**), generating a total of 56 GIs. Thus, given a fragment **X**, each GI can be considered as a function $\boldsymbol{f_n(X)}$. Two fragments **X** and **Y** are considered superimposable (similar) as long as, for each $\boldsymbol{f_n(X)}$, its equivalent $\boldsymbol{f_n(Y)}$ fulfils:

$$f_n(X) = f_n(Y) \pm \delta_n \qquad \textbf{Eq 3.4}$$

where $\boldsymbol{\delta_n}$ is the threshold of the window of confidence for that GI. Thus, through GIs it is possible to evaluate the similarity of **X** versus **Y** without actually perform the superimposition.

By means of this procedure, fragments are located in a 56-dimensional space, each dimension defined by one of the GIs. Clustering is preformed through a breath-first method. Briefly, given thresholds for each GI, fragments are clustered as they satisfy the first GI, the components of each obtained cluster are gathered by satisfying the second GI and so on.

Once the first clusters are created, the centroid of a cluster **k** is defined as the fragment closer to all the rest by all given GIs [Eq 3.5].

$$d(X) = \min\left(\sum_{Y \in K} dist(X,Y)\right) \qquad \textbf{Eq 3.5}$$

Then, those clusters adjacent in the 56-dimensional space are grouped as long as their centroids satisfy a maximum allowed width for each defined GI. Finally, a hierarchical clustering algorithm is applied on the clusters in order to obtain a hierarchical classification.

BriX

BriX [159,160] works with fragments between 4 and 14 residues over a set of non-redundant proteins extracted from ASTRAL40 [161]. All consecutive overlapping fragments are considered, and their secondary structure described with DSSP [125]. Similarly to RRW, clustering steps are applied to each length individually.

The first classification step consists on grouping the fragments according to their secondary structure (DSSP groups). From this point on, each fragment is considered only as the 3D coordinates of its backbone atoms. Now, an iterative process is performed on each DSSP group in order to create the final classification. The process is base upon Hierarchical Agglomerative Clustering (HAC). Briefly,

i. A centroid is defined for each cluster. In the initial step, each fragment is its own centroid, in subsequent steps; it is a representative fragment of the given cluster.
ii. Fast RMSD [162] is used to create a distance matrix between the centroids.
iii. At each iteration the clustering algorithm is applied through the distance matrix and a increasing distance threshold

$$Threshold = 0.5 + k * 0.1 \qquad \textbf{Eq 3.6}$$

where **k** is the iteration number.

iv. By gathering the results at each iteration, the hierarchy is created.

b) Secondary Structure Fragments

These methods rely on the definition of the secondary structure previously to split the protein into fragments. The generated clusters do not need to be limited to segments of similar size. **ArchDB** (see following section) belongs to this category of protein fragment clustering methods.

TPL

The **Taxonomy Protein Loops** clustering method [163] defines loop fragments as regions connecting two secondary structures identified through the consensus (2 out of 3) of DSSP [125], Define [164] and P-curve [165]. The 1586 loops that constitute the database are generated from a non-redundant 50% homology set of PDB structures. Only loops from 3 to 8 residues are considered.

Loops are then separated according to their length and scattered through a 3D landscape described by 3 directions:

The **first direction** is defined by the vector originated in the first C_α atom and finishing in the last C_α atom.

The **second direction** is the orthogonal of the first one. Together, they define the plane that contains the centre of mass of the loop's backbone.

The **third direction** is the vector product of the two others.

Once the first grouping has been performed amongst the loops of the same length; the loops in each group are clustered by means of the complete link algorithm (CLA) over a RMS distance matrix of all versus all. The RMSD is calculated for the backbone residues N, C_α and C, and for a length $N-2$ as the first and last residues are excluded from the calculation. Thus:

$$RMS(X,Y) = \frac{1}{N}\left(\sum_{i+1}^{N-1}(X_i - Y_i)^2\right)^{1/2} \qquad \textbf{Eq 3.7}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the fragments to be compared and $\mathbf{N}$ their shared length. With the application of the CLA, two individual loops are clustered if their RMSD is the minimum amongst all. Each new loop (**X**) is assigned to a cluster (**K**) as long as $D(K,X)$ [Eq 3.8] is smaller than a given threshold.

$$D(K,X) = max_{Y \in K,X}\ RMS(X,Y) \qquad \textbf{Eq 3.8}$$

The process is iterated until all loops have an assigned cluster. RMS thresholds range from 0.6Å for 3 residue loops up to 3.0Å for loops of length 8.

HMM-SA

Structural alphabets such as HMM-SA [166] are another mechanism to approach the protein fragment classification. The initial version of this dictionary, which fits upon the SSW category, is built over 4-residue length overlapping fragments representing the hidden state of a HMM. Constructed over a 30% non-homologous set with a resolution better that 2.5Å of 1.409 sequences, it is composed of 27 structural letters (alphabet space) that can be translated into 3D coordinates. Each protein of the database is, then, defined by L-3 letters, being L the length of the protein.

The basic alphabet goes as follows: [A, a, V, W] are considered α letters, as they appear in the helices. [Z, B, C] belong to the helix termini. β letters, found in strands, are [L, M, N, T, X]. [J, K] belong to the strand termini. Any other letter defines loop regions. The dictionary not only defines loops between secondary structures but also on the N and C terminus of the protein. It is important to note that, from this point on, letters will refer to the HMM-SA alphabet while residues will refer to the amino acid alphabet.

To build a loop classification, loop words of residue length k+3 (being k the letters of the motif), denoted W_k and called coil-words, are created. The exceptionality (statistical significance) of each coil-word is evaluated as a p-value defined by the variation of the occurrence of the word in the dataset

$O[N(w)]$ from the expected occurrence of the word in the background $E[N(w)]$. Those coil-words that appear 150 times more than expected are selected.

The classification is obtained by hierarchical clustering using $RMSd_{dev}$ as distance measure. $RMSd_{dev}$ or structural dissimilarity index is the average C_{α} RMSD of 200 fragment pairs randomly selected from two different words.

Loop Brix

Since its first release, BriX has been extended with Loop BriX [160]. This new classification does not depend on the length of the fragments, but only in their lack of secondary structure.

In this case, the distance matrix required to apply the Hierarchical Agglomerative Clustering (see BriX) is based on (a) the distance between the endpoint of the loop, and (b) the superimposition of the two anchor or stem residues (residues in the flanking secondary structure immediately before and after the loop). The generated clusters might contain loops of different sizes and orientations, and a second round of clustering is performed grouping loops by length and similar structure.

3.2. Smotifs and ArchDB

a) Defining Smotifs

The common underlying concept that connects the different aspects of this thesis is the working definition of protein loops as structural motifs (**Smotif**).

We define and **Smotif** as a protein loop (**C**) of length **M** flanked by $\boldsymbol{n_N}$ and $\boldsymbol{n_C}$ stem residues, i.e. the residues with known structure that precede ($\boldsymbol{n_N}$) and follow ($\boldsymbol{n_C}$) the loop, but are not part of it. The number of residues flanking the loop depends on the secondary structure to which the stem residues belong. The considered secondary structures are α-helix (**H**), 3_{10}helix (**G**) and β-strand (**E**).

$$Smotif = \left.\begin{matrix} H_1 \dots H_4 \\ G_1 \dots G_3 \\ E_1, E_2 \end{matrix}\right\} C_1 \dots C_M \left\{\begin{matrix} H_1 \dots H_4 \\ G_1 \dots G_3 \\ E_1, E_2 \end{matrix}\right. \qquad \textbf{Eq 3.9}$$

The definition of loops as Smotifs allows us to define the geometry of a non-regular protein segment by the local structural arrangement of its flanking regions [Figure 3.1].

There have been slight variations in the algorithm to define Smotifs since the first version was described [167,168]. The most remarkable is the inclusion of the 3_{10}helix (G). The following explanation on how to define Smotifs will be centred in the most recent implementation.

Axis **M1** and **M2** are defined over the N-terminal and the C-terminal flanking secondary structures respectively. Each axis is created according to the shortest moment of inertia of its structure in the segment joint by the loop. This method is used in order to take into account the possibility of a certain degree of curvature in the secondary structure; especially in the β-strand. The number of residues used to calculate the moment of inertia is fixed in both α-helix and the 3_{10}helix at 4 and 3 residues respectively, but it varies between 2 and 3 residues in the β-strand [Eq 3.9]. If **Ij** is the axis built considering the residues up to **j**, we will select:

$$M_x = I_3 \; \boldsymbol{if} \; angle(I_2, I_3) < 10^{\underline{o}} \; \boldsymbol{else} \; I_2 \qquad \textbf{Eq 3.10}$$

Once the to axis are built, we can describe the rest of geometrical variables required. **P1** and **P2** are the start and the end points of the loop and the vector that joins **P1** and **P2** is **L**. The plane **π** is defined by the vector **M1** and L. The plane **τ** is defined by **M1** and the normal to **π**. With all that, we define the geometrical descriptors of a Smotif.

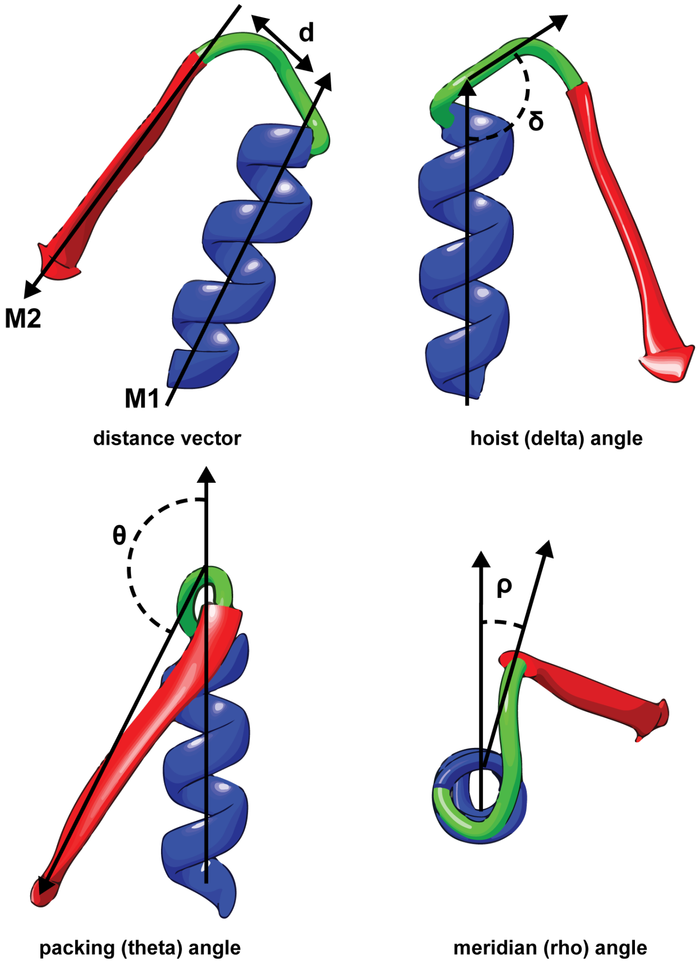


Figure 3.1 Smotif geometry definition. All the flanking secondary structures are represented despite that only selected stem residues do belong to the Smotif.

Distance (d). Is the Euclidian distance between **P1** and **P2**. **(D=|L|)**

Hoist (δ). The hoist or delta angle is the one defined between **M1** and **L**.

Packing(θ). The packing or theta angle is defined between **M1** and **M2**.

Meridian(ρ). The meridian or rho angle is defined between **M2** and **τ**.

In the current distribution of ArchDB, for all Smotifs (both classified and not classified), **d** covers an interval between 1.7 and 187.3Å, with an average value of 9.65Å. Angles **δ** and **θ** range from 0 to 180 degrees, while **ρ** ranges from 0 to 360 degrees.

The last item of Smotif identification is the actual nature of the flanking secondary structures. By considering α-helix (H), β-strand (E) and the 3_{10}helix (G), we can define a total of 10 Smotif **types** by covering all the possible combinations. It is worth noting that the EE combination does not exist as it is spliced into BK (β-links) in which the two flanking β-strands do not contact with each other and BN (β-hairpin) in which they do.

In protein loop classification, geometry is a useful descriptor to group loops and speed up the clustering process. In loop structure prediction and protein design, the geometry allows an optimized hashing and look up of loop conformations given a set of geometrical restraints. Finally in protein structure prediction and design, Smotifs provide a convenient and coherent scheme to break down proteins into a sum of super-secondary elements.

b) Classification of Smotifs

The first classification of Smotifs [167] was generated from 233 high quality crystallographic X-ray structures obtained from the Protein Data Bank (PDB) [169] (resolution better than 2.5Å), after removing redundancy at 25% sequence identity cut-off. For each of these, α-helices and β-strands were defined using DSSP [28,125], which produced a total of 3005 Smotifs. Subsequently, Smotifs were clustered according to their geometrical properties using a density search (DS) algorithm, which is a variant of the single-linkage clustering method [170]. Briefly, a network is built in which the nodes are the Smotifs and the edges are defined by the similarity of the classifying attributes of the Smotifs. The DS algorithm detects regions within that network with high density of Smotifs around a centroid defined by the classifying attributes of the Smotif. In this clustering, loops that belong to the same cluster have a length variation of ±1, similar flanking secondary structures, and similar $[\phi|\psi]$ angles (identified by a consensus conformation). Each cluster is required to have at least 3 Smotifs. The whole process generated a 121 structural subclasses that where further grouped according to their Ramachandran map patterns into 56 classes [Table 3.1].

ArchDB	Method	Source	PDB	Smotifs	Subclasses	Classes
1997	DS	25	233	3005	121	56
2004	DS + rc	40	2310	12665	1496	451
2007	DS + rc	95	5472	36153	4023	2142
2007	DS + rc	40	3640	16957	2550	1119
2007	DS + rc	EC	2349	20260	2686	1338
2014	DS	40	17961	129280	13198	5362
2014	MCL	40	17961	187117	12240	9728

Table 3.1 ArchDB in numbers through time.

ArchDB is organized in a hierarchical fashion: the two first levels of the hierarchy correspond to the flanking secondary structures (type) and the length of the loop (length). The third level of the classification corresponds to classes, which are formed of subclasses with similar Ramachandran map patterns but different geometry. The lowest level of the hierarchy is the subclass, which are the structural cluster of Smotifs, i.e. Smotifs with the same loop conformation and geometry. This schema is used in all versions of ArchDB regardless of the particularities of the algorithm applied for clustering (see in Figure 3.2 in the following section).

The first update of ArchDB [171] made the database available online and introduced minor details in the classification such as the maximum identity between source structures (40%, ArchDB40), the upper limit on resolution (3.0Å) and the minimum number of loops in a cluster (2). A re-clustering algorithm was applied after the first clustering to merge subclasses with shared loops, resulting in an optimized partition of the conformational space [Table 3.1]. Furthermore, it included references to Gene Ontology (GO) [91] and Enzyme [90] annotations. The Enzyme annotation was further exploited for the analysis of kinase super-families and their relation to Smotifs [145]. The number of sub-classes and classes increased significantly [Table 3.1].

The third release of ArchDB included two new sets: ArchDB95, a redundant set, and ArchDB-EC a classification derived from protein enzymes [172]. The new release included extensive functional annotations and cross-references to major biological databases and an increase in the number of classified Smotifs, classes and sub-classes. The new database was used both for modelling of loops (ArchDB40) and study relevant structure-function features in loops (ArchDB95 and ArchDB-EC set). It also included a comprehensive study of the statistical correlation between ArchDB subclasses and GO, EC and SCOP [28] annotations as well as atomic interactions to co-crystallized cofactors and additional functional annotation extracted from PDB [105].

3.3. ArchDB 2014

Bonet J, Planas-Iglesias J, Garcia-Garcia J, Marín-López MA, Fernandez-Fuentes N, Oliva B. ArchDB 2014: structural classification of loops in proteins. Nucleic Acids Res. 2014 Jan;42(Database issue):D315-9. doi: 10.1093/nar/gkt1189.

3.4. Advantages of ArchDB 2014

There are two basic types of contributions of ArchDB 2014. On the one hand, the Smotif definition, all by itself, presents a series of advantages over other loop definitions. On the other hand, it presents several improvements over older versions of the database.

a) With respect to other fragment classification methods

i. **Definition**. Each Smotif has a set of parameters that describes it. Thus, each Smotif has its own identity. Furthermore, these parameters provide a comprehensible description of the actual shape of the Smotif.

ii. **Variable Length**. As Smotifs are defined through their stem residues, there is no real requirement for keeping an identical sequence length amongst all the members of the classification. Although other secondary structure based methods behave similarly, some, by using full backbone RMSD as a clustering parameter, end up enforcing all fragments belonging to a same cluster to be of the same size.

iii. **Contained Dataset Size**. SSW methods, even after filtering the homology of the source PDB set; present a huge overlap of fragments. In fact, each fragment is bound to have at least $(k * 2) - 2$ overlapping fragments (being k the length of the fragment). Overlapping fragments in Smotifs can only happen between consecutive loops separated by small secondary structures, and, at maximum, a Smotif can have two overlaps. This is, though, a benefit shared by all the secondary structure based methods.

iv. **Local Structural Environment.** By considering the moment of inertia of the N- and C-terminal secondary structure instead of just the stem residues, the Smotif can better capture the pre- and post-conditions of the coil arrangement.

b) With respect to previous ArchDB releases

i. **New Smotif Types.** Five new Smotif types are included by considering 3_{10}helix as regular secondary structure – before these were considered part of the loop regions as α-helices were considered only if they exceeded 5 residues. The new Smotifs include 3_{10}helix-3_{10}helix, 3_{10}helix–α-helix, 3_{10}helix–β-strand, β-strand-3_{10}helix, and α-helix-3_{10}helix.

ii. **Fixed and Variable Length Classifications.** Instead of limiting the clusters to a ±1 length variation, we now have a fixed-length classification through the DS algorithm and a variable length classification through MCL. With a variable length classification we can explore the similar effects of loops of different size over the 3D conformation of the protein while, simultaneously, we extend the coverage of the classification over the complete set of Smotifs.

iii. **Scalability.** The new MCL algorithm is capable of processing a substantially bigger data size. This makes it ideal to cope with the expected growth of the PDB.

4. MODELLING WITH STRUCTURAL FRAGMENTS

This chapter will be focused on protein modelling and *de novo* protein design through the use of protein loops. Its structure will be similar to the previous chapter, with an overview of existing methods to bridge missing loops in protein structures followed by Smotifs applications devoted to both global protein structure prediction and loop structure prediction. Only knowledge-based methods are discussed. The section referring to past Smotifs applications is based upon:

Bonet, J., Fiser, A., Oliva, B., & Fernandez-Fuentes, N. **Smotifs as structural local descriptors of super-secondary elements: classification, completeness, and applications.** *BAMS*. (*In press*)

After this introduction, it will focus upon Frag'r'Us through its published article. The supplementary material of the article will be added at the end of its section.

Bonet, J., Segura, J., Planas-Iglesias, J., Oliva, B.,
& Fernandez-Fuentes, N. (2014).
Frag"r"Us: knowledge-based sampling of protein backbone conformations for de novo structure-based protein design.
Bioinformatics (Oxford, England).
http://dx.doi.org/10.1093/bioinformatics/btu129

Finally, a brief comment on the advantages and opportunities that offers the server will close this chapter.

4.1. Fragments applications in Protein Structure

The prediction of loop conformations through knowledge-based approaches is done by searching amongst potentially thousands of conformations extracted from known protein structures. The target loop is flanked by so-called stem residues, i.e. the residues with known structure that precede and follow the loop, but are not part of it. The search implies placing potential loops that fit the restraints of stem residues followed by their ranking based on geometric criteria and/or sequence similarity. Finally, selected loops are superposed and annealed onto the stem regions.

LIP

The **Loops In Proteins** (**LIP**) method [191] is based in a non-clustered dataset of loop fragments. The fragments are created from a <20% homology non-redundant dataset of structures from the PDB with a resolution <3.6Å and the secondary structure defined with DSSP.

Each fragment is defined by:

i. **length**

ii. amino acid **sequence**

iii. (**x**, **y**) values, a two dimensional vector between the $C^{(N)}$ and the $N^{(C)}$ atoms in the ${C_\alpha}^{(N)} - C^{(N)} - N^{(C)}$ plane such as the distance between the two stem residues is:

$$dist_{stem} = (x^2 + y^2)^{1/2} \qquad \textbf{Eq 4.1}$$

Being the atoms labelled **(N)** those belonging to the N-terminal stem and those labelled **(C)** the ones belonging to the C-terminal, that means that (**x**, **y**) are the opposite sites to the hypotenuse that represents de distance between the stem residues.

iv. $\boldsymbol{\beta}$, the angle included by the lines of $C^{(N)} - N^{(C)}$ and $N^{(C)} - C_\alpha^{(C)}$.

v. $\boldsymbol{\gamma}$, the dihedral angle between ${C_\alpha}^{(N)} - C^{(N)} - N^{(C)}$ and $C^{(N)} - N^{(C)} - C_\alpha^{(C)}$.

To select the loop candidates to fill an incomplete structure, all the loops of the required length with the N- to C-terminal distance matching the query gap with a tolerance of 0.75Å are selected. Then **goodness** is calculated for each of the remaining candidates as a correlation of the square differences of the geometrical identifiers:

$$goodness = \Delta x^2 + \Delta y^2 + 2(\Delta\beta^2 + \Delta\gamma^2) \qquad \textbf{Eq 4.2}$$

The putative solutions are, finally, ranked.

LIP prediction predictions reported a $RMSD_{local}$=1.71Å on average on loops 14 residues long, obtaining better results on smaller loops.

Loop BriX

Loop BriX [160] exploits their classification (as explained in the Loop Brix section of Classifying Structural Fragments) to fill a gap in a protein structure. Their bridging algorithm matches subclasses to the gap accordingly to the distance between that and the centroid of the class. The amount of fragments per class and the similarity thresholds applied are managed by the user.

Schomburg

Schomburg's loop prediction method also relies in the position of the anchor groups [192] and the filtering of the selected templates [193].

It depends upon a non-clustered set of protein loops extracted from a 95% homology non-redundant database with a resolution equal or lower to 2Å. Fragments with a RMSD under 0.25Å after the alignment of the stem residues were filtered. A distance (**d**) is defined for each fragment as the Euclidian distance between the middle of ${C_\alpha}^{(N)} - C^{(N)} - O^{(N)}$ and the middle of $N^{(C)} - {C_\alpha}^{(C)}$. Atoms labelled **(N)** belong to the N-terminal stem and those labelled **(C)** belong to the C-terminal. A good fit is assumed if the difference in distance between a fragment and the query gap is smaller that 0.5Å.

After that, loops clashing with the rest of the protein, as well as those presenting unfavourable torsion angles according to a knowledge-based potential are eliminated.

4.2. Smotifs applications in Protein Structure

a) Global Protein Structure Prediction

Smotifs can be used for protein structure prediction, following a fragment assembly approach. The underlying hypothesis is that patterns of indirect structural data characterizing the connecting loop region in a Smotif will determine the relative orientation of flanking secondary structures and thus will be informative for the selection of an entire super-secondary structure element.

As Smotifs are backbone-only defined fragments, a relation needs to be made between a target sequence and the backbone-only library of Smotifs. One possible way to do this is hybrid modelling, where a limited amount of easily obtainable, indirect experimental data is used to select Smotifs for structure modelling. One possible data that can be used in hybrid modelling is to obtain chemical shift (CS) assignments from NMR studies for the target protein. The combination of Smotifs with CS is the base of SmotifCS.

As a clarification, the chemical shift is the resonant frequency of a nucleus relative to a standard. As we explain in the NMR spectroscopy section of the Introduction, some atomic nuclei, like ^{1}H, ^{13}C and ^{15}N, possess a magnetic moment that gives rise to different resonance frequencies and energy levels in a magnetic field. According to the local geometry (bond lengths, binding partners, angles between bonds…) the electron distribution of a given type of nucleus varies, and with it its local magnetic field. This variation of NMR frequencies for a given nucleus type is called the chemical shift. The size of the chemical shift is given with respect to a reference sample.

Thus, the first step of SmotifCS is to calculate the theoretical chemical shift of all backbone atoms N, HN, H_α, C_α, C') and C_β using SPARTA+ [194]. Next, the structure prediction algorithm relies on another pre-calculated database that contains the relative weights of structural information conveyed by a given normalized chemical shift. The construction of this database is as follows: predicted CS values aggregated from all library Smotifs were divided into groups based on atom type (6), residue type (20), and preceding residue type (20), resulting in 2400 categories. For each category, CS values were normalized by subtracting the random coil value. The relative weight of structural information conveyed by a given CS (according to the three parameters described above) is calculated as the difference between the statistical propensities of the "most favoured" and "second-most favoured" secondary structural conformations.

In order to identify the relative orientation of regular secondary structures within a Smotif, the CS patterns of the loop segments and the three flanking secondary structure residues on each side of the loop are analysed. In order to select candidate Smotifs from the library, the experimental CS of each query Smotif and the theoretical CSs of available Smotifs in the library are compared. Theoretical $[\phi|\psi]$ angles predicted with TALOS+ [195] are used to assign each loop residue of the query Smotif in one of the 11 possible locations within the Ramachandran map [182]. The string of Ramachandran Map sub-locations constitutes the "fingerprint" of loop segments that is compared to similar fingerprints derived from the Smotifs of the library. The best matching Smotif fingerprints are then ranked by their CS match "score" calculated as the sum of weighted squared differences between the chemical shifts of the query and library Smotifs.

After a set of suitable candidates is selected for each putative Smotif in the query structure, a full enumeration of the structures is carried out by joining every possible combination of these Smotifs. The lengths of the secondary structures of the sampled Smotifs are extended or shortened as necessary to fit the query sequence. In the process of joining Smotifs, a limited number of steric clashes are allowed. The candidate structures resulting from the full enumeration are evaluated using a linear scoring function with the following components: (a) radius of gyration using C_α carbons; (b) a distance-dependent statistical potential function [196]; (c) an implicit solvation potential [197]; and (d) a knowledge-based long-range backbone hydrogen-bonding potential [198]. All components are converted into statistical Z-scores before combining them with weights optimized on a set of decoy structures. The best 200 structures from this ranking are relaxed using MODELLER [66] to resolve steric clashes and maintain stereochemistry.

The accuracy of the models generated with SmotifCS was evaluated with RMSD and GDT_TS scores [199] against a dataset of 102 NMR structures, each one representing a different SCOP fold [28]. 47 out of the 102 models obtained a $\mathrm{GDT\_TS} >= 50\%$, indicating that, for about half of the models, a high quality homology model was generated. For all cases, at least a topologically correct fold was produced.

b) Loop Structure Prediction

In this section we will focus on a specific knowledge-based approach for loop structure prediction by means of Smotifs: ArchPRED [200]. This method relies on a library of Smotifs and features a selection, filtering and ranking algorithm to select the most suitable conformation for a given target loop sequence.

The **selection** of Smotifs from the library is based on the geometrical restraints imposed by the bracing secondary structures of the missing loop, i.e. Smotifs will be selected if the geometry is similar or fall within the range of tolerance: 2Å in the case of the distance (d) and 30, 30 and 45 degrees in the case of the angles hoist (δ), packing (θ) and meridian (ρ) respectively.

Next, a **filtering** step discards unsuitable Smotifs based on the structural matching of stem residues: $\text{RMSD}_{\text{stems}}$ and unfavourable interactions between Smotifs and the new protein environment such as steric crashes. The $\text{RMSD}_{\text{stems}}$ was shown to correlate with the quality of prediction both for filtering and scoring purposes [193]. However, the correlation is less pronounced in the case of loops longer than 8. Finally, the filtering step evaluates the fitting of Smotifs in the new environment. This aspect is particularly important, as the native structural environment of Smotifs could be very different from one in the target protein.

The last step in the prediction process is the **ranking** of the remaining Smotifs. The scoring function is comprised of a sequence similarity score based on a Conformational Similarity Weight (CSW) matrix [201] and an amino acid $[\phi|\psi]$ dihedral angle propensity term [202]. Given the fact that sequence and propensity scores have different dimensions, these are converted into dimensionless statistical Z-scores, which are obtained in reference to randomly generated sequences and $[\phi|\psi]$ dihedral angles. The final scoring function is then a composite Z-score combining the two types of: sequence and $[\phi|\psi]$ dihedral angles propensity Z-scores. ArchPRED results are ranked by means of that final score.

Compared with other *ab initio* methods, ArchPRED was able to perform similarly to ModLoop [203] for loops of length between 4 and 14 amino acids. However, the results show the dependence of ArchPRED in the Smotif database, thus suggesting that its accuracy and applicability should increase with the increase of information in the structural databases and, the classification of Smotifs.

4.3. Frag’r’Us

Bonet J, Segura J, Planas-Iglesias J, Oliva B, Fernandez-Fuentes N. Frag'r'Us: knowledge-based sampling of protein backbone conformations for de novo structure-based protein design. Bioinformatics. 2014; 30(13): 1935-6. doi: 10.1093/bioinformatics/btu129

4.4. Applications of Frag'r'Us

The methodology behind knowledge-based redesign of a protein's loop structure is quite similar to that of protein loop prediction. In fact, the **Loop BriX** server is theoretically able to provide alternative conformations by superimpose classes instead of subclasses over the stem residues that flank the gapped coil region.

The clear advantage of using Smotifs and its geometry rather than distances and/or structural fitting of stem residues is that the search space is reduced dramatically and the geometry-based filtering is very fast.

Thus, Frag'r'Us makes available a method capable of offering a limited but informative set of putative backbones that (a) is compatible with the local secondary structure of the flanking regions, and (b) is compatible with the global tertiary structure of the protein through the avoidance of residue clashes. All of this increases the probability of the suggested changes of not affecting the global conformation of the protein.

Furthermore, by providing a limited set of alternative conformations, those can be studied in detail in order to look for specific changes that can improve the thermo-stability of the protein, add putative post-transcriptional modification sites or even alter the functionality of the protein (we will see more on that point in the next chapter).

5. PREDICTING PROTEIN FUNCTION WITH STRUCTURAL FRAGMENTS

This fifth chapter is devoted to protein function prediction and functional *de novo* design with Smotifs.

As the methods for functional annotation transfer have already been reviewed in the Introduction, this section will directly the two previous approximations to functional annotation with ArchDB and then focus on Archer as our new approximation to the protein functional annotation problem. The body of this section is contained as an article (*to be submitted*).

Bonet, J., Garcia-Garcia, J., Planas-Iglesias, J.,
Fernandez-Fuentes, N., & Oliva, B.

Archer: Predicting protein function using local structural features. A helpful tool for protein redesign.

(Manuscript to be Submitted)

Finally, a brief comment on the advantages and opportunities that offers the application will close this chapter.

5.1. Smotif functional correlation

ArchKI

ArchKI [145] was one of the first approximations to tackle the functional annotation of Smotifs. It was specifically devoted to kinase proteins.

The initial construction of the database was parallel to that of ArchDB [171], with the exception that source structures were obtained from a non-homologous subset of the PDB with assigned EC number 2.7.X.X (i.e. kinase or phospho-transferase function) [90].

Individual Smotifs were assigned functional annotation if they contained: (a) residues within a cut-off distance of 6Å from an heteroatom, ligand, inhibitor, cofactor or complex partner molecule (i.e. interactor), (b) residues identified as active sites by the PDB annotation or (c) residues identified by the functional annotation collected from literature and assigned to specific motifs of kinases.

The functional annotation of the residues was divided in 4 functional categories; namely, (a) adenosine triphosphate (ATP) binding, (b) substrate binding (except ATP), (c) ion interaction and (d) catalytic for residues involved in the catalytic reaction or the stabilization of a transition state.

After the classification, a PROSITE-like pattern [106] was obtained by the alignment of the members of each cluster. Similarly, a position specific scoring matrix (PSSM) was derived from the alignment to quantify the degree of conservation of the different positions in the alignment.

Finally, clusters were categorized as “functional” if there was a meaningful conservation of the functional residues and more than 50% of the loops in the cluster belonged to the same SCOP superfamily [28]. They were labelled as “structural” otherwise.

The potential application of ArchKI to loop modelling was tested as a n-fold cross-validation in which loops were taken out of the clusters. The

PSSMs of the new clusters were re-calculated and the extracted loops were aligned and given a normalized z-score of the sequence-PSSM fitting.

ArchFun

ArchFun [105] directly exploited what at that time was the current release of ArchDB [171] to perform a GO [91] correlation analysis with the Smotif clusters and develop a GO functional annotation method.

Functional transfer of GO terms to Smotifs was performed by direct protein association. That is, all Smotifs belonging to a given protein were given the same GO terms as that protein, as well as their parent terms in the ontology hierarchy (with the exceptions of the first two levels and the terms occurring in more than 10% of the classified Smotifs).

Three different function-to-cluster association values were calculated as frequency [Eq 5.1], Log-odds [Eq 5.2] and Mutual Information [Eq 5.3].

$$F = \frac{k}{n} \qquad \textbf{Eq 5.1}$$

$$Logodd = \log\left(\frac{F}{K/N}\right) \qquad \textbf{Eq 5.2}$$

$$MI = \left(\frac{k}{N}\right) Logodd \qquad \textbf{Eq 5.3}$$

For each of these metrics, a significance threshold was devised by comparing them with the distribution of 500 random classifications, that is, against clusters created by randomly aggregate Smotifs of the same type. A p-value of obtaining an association score equal or better than a certain value was defined as the average of frequencies observed in the 500 random classifications.

The creation of functional enriched sequence patterns **fp** derived from the Smotifs clusters was performed as follows. For each cluster **k** enriched in a function **f**, an alignment was produced by selecting the Smotifs annotated with **f**. Afterwards, they were expanded by Swiss-Prot/HSSP [118] homologues annotated in the same function. Homologue sequences producing or having gaps in the alignment were removed; as well as any alignment with less than 10 sequences.

In order to perform transfer annotation, random sequences from each alignment were selected and used as queries for a BLAST [73] search over Swiss-Prot. For each putative homologue found, if it also matched any of the **fp** of the cluster from witch the query sequence was extracted, the homologue is assigned the function **f**.

The accuracy of the method was around 97% when assigning level 3 GO terms with a sequence identity of 60%, decreasing depending on the level of GO term (85% at level 5). At lower percentages of sequence identity, the method showed a decreased applicability but maintained its accuracy, while direct BLAST functional transfer did produce a high number of false positives.

5.2. Archer

Archer: Predicting protein function using local structural features. A helpful tool for protein redesign.

Jaume Bonet[1,+], Javier Garcia-Garcia[1,+], Joan Planas-Iglesias[1], Narcis Fernandez-Fuentes[1,2,*] and Baldo Oliva[1,*]

[1] Structural Bioinformatics Lab (GRIB-IMIM). Department of Experimental and Life Sciences. Universitat Pompeu Fabra, Barcelona, Catalonia (Spain).

[2] Institute of Biological, Environmental and Rural Sciences (IBERS) Aberystwyth University Aberystwyth, Ceredigion, UK.

[+] Both authors contributed equally.

[*] To whom correspondence should be addressed. Tel: +44 1970 621 680; Fax: +44 1970 622 350; Email: narcis.fernandez@gmail.com

Correspondence may also be addressed to Baldo Oliva. Tel: +34 933 160 509; Fax: +34 933 160 550; Email: baldo.oliva@upf.edu

a) Abstract

The advance of high-throughput sequencing methodologies has led to an exponential increase of new protein sequences, a large proportion of which remain unannotated. The gap between the number of known proteins and those with assigned function is increasing. In light of this situation, computational methods to predict the function of proteins have become a valid and necessary strategy. Here we present Archer, a server that exploits ArchDB's hierarchy of super-secondary structures to map GO and Enzyme functions upon protein regions and, thus, infer the function of a protein. The server relies on either the sequence or structure of the protein of interest and returns the mapping of functional subclasses extracted from ArchDB. Moreover, it computes the functional enrichment and significance of each subclass, combines the functional descriptors and predicts the function of the query-protein. Furthermore, users can select variants of the target sequence that swap the region of a super-secondary structure by another that putatively fits in the same scaffold. Only variants that modify the predicted function are offered for selection, thus providing a rational, knowledge-based, approach for protein design and functionalization. The Archer server is accessible at **http://sbi.imim.es/archer**.

b) Introduction

Whole genome sequencing projects have become a source of proteins whose function is unknown. Consequently, the annotation of protein function has become one of the most important challenges, particularly in computational-based methods [223]. Global sequence similarity has been extensively used to annotate protein function, considering the relationship between sequence and function similarity [224]. Despite the fact that homology annotation transfer is reliable for very high sequence similarity [225], its predictive power quickly diminishes when sequence homology falls below 70% [226].

One of the main problems of functional annotation based on homology is the functional specificity [225]. To overcome this limitation, some studies have proposed the functional annotation using protein-domains [227,228]. However, the transfer of functional information still faces several problems, as small protein changes can result in new protein functions [229,230]. This effect has been clearly shown in the enabling and disabling loops found in homo-domain interactions [176] or in the specificity of the RGD motif [231]. Following upon these findings, it seems a logical evolution to split the protein sequence into super-secondary structures, or local structural features, as they might be key to identify specific protein functions.

ArchDB [168] is a classification of super-secondary structures (loops henceforth) according to their geometrical properties [167]. It is well documented that loops play a central role in protein functions (from ATP binding [133] to enzyme activity [136] or DNA-binding [232] amongst others), and several methods have proposed to annotate protein function based on loops [145,172,233,234]. Moreover, we have recently shown their importance in protein-protein recognition [142,184].

In this work we present Archer, a novel web server designed to infer the function of proteins based on functional subclasses extracted from ArchDB.

The server allows to explore Gene Ontology (GO) molecular function [91] and the Enzyme (EC) database [90] functions enriched on the loops of a query protein. Then, it uses this enrichment to infer the function of the protein. Furthermore, the server offers variants of the query sequence with changes on the regions mapped by ArchDB subclasses yielding a new putative function. Each sequence fragment is substituted by the sequence of a subclass from ArchDB with similar geometry. This guarantees that the substitution will not change the main scaffold of the protein. Thus, each variant can be used in different aspects of computational protein design such as grafting of novel protein functionalities and/or redesign of the existing ones (see Section 1: Generation of protein variants with different function).

c) Methods

Functional association of super-secondary structures. Archer exploits the Markov Cluster (MCL) classification of ArchDB [168]. Four different metrics are used to evaluate the relation between a loop-subclass (C) from ArchDB and a function ϕ (either from GO or EC), extending the previous work of Espadaler et al. [6,105]: frequency (F), log-odd (Log-Odd), mutual information (MI) and the p-value of the hyper-geometric distribution of the enrichment of the function ϕ in the subclass C. A detailed explanation on how to calculate these metrics is described in Section 2: Metrics for functional association of super-secondary structures.

Assigning ArchDB subclasses to sequence. Given a protein sequence, the mapping of ArchDB subclasses is done by sequence homology search with BLAST [73]. The returned hits are filtered by the percentage of sequence identity as a function of the length of the aligned regions [51] (see Section 3: Assignation of ArchDB subclasses to a query protein (mapping)).

Assigning ArchDB subclasses to structure. For a given protein structure, all secondary structures contained in the protein are identified with DSSP [125]. All loops are then defined as two correlative regular secondary structures and the flexible region between them [167]. Each loop-region is then assigned to a subclass for which it fulfils its geometrical and structural constraints (see Section 3: Assignation of ArchDB subclasses to a query protein (mapping)).

Protein function prediction. For each subclass of ArchDB assigned to a protein, the four metrics of functional association and the number of subclasses supporting each function are used to predict the function of the query-protein. A 6D vector is built describing the association between the protein and a function ϕ, formed by the metrics specifically associated with that function: 1) S, the number of subclasses mapped in the protein and associated with function ϕ; 2) MaxF, maximum value of frequency (metric F) among the loops associated with ϕ mapped in the protein; 3) MaxLO, maximum value of log-odds (metric Log-Odd) among the loops mapped in the protein (associated with ϕ); 4) MaxMI, maximum value of mutual information (metric MI) among the loops associated with ϕ mapped in the protein; 5) MinPV, minimum p-value of the hyper-geometric distribution (metric p-value) among the loops associated with ϕ mapped in the protein; and 6) SUM, the sum of p-values of the hyper-geometric distribution of the function for all the loops mapped in the protein and associated with function ϕ. Then, the prediction is performed through a J48 trained pruned tree using WEKA [235] and the 6D vector that associates proteins and functions. See Section 4: Protein function prediction for more details.

d) Server Usage

Archer admits two types of input. The user can either upload the sequence (FASTA format) or the structure (atomic coordinates in standard Protein Data Bank format [169]) of the protein of interest. After the query has been submitted, a window will display the query and a unique code assigned to it. This code can be used to retrieve the results of that particular query from the main page of the server. The results page consists of a summary section and three tabs: subclass mapping, function prediction and sequence variants. It also allows downloading all the data for further study.

Subclass mapping. A tab lists all subclasses from ArchDB mapped in the query. For each subclass, the user can select the mapped region in the query-sequence and explore the GO and Enzyme annotations associated with the subclass. Furthermore, heteroatom contacts and functional PDB-sites assigned to particular loops within the subclass are also listed.

Function Prediction. This tab displays the results of function prediction for the query protein, highlighting the regions of the query supporting the prediction.

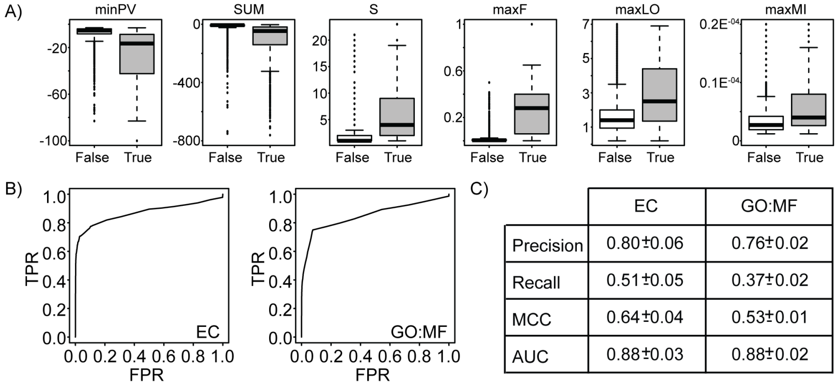
Sequence Variants. This tab displays alternative sequences (variants) of the query protein. Each variant is built by substituting at least one loop-subclass from the original sequence by an ArchDB mapped region with similar geometry. Variants are grouped according to the changes predicted on the original function, such that only variants yielding a different prediction are displayed. Deleterious variants, i.e. variants that simply remove the predicted function of the query protein, and invariants, i.e. variants yielding the same prediction as the original query-sequence, are not shown. The new putative sequence can be selected for display and, if the structure of the query protein is known, also the superposition with the loop-conformation of the proposed new sequence can be seen (or downloaded for further use in modelling).

e) Evaluation

Benchmark Dataset. Protein function prediction was evaluated in two different datasets: one for the prediction of Enzyme annotations (EC) and the other for GO Molecular Function terms (GO:MF). Both datasets were derived from all human proteins in Uniprot Swiss-Prot [10] (the largest curator annotated available set) with EC or GO:MF assigned, respectively. We used CD-HIT [50] to remove proteins with more than 40% sequence identity within the dataset in order to avoid the bias of highly populated homolog family members. The resulting datasets were named hEC (containing 2,054 proteins) and hGO (consisting of 9,305 proteins) for EC and GO:MF annotations, respectively. Additionally, we also used CD-HIT to remove sequences with more than 40% of sequence identity between these two sets and any of the sequences used in the construction of the ArchDB dataset. This avoided the homology between the training and the testing sets (see below). The resulting datasets were named hEC40 (containing 1,357 proteins) and hGO40 (consisting of 7.529 proteins).

Training. ArchDB subclasses were assigned to proteins as described in Methods. We were able to assign ArchDB subclasses to 87% of proteins in the EC dataset and 70% in the GO:MF dataset. We considered for each subclass all functions enriched with a p-value lower than 10^{-3} if, at least, two members of the subclass were associated to that function. We were able to predict at least one function for 98% of the proteins in the EC dataset and 99% in GO:MF. We trained two J48 pruned decision trees (one for EC and another for GO:MF datasets) using WEKA [235] (see Methods). All metrics used in the 6D training vector showed significantly different distributions between correct (true) and false function associations in the EC and GO:MF datasets (see Figure 5.1A). For example, from metric S in Figure 5.1A, most proteins had about 5 or more loops correctly assigned to the enzyme function, while the majority of them had less than two loops associated with an EC code different than the real (wrong association).

Evaluation. Decision trees in WEKA were trained and tested using a ten-fold cross-validation for each dataset (hEC40 and hGO40 with mapped loops in ArchDB). The association of a protein to a function was considered a true positive if the function is actually associated with the protein. It was considered a false positive otherwise. Figure 5.1B shows the ROC curves and the quality of the predictions of the server for EC and GO:MF datasets. Precision, recall, Mathew Correlation Coefficient and AUC are shown in the table within Figure 5.1C. Remarkably, the precision of Enzyme annotation at the third level of classification reached values of almost 80% while the recall was still about 50%, which means that we could apply a reliable putative redesign of function to 50% of enzymes.



	EC	GO:MF
Precision	0.80±0.06	0.76±0.02
Recall	0.51±0.05	0.37±0.02
MCC	0.64±0.04	0.53±0.01
AUC	0.88±0.03	0.88±0.02

Figure 5.1 Archer attributes and predictor analysis. A. Distribution of different metrics in the EC dataset for the 6D-vector classifier: Distributions of false positive (FP, white boxplots) and true positives (TP, grey boxplots) in the training sets for mono-dimensional vector metrics (S, MaxF, MaxLO, MaxMI, MinPV and SUM, as described in Methods). B. ROC curve of the average TPR versus FPR obtained with the 10-fold validation using classifiers trained with WEKA for EC dataset (left) and GO:MF dataset (right). C. Table showing the average and standard deviation of statistic measurements in the 10-fold validation: MCC: Mathews Correlation Coefficient, AUC: Area Under the Curve, recall and precision

Comparison. We compared our results with sequence-based annotation methods, such as Best-BLAST [103], BLAST2GO [236] and BLAST2GO+InterPro [117]. We used the datasets hEC and hGO to test the prediction of function. Functions were searched in different pools of proteins with known function: for Best-BLAST we used UniProt, for BLAST2GO and BLAST2GO+InterPro we used the dataset of all non-redundant sequences (NR) from NCBI. We tested Archer with the sets hEC, hGO, hEC40 and hGO40. We tested hEC and hGO against SwissProt with Best-BLAST and against NR with BLAST2GO and BLAST2GO+InterProt. For the last three methods, we tested different levels of redundancy between the test set and the template set (from 100 to 40; see Section 5: Comparison with other methods). Our results show that Best-BLAST, BLAST2GO and

BLAST2GO+InterPro have a good performance if there are close homologs (sequences with more than 70% sequence identity) in the pool of proteins with associated functions, but PPV and Recall decrease dramatically otherwise, specially in GO:MF term prediction. Conversely, Archer's classifier is almost unaffected by the reduction of putative homologs (hEC versus hEC40, and hGO versus hGO40).

f) Conclusions

In this work we have presented a server to predict protein functions through the identification of structural features. The server is able to maintain a high degree of accuracy on the prediction regardless of the percentage of sequence identity of the query-protein with any of the proteins used in the server for the annotation. Archer provides a simple user interface and a comprehensive results-page. The results include the mapping of function-associated annotations to the query by means of structural similarity, based on super-secondary structures. All the results can be traced back to the original databases, thus providing means to understand how the prediction was obtained. Additionally, it highlights the regions that might be implicated in the protein function, which allows the users to better comprehend the results and devise new experiments. In particular, Archer offers sequence variants of the original query protein that could have a different function. On the one hand, each variant should be able to maintain the original scaffold of the query protein, as the substitution is selected among super-secondary structures fitting on the geometry of the flanking original secondary structures (as in Frag'r'Us [43]). On the other hand, the new fragment is associated with a different function and, within the margins of accuracy demonstrated for the server, it modifies the function of the protein. Therefore, the server helps to redesign the query-protein and change its function. Furthermore, when the structure of the query protein is known, the structural information on the putative variants can also be retrieved, which helps to ensure the structural fitting of the new fragment(s) and it suggests a template for modelling. We believe that Archer can become a very useful tool for the scientific community, predicting function for proteins with reasonable accuracy. Furthermore, it is, to our knowledge one of the first services to provide a rational approach to derive new functionalities and/or redesign existing ones by pinpointing suitable regions on the query protein and the potential candidate sequences to be grafted, having a background reliability based on the original capability of protein function predictor.

g) Funding

Spanish Ministry of Science and Innovation (MICINN) [FEDER BIO2008-0205, FEDER BIO2011-22568, EUI2009-04018]. Funding for open access charge: Spanish Ministry of Science and Innovation.

h) Supplementary Material

Section 1: Generation of protein variants with different function

Archer generates possible protein variants from a protein query by applying changes to the regions of super-secondary structures assigned to a subclass in ArchDB. The proposed new sequence belongs to some subclass in ArchDB with similar geometry of the flanking secondary structures. Thus, the proposed protein variant should not distort the backbone conformation of the original query protein. Only the variants that predict a different function from the original query (according to Archer's predictor) are presented to the user. The confidence of that prediction is the same as for the test (see Section 4: Protein function prediction). The creation of the variants follows a series of steps that are applied for each region mapped with a subclass from ArchDB:

Search of similar subclasses: Two subclasses are considered similar if their geometrical properties are within certain degrees of similitude. The allowed variability is ±1.5 Angstroms for the distance between the borders of the two flanking secondary structures of an arch, ±6 degrees for the theta/delta angles and ±15 degrees for the rho angle.

Selection of representatives for each new function: If several subclasses are found, all the enriched functions of all the subclasses are listed, and the better-scored subclass for each function (lowest p-value) is selected as representative for each function. The best-scored arch inside the subclass, from a protein annotated with the selected function, is selected to substitute the original query-sequence.

All possible changes are explored for all protein regions with mapped subclasses. Then, all combinations are tested with Archer. Those combinations yielding a prediction of function different from the original are selected and grouped according to the new set of functions.

Sequence variants are displayed to the user clustered by the putative new function and ordered accordingly to the number of regions that need to be changed. If the user provides the structure of the query protein, the coordinates of the variant loop are superimposed over it.

Section 2: Metrics for functional association of super-secondary structures

Four different metrics are used to evaluate the statistical relation between a function and a subclass form ArchDB. They are all based on the following parameters:

N: Total number of loops in ArchDB. Population.

K: Total number of loops in ArchDB with a given function. *Success in population.*

n: Total number of loops in a subclass. Sample.

k: Total number of loops in a subclass with a given function. *Success in sample.*

The first and easiest metric we can develop is the raw frequency, based on the idea that the relationship between a function and a subclass is proportional to its frequency among the loops of the subclass:

$$F = \frac{k}{n} \qquad \textbf{Eq 5.4}$$

We can further extend the relationship by correcting the success of a function in a subclass according to its success in the full population. Those two metrics are log-odds:

$$Logodd = \log\left(\frac{F}{K/N}\right) \qquad \textbf{Eq 5.5}$$

And the hyper-geometric p-value distribution, which measures the probability of obtaining by chance a subclass at least as enriched in a given structure:

$$p - value = \sum_{i=k}^{i=n} \frac{\binom{n}{i}\binom{N-n}{K-i}}{\binom{N}{K}} \qquad \textbf{Eq 5.6}$$

Finally, a method derived from Mutual Information is also used to evaluate the information content of a given function in a subclass:

$$MI = \binom{k}{N} Logodd \qquad \textbf{Eq 5.7}$$

The combination of these metrics form the 6D vector to feed the WEKA [235] classifier and predict the protein function. Section 4: Protein function prediction focuses on that issue.

Section 3: Assignation of ArchDB subclasses to a query protein (mapping)

To a query protein sequence:

We mapped ArchDB subclasses [168] to protein sequences extending the process described in a previous work [142]. Succinctly, the process consists on searching homologues for the query protein amongst proteins with classified loops in ArchDB using BLAST (2.2.28+ version). ArchDB subclass annotation is transferred if three conditions are met:

a) The hits satisfy a minimum percentage of identity according to the length of the alignment (above the twilight-zone curve, as described by Rost [51]).

b) There are no gaps in the flanking secondary structures of the region. Only the loop region can contain up to 2 residue-gaps as long as they do not correspond with the first or last amino acid of the aperiodic region.

c) Once a subclass has been assigned in a region of the query protein, the rest of BLAST hits in that region are neglected.

To a query protein structure:

Given the structure of a query protein, all loops are calculated using the same algorithm as in ArchDB [168]. This is, the secondary structure is calculated with DSSP [125] and the vectors defining the geometry between the flanking secondary structures are calculated (distance, hoist angle, packing angle and meridian angle).

For the substitution of a sequence fragment of the query protein in order to predict a new function (see Section 1: Generation of protein variants with different function) all subclasses with similar geometry are assigned and all putative new sequences too.

Section 4: Protein function prediction

Evaluation measures

A prediction consists on a protein-function association pair. We used the accuracy metrics as defined by Jones et al. [103], in which they modified the definitions of precision and recall measures to assess the accuracy of pair-association predictions. PPV is defined as the proportion of correct predicted pairs (true positive) over the total number of predicted pairs (positives). Recall is defined as the proportion of correctly predicted pairs over the total number of correct pairs (i.e., sensitivity of the method). For the prediction of GO Molecular Function terms, different terms can be related and/or highly similar. However, we followed Jones et al. and considered a correct prediction (true positive) only when there is an exact match (i.e. similar, but not equal, functional terms are considered different).

Pruned decision trees

The metrics described in section 3 showed significantly different distributions between correct (true) and incorrect (false) function assignments [Figure 5.1]. Then, we trained a decision tree to decipher the best combination allowing us to separate between correct and incorrect assignments. First, a 6D vector was built to describe the association pair of a protein and a function as in the main text. This forms a bijective application between a 6D vector and a pair (protein-function) association. Next, we used a J48 decision tree, a variant of a C4.5 decision tree [237] implemented in WEKA (version 3.7.10) [235] to predict the protein-function pair with a 6D vector. We used 0.25 as confidence factor used for pruning and 2 as the minimum number of instances per leaf. All options were set to default values. The sets hEC40 and hGO40 (see definition in the main text) were used to train and test WEKA using a 10-fold cross-validation. For each protein in the dataset, we considered all functions (EC or GO:MF, respectively) enriched with their mapped subclasses having a p-value lower than 10^{-3} if at least two members of the subclass were associated to that function. Next, the set was split in 10 non-overlapping groups and all pairs in one group were tested using the decision tree constructed with the remaining nine, repeating the procedure up to ten times.

Section 5: Comparison with other methods

In order to benchmark our method against the current state of the art, we compared the results of EC/GO:MF term prediction with sequence-based annotation methods, such as the Best-BLAST method [103] and BLAST2GO [236]. We selected a random set of 400 sequences from hEC and 900 from hGO to do the comparison.

Best-BLAST uses UniProt as its pool of protein-templates with known function. The method assigns the functional terms of the best matching protein sequence returned by BLAST.

Similarly, BLAST2GO uses BLAST to search homologs to a query protein in a pool of non-redundant protein-templates from NCBI(NR). BLAST2GO uses different methods to transfer the functional annotation, and it includes the possibility to use InterPro annotation [236] to improve the prediction (BLAST2GO+InterPro).

We define the "template sequence dataset" as the pool of protein sequences with know function that we use as a template to transfer their annotation to a protein query. The template sequence datasets used in each method are:

i. ArchDB database for Archer;
ii. UP: Uniprot proteins with functional annotation for Best-BLAST;
iii. NR: non-redundant NCBI proteins for BLAST2GO.

We tested the effect of redundancy between the different benchmark datasets of query proteins and the template sequence datasets. For Archer, we benchmarked the protein sequences from hEC40 and hGO40. For Best-BLAST and BLAST2GO we removed the template sequences obtained with BLAST having more than 80%, 60% and 40% of sequence identity with their corresponding query sequence of the benchmark. Thus, we evaluated the behaviour of the different methods at different levels of sequence homology between the queries of the benchmark and the template set. A brief summary of the results and a distribution of the results are shown in Figure 5.2. Although for the study of enzyme function our method is surpassed by both Best-BLAST and BLAST2GO, when studying the prediction of GO:MF terms our approach outperforms all methods until the pool of templates contains close homologs (with percentage of identity largest than 70%).

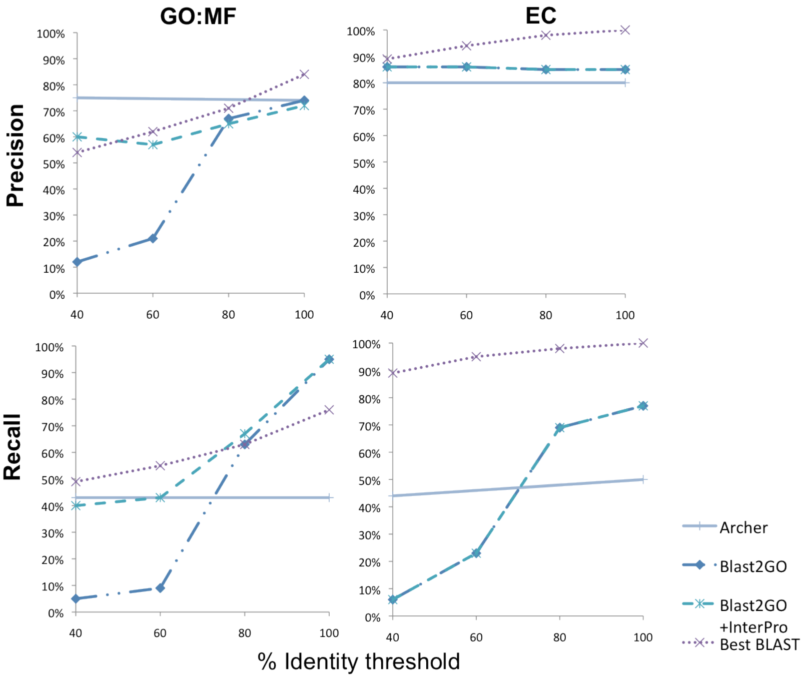


Figure 5.2 Comparison of the curves of Precision and Recall with four different methods: 1) Archer; 2) Blast2GO (executed with default options); 3) Blast2GO+InterPro (executing Blast2GO integrated with InterProScan annotation); and 4) Best BLAST. Precision and recall are calculated with the corresponding template database of sequences for each method and removing homologs from the benchmark. The threshold on percentage of identity shown in the X axis identifies the maximum percentage of identity between the sequences of the benchmark and the sequences from the template database used to calculate precision and recall. Recall with the method "Archer" is calculated with sequences of the benchmark for which the method can be applied (i.e. 82% of the EC and 19% of the GO:MF benchmarks).

5.3. Applications of Archer

We have previously discussed the successes to alter the functional specificity of a protein (protein re-design). Most published examples, though, are very specific and focus on particular cases [207,208]. Archer has the potential to help, in a generic way; making decisions regarding the nature of sequence and conformational changes towards a new selected function.

As a function predictor, its main benefit relies on its apparent lack of dependence in sequence homology. As we have seen, Archer's performance is maintained regardless of the homology with the test set. Thus, a combination of global homology search (Best-BLAST approach) and Archer, prioritizing one or the other according to the similarity of the query with other proteins with known function [Figure 5.2] should be a more reliable predictor than each individual method. Furthermore, the full traceability of the annotation transfer facilitates a deeper understanding of the prediction process.

As a method for loop re-design, we believe it covers a *niche* not already exploited. By itself, there is a clear benefit in the fact that it can highlight the regions of the protein that are statistically supporting a given function. It is able to offer a limited, thus experimentally manageable, set of putative mutations for changing the function without lost of structures ranked by a confidence reliability score. Furthermore, this limited set of changes can help to reduce the experimental requirements in order to develop designed proteins both in time and cost.

6. DISCUSSION

6.1. Working with Smotifs

This thesis presents a series of resources and methods to exploit protein fragments from a computational point of view. In the Introduction, I have discussed the growing gap between identified proteins and proteins with know structure and function, the limitations of the experimental approximations in closing that gap at the required pace, and the basis that sustain the computational approaches to tackle that same problem. Finally, I have introduced our own approximation to the problem.

I have started by defining our working unit, the Smotif, its properties and its main descriptors (see section Defining Smotifs). This definition is immediately followed by a description of the mechanisms devised to exploit these descriptors as means to clusterize the Smotifs according to their similarity.

By applying two different clustering protocols, the Density Search (DS) and the Markov Clustering (MCL), we were able to create the two apparently independent classification trees that conform ArchDB. While the DS algorithm provided us with a length-dependent classification, the MCL clustering produced a length free classification that allowed us a much better coverage of the Smotif space (that is, a higher percentage of classified Smotif against non classified ones). As a matter of fact, by mapping this version of ArchDB backwards in time (starting from 1972), Figure 6.1 shows an increasing tendency over the years to progressively classify a larger percentage of all the available Smotifs at that particular distribution of PDB, as if the increase of known structures confirmed that the total number of possible Smotif conformations is limited. This has a clear impact in the improvement of predictions in knowledge-based methods.

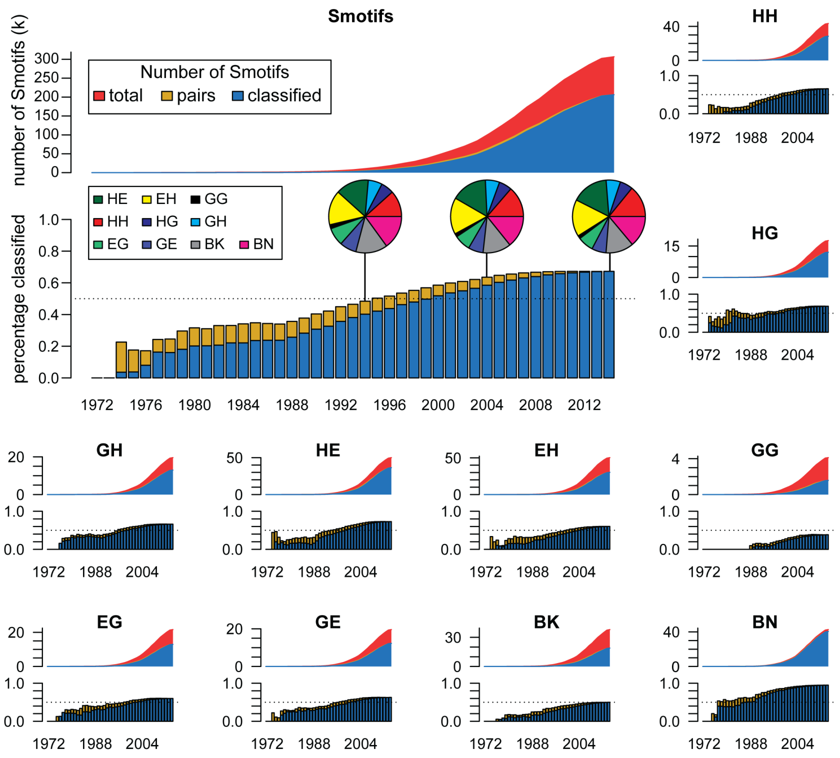


Figure 6.1 Mapping of Smotifs in the PDB over the years. Data is shown for each Smotif type and for all the Smotifs (top left double graphic). For each graph, the top section represents the total number of Smotifs (in thousands) mapped over PDB in that particular year (red), the number of those belonging to subclasses (blue) and the number of those that have at least one geometrically similar Smotif. The bottom section represents the percentage of Smotifs clustered (red) or with a similar one (green). The pie charts represent the percentage of the different Smotifs, according to their flanking secondary structures, found in 1994, 2004 and 2014.

After revising ArchDB, I have focused upon the assisted modelling of loop structures and the suggestion of alternative backbone conformations for loops that are compliant with their flanking secondary structure.

Frag'r'Us is presented as a service able to provide this sampling of loop backbone conformations using the Smotifs definitions. The web application has been developed to generate a comprehensive and controlled set of realistic conformations to be used as a starting point for further manipulation and optimization using protein design techniques. To further demonstrate that, several examples in which both the original structure and its computational counterpart are known were selected and executed with Frag'r'Us. Our results clearly demonstrate the ability of the algorithm to closely mimic the results obtained by those methods.

Finally, I tackled the subject of functional annotation transfer and the putative advantage of a structural-based fragment method that, not only localizes putative regions related to the function, but offers fragment substitutions that could affect the protein's function while minimizing their effect upon the global secondary structure of the protein.

With this, we have provided a service able to provide a controlled set of alternative conformations, which implies the possibility of study each one of these putative replacements in detail to alter the protein towards different interests, from the protein's stability to its function. Of course, we have tried to automatize that last process through Archer.

By exploiting statistical correlations between EC and GO annotations and Smotifs clusters in ArchDB, Archer was develop as means to perform the above described tasks. In its own level of specificity and applicability, which can be increased by combining its predictions with those of Best-BLAST in high homology annotation transfer, Archer is not only able to predict the putative function of a protein but also to suggest fragment substitutions and evaluate their possible functional effect, both regarding the original predicted function and the arisen of new functionalities. Furthermore, it offers a full traceability of the origin of the assignations and the decision process to reach them; this property is key if it is to help researchers devise new designed sequences.

6.2. Developing Software and Web Services

Computational biology should be about creating knowledge, grasp a new understanding able to adequately explain both old and new experimental evidences to help move forward basic research (both experimental or from other computational groups) towards medical, technological or even industrial applications. Thus, it is imperative to devote part of our time towards the design of sustainable and user friendly software that should not require of our continued presence for others to use. At the end, that includes everything from well-described pipelines, to graphical user's interfaces and web applications and services. There are several problems related to those requirements.

Firstly, computational publications are considered in a similar manner as experimental ones in terms of reproducibility. That should not be so. It is obvious that, with experimental research, a code of honour was required between the researchers and the journals. One can describe experimental methods, show their results, even share the protocols, but there is no way to pack everything required to make a 100% reproducible article. That opportunity is available to computational research. Logically, researchers in the field are becoming increasingly aware of that fact, proposing descriptive labels in computational articles regarding the level of openness of their code and source data [238], as well as methods on how that data should be presented, structured and prepared [239,240]. Despite this seed of awareness, I am yet to find a journal, even a specialized one, to request for the code and the data to be available or that includes a system to share that type of data other that a compressed supplementary material file. It is my opinion that journals should lead changes regarding this particular issue by (a) requesting the availability of the code and the data, (b) making available systems for code distribution and (c) ensuring a way to cite that code when others are using it, even if it is for different applications. This last point would be vital in order to encourage the researchers to post their full data.

Secondly, web services are not easy to maintain. Developers of a particular service might move to other labs, and the ability of the hosting lab to maintain the server will be directly proportional to its level of documentation. And that is assuming the lab either has a technician devoted to such tasks or a new researcher that wants to take over the project. Otherwise there is a risk of not even trying. Schultheiss *et al.* [241] studied this phenomena checking, in 2011, the availability and functionality of all the web services published in the special server issue of the Nucleic Acids Research (NAR) journal. Although they found a continuous increase in the maintenance of the servers (newer servers were working better than older ones), which might be simply due to their chronological proximity to the review, and a clear increase in some specific categories such as example data or online help (which are nowadays requested by NAR as part of the features of the server), only 54% of the servers from 2009 were considered fully operational. Sadly, an increase of the budget in science, that can be devoted to hire technical personnel, is the only real solution I can think of to tackle this particular issue.

The last problem is the seemingly lack of interest from experimental labs over the resources available. With the exception of databases, web services and applications are mostly cited by other computational researchers either prising or simply comparing against them, or by research articles performed either by people form the same group or by experimental articles with computational collaborators. In my opinion, this has to do with the virtual fragmentation created between experimentalist and computational researchers. Although using different techniques, we kind of move towards the same objectives. But, as a rule, we are not only physically separated in the lab, but also in our publishing ecosystem. Except for some wide range journals like PLoS ONE, most journals are clearly devoted to one field or the other. This is why initiatives like the Database issue and Server issue of NAR are such good ideas. Of course, as long as the number of mixed collaborations keeps growing, at least we know that our developments and expertise are going to have a final impact in the experimental field.

6.3. Future Perspectives

a) Prediction of Drug Targets

As previous works have shown [109], drug targeting can be described by a series or protein fragments that do not need to be interconnected between them in the same manner (that is, they are separated fragments that fall similarly together in the 3D space regardless of the fragments joining them).

Internally, ArchDB contains cross-references with PDBTM [242], a database of transmembrane proteins. Furthermore, it also contains scored correlations between Smotifs clusters and DrugBank, a correlation that could be extended by further categorize the drugs through databases such as FragmentStore [243].

- Thus, the combination of drug assignments and transmembrane data already contained in the database could be extended and exploited to identify drug targets in the cell's membrane, which could be useful for the analysis of drug physiological activities.

b) Protein-Protein Interactions

Protein-protein interaction (PPI) is one type of protein function that spans through every biological process of the cell. As we commented in the Introduction, through their interactions proteins achieve their quaternary structure. Furthermore, by creating both transient (temporal) and permanent physical connections between them, proteins conform the PPI network [12].

Thus, extending the study towards PPIs seems like the natural thing to do. On the one hand, PPIs are a more specialized functional descriptor; which means that progressing towards its study will represent a more detailed view over the protein function. On the other hand, it is because of PPIs that proteins acquire their quaternary structure and perform biological processes. As, up to now, we have been focusing in tertiary structure and molecular functions, developing methods to explore PPI through protein fragments will actually broaden the field of applications of Smotifs.

Protein Design

In a previous work by Planas-Iglesias *et al.* (Appendix 8.4) [142] we showed the possibility to exploit ArchDB's classification in order to predict non-biological PPIs (i.e. PPIs that can happen given the chance for both proteins to meet –like in an *in vitro* assay, but do not have to actually happen in a living cell). Using both statistically correlated interacting and non-interacting pairs of subclasses from ArchDB interacting protein pairs could be predicted. As the location in which the subclasses were mapped on the proteins were irrelevant, we hypothesised that our results could be explained by the funnel-like intermolecular energy landscape theory [244]. The theory developed in that work yielded in the development of the iLoops server (Appendix 8.3) [184], thus proving the applicability of the method.

One possible expansion of this method should be the ability to predict changes that will hinder the ability of two proteins to interact. As the interface is not identified, changes might not always relate to a full lose of the interaction but they might affect the efficiency of the partner recognition; as a matter of fact very recent approximation to that same objective have started to appear [245].

- Thus, by combining **iLoops** and **Archer**, we should be able to develop a method to suggest local conformational changes that might affect a PPI without affecting the molecular function of the protein or the other way around. In other words, we could develop a method to suggest protein redesign affecting at both molecular function and/or biological process levels.

Protein Interface Prediction

Proteins interact between themselves through a limited set of interface types. Therefore, these interfaces are key to determine the protein function.

Due to its relevance, the computational identification of protein-protein interacting interfaces is important for a more accurate characterization of the function of proteins and their complexes [246]. This can also allow the prediction of new unknown protein-protein interactions [247] and putative drug targets [248] amongst other applications.

In a previous work by Aragues *et al.* (Appendix 8.12) [249] a protocol for describing interactive motifs (**iMotif**) was developed. The article was focused on the characterization of protein hubs (i.e. highly connected proteins in the PPI network) and their putative number of interfaces relying exclusively in their sequence and connectivity information. Even by limiting the study to that information, the method was able to strongly correlate its findings in a similar study performed with structural information [250]. Although sequence motifs were generated with PRATT [251], they were exclusively used to merge detected iMotifs after the first step of the procedure, but the applicability of the detected regions as means for PPI prediction were not tested.

The protocol is based on the assumption that proteins with overlapping sets of interacting partners would tend to interact with such partners through the same protein region (interacting motif). To do so, the method relies in the analysis of protein-protein interaction networks, in which each node represents a protein and each bidirectional edge the interaction between a pair of proteins.

Briefly, the clustering algorithm (iM clustering) can be summarized in two steps [Figure 6.2]. Firstly, for each protein of the network an interaction cluster is assigned. Each cluster is linked to the clusters of the interacting protein. Thus, the initial cluster interaction network is created with the same topology of the protein interaction network. Once this initial cluster network is created, new clusters are created iteratively by merging similar clusters until their similarity score drops under a predetermined threshold. The similarity score between clusters is defined as the number of common interacting partners that two clusters share in the cluster interaction network. At this point, a protein can be found in more than one cluster.

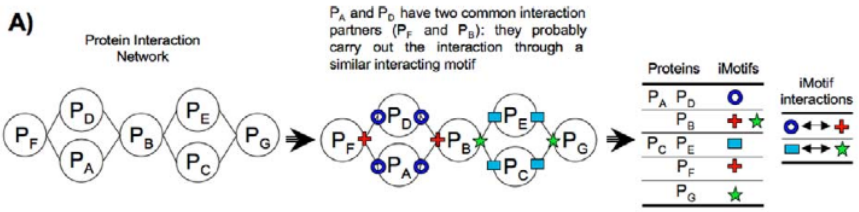

A)
Protein Interaction Network
P_A and P_D have two common interaction partners (P_F and P_B): they probably carry out the interaction through a similar interacting motif
Proteins
iMotifs
iMotif interactions
P_A P_D
P_B
P_C P_E
P_F
P_G

Figure 6.2 iMotif clustering protocol. Image extracted from [249].

Due to the ability of the method to limit putative interacting regions between pairs of proteins, it is worth to explore the possibilities of the iMotif protocol combined with Smotifs and sequence pattern generators to obtain interface-related sequence motifs. This can be applied for both globular and intrinsically unstructured (IUPs) protein regions. Thus, we believe the method can be exploited to:

- Prediction of interfaces over structured regions, by mapping **Smotifs** as their representative sequence patterns. By combining this with the pattern scoring system of **iMotifs** and the interaction correlation in **iLoops**, we should be able to better detect the interacting surface.
- Prediction of interfaces in intrinsically unstructured proteins (IUPs), through generation of sequence motifs amongst clustered sets of proteins and the improvement of the scoring filters to remove spurious motifs.

As a matter of fact, a preliminary pilot study has been initiated towards these aims.

iMotifs Applicability: Preliminary Study

Due to its potential, it is worth to explore the possibilities of the iMotif protocol as means of detecting PPI interfaces. A preliminary pipeline [Figure 6.3] has been developed in which sequence patterns are extracted from sets of proteins belonging to a same iMotif. Those patterns are then scored according to their presence in the background database in order to select those that more reliably represent the protein interface.

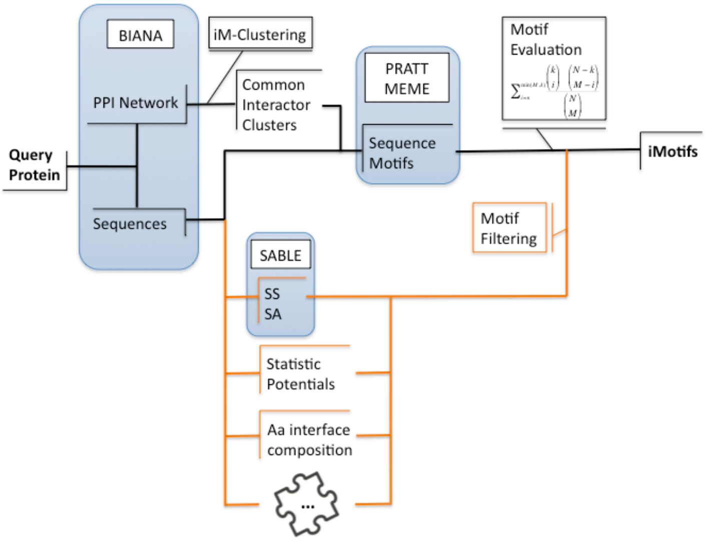


Figure 6.3 Preliminary iMotif interface prediction protocol.

This preliminary pipeline (specifically the black line section [Figure 6.3]) has already been tested against the *C. elegans* fragmentome [252]. Briefly, the fragmentome is a PPI network in which interacting proteins have been spliced and the interactions re-tested. This way, they have been able to define the minimal interacting regions (MIR) required to perform the alignment [Figure 6.4]. Sadly, almost half of the MIRs cover around 90% of the protein sequence [Figure 6.4B]. This is probably related to the folding requirements to perform the interaction.

In any case, by applying the iMotif pattern generation pipeline over those proteins with MIR smaller than 20% of the protein, the method was able to improve prediction of fragments inside the MIR region over the random selection of protein fragments. The improvement was directly related with the number of common interactors shared by the proteins in a given iMotif cluster [Figure 6.5].

One of the main drawbacks of the method, as now implemented, is the generation of noisy predictions; there are several improvements that can be implemented.

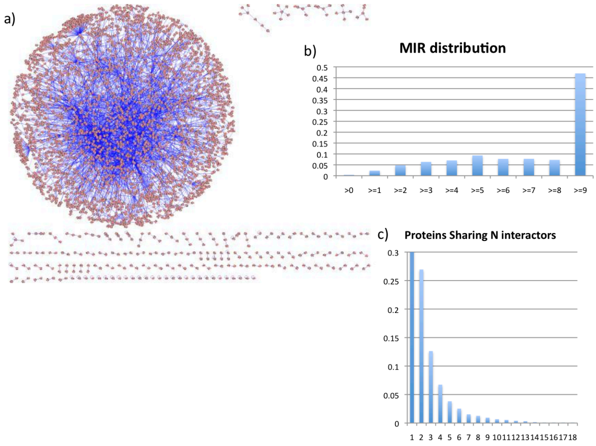


Figure 6.4 The fragmentome database. a) is a representation of the fragmentome database, b) shows the distribution of protein coverage of the MIR over the proteins in the dataset, c) represents the accumulative number of interactors for each node of the network (frequency of 1 interactor reaches 1 –not shown).

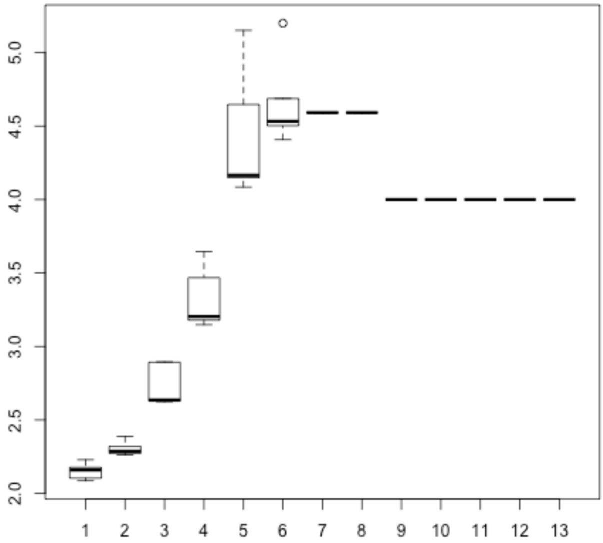


Figure 6.5 iMotif prediction improvement over random. The x-axis represents the number of shared interactors between proteins in an iMotif. The y-axis represents the improvement of the prediction method over a random selection of protein fragments.

7. CONCLUSIONS

In this work, we have studied the properties of structural fragments and its application into completing gaps in these known data through the use of knowledge-base extrapolation of the data. Specifically, we have been able to:

i. Update the classification of protein fragments and apply a new clustering algorithm to allow a higher range of backbone flexibility in the loop region while maintaining its descriptive properties.

ii. Assess the functional correlation of clustered sets of geometrically similar protein fragments. Including the addition of new functional databases and the definition of a new correlation score.

iii. Develop a method for the prediction of loop conformation in incomplete protein structures through geometrical constraints.

iv. Develop a system to transfer protein function annotation by fragment homology search and highlight the regions more probably related to the assigned function.

v. Develop a recommendation algorithm for protein redesign able to propose structurally conservative sequence substitutions that can affect the function of the protein.

vi. Make each single improvement presented in this work available through a collection of web interfaces.

8. APPENDIX

In addition to the work and the articles presented as the main body of this thesis, during my time in the Structural BioInformatics Lab I have participated in several other pieces of research in collaboration either with other members of the lab or with external experimental groups. With some of them, I have actually built a continued collaboration that have yielded in multiple projects published together:

1. **Molecular Immunology Unit.** *Hospital Universitario Puerta de Hierro, Madrid (Spain).* We have already published 2 research articles and 1 review, and we have, still, 1 project with them.
2. **Chromatin and Gene Expression Group.** *Centre de Regulació Genòmica (CRG), Barcelona (Spain).* We have 1 published and 1 submitted article with them, and we have 1 project standing.
3. **Protein Folding and Stability Group.** *Universitat Autònoma de Barcelona; Barcelona (Spain).* With whom we have published 2 research articles.

Most of these collaborations have been centred in the modelling of newly design antibodies (Appendix 8.5, Appendix 8.10, Appendix 8.11), the effect of protein elongation in its stability (Appendix 8.2), the effect of post-transcriptional modifications in their function efficiency (Appendix 8.6) or the transference of protein functionality through domain structural similarity (Appendix 8.9).

8.1. On the Use of Knowledge-Based Potentials for the Evaluation of Models of Protein-Protein, Protein-DNA, and Protein-RNA Interactions.

Fornes O, Garcia-Garcia J, Bonet J, Oliva B. On the Use of Knowledge-Based Potentials for the Evaluation of Models of Protein-Protein, Protein-DNA, and Protein-RNA Interactions. Adv Protein Chem Struct Biol. 2014;94:77-120. doi: 10.1016/B978-0-12-800168-4.00004-4.

8.2. Elongation of the C-terminal domain of an anti-amyloid β single-chain variable fragment increases its thermodynamic stability and decreases its aggregation tendency.

Rivera-Hernández G, Marin-Argany M, Blasco-Moreno B, Bonet J, Oliva B, Villegas S. Elongation of the C-terminal domain of an anti-amyloid β single-chain variable fragment increases its thermodynamic stability and decreases its aggregation tendency. MAbs. 2013 Sep-Oct;5(5):678-89. doi: 10.4161/mabs.25382.

8.3. iLoops: a protein-protein interaction prediction server based on structural features.

Planas-Iglesias J, Marin-Lopez MA, Bonet J, Garcia-Garcia J, Oliva B. iLoops: a protein-protein interaction prediction server based on structural features. Bioinformatics. 2013 Sep 15;29(18):2360-2. doi: 10.1093/bioinformatics/btt401.

8.4. Understanding Protein-Protein Interactions Using Local Structural Features.

Planas-Iglesias J, Bonet J, García-García J, Marín-López MA, Feliu E, Oliva B. Understanding protein-protein interactions using local structural features. J Mol Biol. 2013 Apr 12;425(7):1210-24. doi: 10.1016/j.jmb.2013.01.014

8.5. Generation and characterization of monospecific and bispecific hexavalent trimerbodies.

Blanco-Toribio A, Sainz-Pastor N, Álvarez-Cienfuegos A, Merino N, Cuesta ÁM, Sánchez-Martín D, Bonet J, Santos-Valle P, Sanz L, Oliva B, Blanco FJ, Álvarez-Vallina L. Generation and characterization of monospecific and bispecific hexavalent trimerbodies. MAbs. 2013 Jan-Feb;5(1):70-9.
doi: 10.4161/mabs.22698.

8.6. CDK2-dependent activation of PARP-1 is required for hormonal gene regulation in breast cancer cells.

Wright RH, Castellano G, Bonet J, Le Dily F, Font-Mateu J, Ballaré C, Nacht AS, Soronellas D, Oliva B, Beato M. CDK2-dependent activation of PARP-1 is required for hormonal gene regulation in breast cancer cells. Genes Dev. 2012 Sep 1;26(17):1972-83. doi: 10.1101/gad.193193.112.

8.7. Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence.

Baldo Oliva, Joan Planas-Iglesias, Jaume Bonet, Manuel A. Marín-López, Elisenda F, Gursoy A. Structural bioinformatics of proteins: predicting the tertiary andquaternary structure of proteins from sequence. Dins: Protein-protein interactions : computational and experimental tools. Cai W, Hong H (ed.). ISBN: 978-953-51-0397-4. InTech, 2012. DOI: 10.5772/2679.

8.8. Networks of Protein-Protein Interactions:

García-García J, Bonet J, Guney E, Fornes O, Planas-Iglesias J, Oliva B. Networks of protein-protein interactions : from uncertainty to molecular details. Mol inform. 2012; 31(5):342-62. DOI: 10.1002/minf.201200005

8.9. Prediction of a new class of RNA recognition motif.

Cerdà-Costa N, Bonet J, Fernández MR, Avilés FX, Oliva B, Villegas S. Prediction of a new class of RNA recognition motif. J Mol Model. 2011 Aug;17(8):1863-75.
doi: 10.1007/s00894-010-0888-0.

8.10. Multivalent antibodies: when design surpasses evolution.

Cuesta AM, Sainz-Pastor N, Bonet J, Oliva B, Alvarez-Vallina L. Multivalent antibodies: when design surpasses evolution. Trends Biotechnol. 2010 Jul;28(7):355-62.
doi: 10.1016/j.tibtech.2010.03.007.

8.11. In vivo tumor targeting and imaging with engineered trivalent antibody fragments containing collagen-derived sequences.

Cuesta AM, Sánchez-Martín D, Sanz L, Bonet J, Compte M, Kremer L, Blanco FJ, Oliva B, Alvarez-Vallina L. In vivo tumor targeting and imaging with engineered trivalent antibody fragments containing collagen-derived sequences. PLoS One. 2009;4(4):e5381. doi: 10.1371/journal.pone.0005381

8.12. Characterization of protein hubs by inferring interacting motifs from protein interactions.

Aragues R, Sali A, Bonet J, Marti-Renom MA, Oliva B. Characterization of protein hubs by inferring interacting motifs from protein interactions. PLoS Comput Biol. 2007 Sep;3(9):1761-71. DOI: 10.1371/journal.pcbi.0030178

8.13. The role of residue stability in transient protein-protein interactions involved in enzymatic phosphate hydrolysis. A computational study.

Bonet J, Caltabiano G, Khan AK, Johnston MA, Corbí C, Gómez A, Rovira X, Teyra J, Villà-Freixa J. The role of residue stability in transient protein-protein interactions involved in enzymatic phosphate hydrolysis. A computational study. Proteins. 2006 Apr 1;63(1):65-77. DOI: 10.1002/prot.20791

9. REFERENCES

References

[1] Garcia-Garcia J, Bonet J, Guney E, Fornes O, Planas-Iglesias J, Oliva B. Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details. Molecular Informatics 2012;31:342–62.

[2] Voet D, Voet JG. Biochemistry, 4th Edition. Wiley; 2011.

[3] Gerardus Johannes M. On the composition of some animal substances. Journal Für Praktische Chemie 1838;129.

[4] Crick F. Central dogma of molecular biology. Nature 1970;227:561–3.

[5] Kircher M, Kelso J. High-throughput DNA sequencing--concepts and limitations. Bioessays 2010;32:524–36.

[6] Li G-W, Xie XS. Central dogma at the single-molecule level in living cells. Nature 2011;475:308–15.

[7] Temperley R, Richter R, Dennerlein S, Lightowlers RN, Chrzanowska-Lightowlers ZM. Hungry codons promote frameshifting in human mitochondrial ribosomes. Science 2010;327:301.

[8] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291:1304–51.

[9] Lawrence RT, Villén J. Drafts of the human proteome. Nat Biotechnol 2014;32:752–3.

[10] UniProt Consortium. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 2014;42:D191–8.

[11] O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinformatics 2002;3:275–84.

[12] Garcia-Garcia J, Guney E, Aragues R, Planas-Iglesias J, Oliva B. Biana: a software framework for compiling biological interactions and analyzing networks. BMC Bioinformatics 2010;11:56.

[13] Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 2012;40:D130–5.

[14] Harte RA, Farrell CM, Loveland JE, Suner M-M, Wilming L, Aken B, et al. Tracking and coordinating an international curation effort for the CCDS Project. Database (Oxford) 2012;2012:bas008.

[15] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database--2009 update. Nucleic Acids Res 2009;37:D767–72.

[16] Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE, Mouse Genome Database Group. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res 2014;42:D810–7.

[17] Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelets VB, et al. FlyBase: improvements to the bibliography. Nucleic Acids Res 2013;41:D751–7.

[18] Harris TW, Baran J, Bieri T, Cabunoc A, Chan J, Chen WJ, et al. WormBase 2014: new views of curated biology. Nucleic Acids Res 2014;42:D789–93.

[19] Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic Acids Res 2014;42:D749–55.

[20] Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, et al. DisProt: a database of protein disorder. Bioinformatics 2005;21:137–40.

[21] Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. Nature 2007;450:973–82.

[22] Brocchieri L, Karlin S. Protein length in eukaryotic and prokaryotic proteomes. Nucleic Acids Res 2005;33:3390–400.

[23] Gutteridge A, Thornton JM. Understanding nature's catalytic toolkit. Trends Biochem Sci 2005;30:622–9.

[24] Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USa 1951;37:205–11.

[25] Eisenberg D. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. Proc Natl Acad Sci USa 2003;100:11207–10.

[26] Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci USa 1973;70:697–701.

[27] Sprinzak E, Margalit H. Correlated sequence-signatures as markers of protein-protein interaction. J Mol Biol 2001;311:681–92.

[28] Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:D226–9.

[29] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res 2014;42:D222–30.

[30] Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, et al. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res 2013;41:D490–8.

[31] Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–33.

[32] Mészáros B, Dosztányi Z, Simon I. Disordered binding regions and linear motifs--bridging the gap between two models of molecular recognition. PLoS ONE 2012;7:e46829.

[33] Papsdorf K, Richter K. Protein folding, misfolding and quality control: the role of molecular chaperones. Essays Biochem 2014;56:53–68.

[34] Bonet J, Caltabiano G, Khan AK, Johnston MA, Corbí C, Gómez A, et al. The role of residue stability in transient protein-protein interactions involved in enzymatic phosphate hydrolysis. A computational study. Proteins 2006;63:65–77.

[35] Berman HM, Kleywegt GJ, Nakamura H, Markley JL. The Protein Data Bank archive as an open data resource. J Comput Aided Mol Des 2014.

[36] Holm L, Rosenström P. Dali server: conservation mapping in 3D. Nucleic Acids Res 2010;38:W545–9.

[37] Mosca R, Ceol A, Stein A, Olivella R, Aloy P. 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res 2014;42:D374–9.

[38] Jefferson ER, Walsh TP, Roberts TJ, Barton GJ. SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions. Nucleic Acids Res 2007;35:D580–9.

[39] Montelione GT. The Protein Structure Initiative: achievements and visions for the future. F1000 Biol Rep 2012;4:7.

[40] Crowfoot D. X-ray single crystal photographs of insulin. Nature 1935.

[41] Wüthrich K. The way to NMR structures of proteins. Nat Struct Biol 2001;8:923–5.

[42] Ochi T, Bolanos-Garcia VM, Stojanoff V, Moreno A. Perspectives on protein crystallisation. Prog Biophys Mol Biol 2009;101:56–63.

[43] Bonet J, Segura J, Planas-Iglesias J, Oliva B, Fernandez-Fuentes N. Frag"r"Us: knowledge-based sampling of protein backbone conformations for de novo structure-based protein design. Bioinformatics 2014.

[44] Braun W, Wider G, Lee KH, Wüthrich K. Conformation of glucagon in a lipid-water interphase by 1H nuclear magnetic resonance. J Mol Biol 1983;169:921–48.

[45] Frank J. Single-particle imaging of macromolecules by cryo-electron microscopy. Annu Rev Biophys Biomol Struct 2002;31:303–19.

[46] Neylon C. Small angle neutron and X-ray scattering in structural biology: recent examples from the literature. Eur Biophys J 2008;37:531–41.

[47] Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008;26:1135–45.

[48] Sheehan D, O'Sullivan S. Online homology modelling as a means of bridging the sequence-structure gap. Bioeng Bugs 2011;2:299–305.

[49] Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. Annu Rev Biophys Biomol Struct 1996;25:113–36.

[50] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28:3150–2.

[51] Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.

[52] Planas-Iglesias J, Bonet J, Marín-López M, Feliu E, Gursoy A, Oliva B. Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence. *Protein-Protein Interactions - Computational and Experimental Tools*, InTech; 2012.

[53] Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 2005;59:467–75.

[54] Pollastri G, McLysaght A. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 2005;21:1719–20.

[55] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.

[56] Kountouris P, Hirst JD. Prediction of backbone dihedral angles and protein secondary structure using support vector machines. BMC Bioinformatics 2009;10:437.

[57] Dosztányi Z, Csizmok V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 2005;347:827–39.

[58] Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res 2003;31:3701–8.

[59] Chen H, Skolnick J. M-TASSER: an algorithm for protein quaternary structure prediction. Biophys J 2008;94:918–28.

[60] Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Meth Enzymol 2011;487:545–74.

[61] Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol 2012;30:1072–80.

[62] Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 2012;80:1715–35.

[63] Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.

[64] Kelley LA, Sternberg MJE. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 2009;4:363–71.

[65] Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–60.

[66] Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Meth Enzymol 2003;374:461–91.

[67] Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 2014;42:W252–8.

[68] Krieger E, Koraimann G, Vriend G. Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. Proteins 2002;47:393–402.

[69] Andrec M, Harano Y, Jacobson MP, Friesner RA, Levy RM. Complete protein structure determination using backbone residual dipolar couplings and sidechain rotamer prediction. J Struct Funct Genomics 2002;2:103–11.

[70] Schwede T. Protein modeling: what happened to the "protein structure gap"? Structure 2013;21:1531–40.

[71] Colak R, Kim T, Michaut M, Sun M, Irimia M, Bellay J, et al. Distinct types of disorder in the human proteome: functional implications for alternative splicing. PLoS Comput Biol 2013;9:e1003030.

[72] Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. Electrophoresis 2009;30 Suppl 1:S162–73.

[73] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.

[74] Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol 2011;7:e1002195.

[75] Fornes O, Aragues R, Espadaler J, Marti-Renom MA, Sali A, Oliva B. ModLink+: improving fold recognition by using protein-protein interactions. Bioinformatics 2009;25:1506–12.

[76] Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.

[77] Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, et al. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003;31:3497–500.

[78] Magis C, Taly J-F, Bussotti G, Chang J-M, Di Tommaso P, Erb I, et al. T-Coffee: Tree-based consistency objective function for alignment evaluation. Methods Mol Biol 2014;1079:117–29.

[79] Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, et al. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res 2010;38:W695–9.

[80] Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963;7:95–9.

[81] Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 1996;8:477–86.

[82] Melo F, Feytmans E. Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 1998;277:1141–52.

[83] Aloy P, Oliva B. Splitting statistical potentials into meaningful scoring functions: testing the prediction of near-native structures from decoy conformations. BMC Struct Biol 2009;9:71.

[84] Fernandez-Fuentes N, Rai BK, Madrid-Aliste CJ, Fajardo JE, Fiser A. Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. Bioinformatics 2007;23:2558–65.

[85] Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 2013;29:845–54.

[86] Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nat Methods 2013;10:221–7.

[87] Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys 2003;36:307–40.

[88] Erdin S, Lisewski AM, Lichtarge O. Protein function prediction: towards integration of similarity metrics. Curr Opin Struct Biol 2011;21:180–8.

[89] Andrade MA, Ouzounis C, Sander C, Tamames J, Valencia A. Functional classes in the three

domains of life. J Mol Evol 1999;49:551–7.

[90] Bairoch A. The ENZYME database in 2000. Nucleic Acids Res 2000;28:304–5.

[91] The Gene Ontology Consortium. The Gene Ontology: enhancements for 2011. Nucleic Acids Res 2012;40:D559–64.

[92] Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 2013;42:D503–9.

[93] Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 2010;38:D355–60.

[94] Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, et al. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 2004;32:D431–3.

[95] Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 2012;41:D377–86.

[96] Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martínez C, et al. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res 2012;41:D605–12.

[97] Boyce S, Tipton KF. Enzyme Classification and Nomenclature. Chichester: John Wiley & Sons, Ltd; 2001.

[98] Omelchenko MV, Galperin MY, Wolf YI, Koonin EV. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biol Direct 2010;5:31.

[99] Schulz S, Stenzhorn H, Boeker M, Smith B. Strengths and limitations of formal ontologies in the biomedical domain. Rev Electron Comun Inf Inov Saude 2009;3:31–45.

[100] Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics 2007;23:1274–81.

[101] Wang Z, Zhang X-C, Le MH, Xu D, Stacey G, Cheng J. A protein domain co-occurrence network approach for predicting protein function and inferring species phylogeny. PLoS ONE 2011;6:e17906.

[102] Rost B. Enzyme function less conserved than anticipated. J Mol Biol 2002;318:595–608.

[103] Jones CE, Baumann U, Brown AL. Automated methods of predicting the function of biological sequences using GO and BLAST. BMC Bioinformatics 2005;6:272.

[104] Bork P, Koonin EV. Protein sequence motifs. Curr Opin Struct Biol 1996;6:366–76.

[105] Espadaler J, Querol E, Avilés FX, Oliva B. Identification of function-associated loop motifs and application to protein function prediction. Bioinformatics 2006;22:2237–43.

[106] Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Acids Res 2013;41:D344–7.

[107] Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, et al. Phospho.ELM: a database of phosphorylation sites--update 2011. Nucleic Acids Res 2011;39:D261–7.

[108] Segura J, Oliva B, Fernandez-Fuentes N. CAPS-DB: a structural classification of helix-capping motifs. Nucleic Acids Res 2012;40:D479–85.

[109] Jalencas X, Mestres J. Chemoisosterism in the proteome. J Chem Inf Model 2013;53:279–92.

[110] Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 2007;8:995–1005.

[111] Lage K. Protein-protein interactions and genetic diseases: The interactome. Biochim Biophys Acta 2014.

[112] Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, et al. HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res 2013;41:D584–9.

[113] Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 2010;38:D196–203.

[114] Xie H, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A, et al. Large-scale protein annotation through gene ontology. Genome Res 2002;12:785–94.

[115] Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, et al. The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. Database (Oxford) 2012;2012:bas019.

[116] Meereis F, Kaufmann M. Extension of the COG and arCOG databases by amino acid and nucleotide sequences. BMC Bioinformatics 2008;9:479.

[117] Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, et al. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 2012;40:D306–12.

[118] Dodge C, Schneider R, Sander C. The HSSP database of protein structure-sequence alignments

and family profiles. Nucleic Acids Res 1998;26:313–5.

[119] Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–47.

[120] Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208:1–22.

[121] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 2011;39:D561–8.

[122] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 2014;42:D358–63.

[123] Krull M, Pistor S, Voss N, Kel A, Reuter I, Kronenberg D, et al. TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. Nucleic Acids Res 2006;34:D546–51.

[124] Ho BK, Thomas A, Brasseur R. Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. Protein Sci 2003;12:2508–22.

[125] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–637.

[126] Linhananta A, Zhou H, Zhou Y. The dual role of a loop with low loop contact distance in folding and domain swapping. Protein Sci 2002;11:1695–701.

[127] Chintapalli SV, Illingworth CJR, Upton GJG, Sacquin-Mora S, Reeves PJ, Mohammedali HS, et al. Assessing the effect of dynamics on the closed-loop protein-folding hypothesis. J R Soc Interface 2014;11:20130935.

[128] Fersht AR. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. Proc Natl Acad Sci USa 2000;97:1525–9.

[129] Zhou H-X. Loops, linkages, rings, catenanes, cages, and crowders: entropy-based strategies for stabilizing proteins. Acc Chem Res 2004;37:123–30.

[130] Kumar S, Nussinov R. How do thermophilic proteins deal with heat? Cell Mol Life Sci 2001;58:1216–33.

[131] Hoedemaeker FJ, van Eijsden RR, Díaz CL, de Pater BS, Kijne JW. Destabilization of pea lectin by substitution of a single amino acid in a surface loop. Plant Mol Biol 1993;22:1039–46.

[132] Fetrow JS. Omega loops: nonregular secondary structures significant in protein function and stability. Faseb J 1995;9:708–17.

[133] Saraste M, Sibbald PR, Wittinghofer A. The P-loop--a common motif in ATP- and GTP-binding proteins. Trends Biochem Sci 1990;15:430–4.

[134] Kawasaki H, Kretsinger RH. Calcium-binding proteins 1: EF-hands. Protein Profile 1995;2:297–490.

[135] Wlodawer A, Miller M, Jaskólski M, Sathyanarayana BK, Baldwin E, Weber IT, et al. Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. Science 1989;245:616–21.

[136] Johnson LN, Lowe ED, Noble ME, Owen DJ. The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases. FEBS Lett 1998;430:1–11.

[137] Gunasekaran K, Nussinov R. Modulating functional loop movements: the role of highly conserved residues in the correlated loop motions. Chembiochem 2004;5:224–30.

[138] Adams JA. Activation loop phosphorylation and catalysis in protein kinases: is there functional evidence for the autoinhibitor model? Biochemistry 2003;42:601–7.

[139] Bernstein LS, Ramineni S, Hague C, Cladman W, Chidiac P, Levey AI, et al. RGS2 binds directly and selectively to the M1 muscarinic acetylcholine receptor third intracellular loop to modulate Gq/11alpha signaling. J Biol Chem 2004;279:21248–56.

[140] Wright RHG, Castellano G, Bonet J, Le Dily F, Font-Mateu J, Ballaré C, et al. CDK2-dependent activation of PARP-1 is required for hormonal gene regulation in breast cancer cells. Genes Dev 2012;26:1972–83.

[141] Feng W, Shi Y, Li M, Zhang M. Tandem PDZ repeats in glutamate receptor-interacting proteins have a novel mode of PDZ domain-mediated target binding. Nat Struct Biol 2003;10:972–8.

[142] Planas-Iglesias J, Bonet J, Garcia-Garcia J, Marín-López MA, Feliu E, Oliva B. Understanding Protein-Protein Interactions Using Local Structural Features. J Mol Biol 2013.

[143] Stella S, Molina R, López-Méndez B, Juillerat A, Bertonati C, Daboussi F, et al. BuD, a helix-loop-helix DNA-binding domain for genome modification. Acta Crystallogr D Biol Crystallogr 2014;70:2042–52.

[144] Gabrielli E, Pericolini E, Cenci E, Ortelli F, Magliani W, Ciociola T, et al. Antibody complementarity-determining regions (CDRs): a bridge between adaptive and innate immunity. PLoS ONE 2009;4:e8187.

[145] Fernandez-Fuentes N, Hermoso A, Espadaler J, Querol E, Avilés FX, Oliva B. Classification of common functional loops of kinase super-families. Proteins 2004;56:539–55.

[146] Insuline glusine (Apidra): a new rapid-acting insulin. Med Lett Drugs Ther 2006;48:33–4.

[147] Ugochukwu UC, Manning DAC, Fialips CI. Microbial degradation of crude oil hydrocarbons on organoclay minerals. J Environ Manage 2014;144:197–202.

[148] Wojcieszyńska D, Domaradzka D, Hupert-Kocurek K, Guzik U. Bacterial degradation of naproxen - Undisclosed pollutant in the environment. J Environ Manage 2014;145:157–61.

[149] Bahassi EM, Stambrook PJ. Next-generation sequencing technologies: breaking the sound barrier of human genetics. Mutagenesis 2014;29:303–10.

[150] Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, et al. High-quality binary protein interaction map of the yeast interactome network. Science 2008;322:104–10.

[151] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USa 1992;89:10915–9.

[152] Yamada K, Tomii K. Revisiting amino acid substitution matrices for identifying distantly related proteins. Bioinformatics 2013.

[153] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014;42:D1091–7.

[154] Zhu Q, Vajda S, Smith TF. Beta-turn new classification and its some features in proteins. Chin Med Sci J 1997;12:84–91.

[155] Edwards MS, Sternberg JE, Thornton JM. Structural and sequence patterns in the loops of beta alpha beta units. Protein Eng 1987;1:173–81.

[156] Wintjens RT, Rooman MJ, Wodak SJ. Automatic classification and analysis of alpha alpha-turn motifs in proteins. J Mol Biol 1996;255:235–53.

[157] Rooman MJ, Rodriguez J, Wodak SJ. Automatic definition of recurrent local structure motifs in proteins. J Mol Biol 1990;213:327–36.

[158] Tendulkar AV, Joshi AA, Sohoni MA, Wangikar PP. Clustering of protein structural fragments reveals modular building block approach of nature. J Mol Biol 2004;338:611–29.

[159] Baeten L, Reumers J, Tur V, Stricher F, Lenaerts T, Serrano L, et al. Reconstruction of protein backbones from the BriX collection of canonical protein fragments. PLoS Comput Biol 2008;4:e1000083.

[160] Vanhee P, Verschueren E, Baeten L, Stricher F, Serrano L, Rousseau F, et al. BriX: a database of protein building blocks for structural analysis, modeling and design. Nucleic Acids Res 2011;39:D435–42.

[161] Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 2014;42:D304–9.

[162] Brüschweiler R. Efficient RMSD measures for the comparison of two molecular ensembles. Root-mean-square deviation. Proteins 2003;50:26–34.

[163] Kwasigroch JM, Chomilier J, Mornon JP. A global taxonomy of loops in globular proteins. J Mol Biol 1996;259:855–72.

[164] Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 1988;3:71–84.

[165] Sklenar H, Etchebest C, Lavery R. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins 1989;6:46–60.

[166] Regad L, Martin J, Camproux AC. Identification of non random motifs in loops using a structural alphabet. Audio, Transactions of the IRE Professional Group on 2006:1–9.

[167] Oliva B, Bates PA, Querol E, Avilés FX, Sternberg MJ. An automated classification of the structure of protein loops. J Mol Biol 1997;266:814–30.

[168] Bonet J, Planas-Iglesias J, Garcia-Garcia J, Marín-López MA, Fernandez-Fuentes N, Oliva B. ArchDB 2014: structural classification of loops in proteins. Nucleic Acids Res 2014;42:D315–9.

[169] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.

[170] Everitt B. Chapter 3. Cluster Analysis, London: 1974, pp. 45–60.

[171] Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Avilés FX, Sternberg MJE, et al. ArchDB: automated protein loop classification as a tool for structural genomics. Nucleic Acids Res 2004;32:D185–8.

[172] Hermoso A, Espadaler J, Enrique Querol E, Avilés FX, Sternberg MJE, Oliva B, et al. Including Functional Annotations and Extending the Collection of Structural Classifications of Protein Loops (ArchDB). Bioinform Biol Insights 2009;1:77–90.

[173] Tyagi M, Hashimoto K, Shoemaker BA, Wuchty S, Panchenko AR. Large-scale mapping of human protein interactome using structural complexes. EMBO Rep 2012;13:266–71.

[174] Mosca R, Pons T, Ceol A, Valencia A, Aloy P. Towards a detailed atlas of protein-protein

interactions. Curr Opin Struct Biol 2013.

[175] Fitzkee NC, Fleming PJ, Gong H, Panasik N, Street TO, Rose GD. Are proteins made from a limited parts list? Trends Biochem Sci 2005;30:73–80.

[176] Akiva E, Itzhaki Z, Margalit H. Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. Proc Natl Acad Sci USa 2008;105:13292–7.

[177] Hildebrand PW, Goede A, Bauer RA, Gruening B, Ismer J, Michalsky E, et al. SuperLooper--a prediction server for the modeling of loops in globular and membrane proteins. Nucleic Acids Res 2009;37:W571–4.

[178] Moon HS, Bhak J, Lee KH, Lee D. Architecture of basic building blocks in protein and domain structural interaction networks. Bioinformatics 2005;21:1479–86.

[179] de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 2000;41:271–87.

[180] Regad L, Saladin A, Maupetit J, Geneix C, Camproux A-C. SA-Mot: a web server for the identification of motifs of interest extracted from protein loops. Nucleic Acids Res 2011;39:W203–9.

[181] Fernandez-Fuentes N, Querol E, Avilés FX, Sternberg MJE, Oliva B. Prediction of the conformation and geometry of loops in globular proteins: testing ArchDB, a structural classification of loops. Proteins 2005;60:746–57.

[182] Fernandez-Fuentes N, Oliva B, Fiser A. A supersecondary structure library and search algorithm for modeling loops in protein structures. Nucleic Acids Res 2006;34:2085–97.

[183] Van Dongen S. Graph Clustering Via a Discrete Uncoupling Process. SIAM J Matrix Anal & Appl 2008;30:121–41.

[184] Planas-Iglesias J, Marín-López MA, Bonet J, Garcia-Garcia J, Oliva B. iLoops: a protein-protein interaction prediction server based on structural features. Bioinformatics 2013.

[185] Carter P, Andersen CAF, Rost B. DSSPcont: Continuous secondary structure assignments for proteins. Nucleic Acids Res 2003;31:3293–5.

[186] Fernandez-Fuentes N, Fiser A. Saturating representation of loop conformational fragments in structure databanks. BMC Struct Biol 2006;6:15.

[187] Russell RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 1992;14:309–23.

[188] Regad L, Martin J, Nuel G, Camproux A-C. Mining protein loops using a structural alphabet and statistical exceptionality. BMC Bioinformatics 2010;11:75.

[189] Hu XZ, Li QZ. Prediction of the beta-hairpins in proteins using support vector machine. Protein J 2008;27:115–22.

[190] Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, et al. DrugBank 3.0: a comprehensive resource for "omics" research on drugs. Nucleic Acids Res 2011;39:D1035–41.

[191] Michalsky E, Goede A, Preissner R. Loops In Proteins (LIP)--a comprehensive loop database for homology modelling. Protein Eng 2003;16:979–85.

[192] Wohlfahrt G, Hangoc V, Schomburg D. Positioning of anchor groups in protein loop prediction: the importance of solvent accessibility and secondary structure elements. Proteins 2002;47:370–8.

[193] Heuser P, Wohlfahrt G, Schomburg D. Efficient methods for filtering and ranking fragments for the prediction of structurally variable regions in proteins. Proteins 2004;54:583–95.

[194] Shen Y, Bax A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J Biomol NMR 2010;48:13–22.

[195] Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 2009;44:213–23.

[196] Rykunov D, Fiser A. New statistical potential for quality assessment of protein models and a survey of energy functions. BMC Bioinformatics 2010;11:128.

[197] Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–52.

[198] Morozov AV, Kortemme T. Potential functions for hydrogen bonds in protein structure prediction and design. Adv Protein Chem 2005;72:1–38.

[199] Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–4.

[200] Fernandez-Fuentes N, Zhai J, Fiser A. ArchPRED: a template based loop structure prediction server. Nucleic Acids Res 2006;34:W173–6.

[201] Kolaskar AS, Kulkarni-Kale U. Sequence alignment approach to pick up conformationally similar protein fragments. J Mol Biol 1992;223:1053–61.

[202] Shortle D. Composites of local structure propensities: evidence for local encoding of long-range structure. Protein Sci 2002;11:18–26.

[203] Fiser A, Sali A. ModLoop: automated modeling of loops in protein structures. Bioinformatics

2003;19:2500–1.
[204] Khare SD, Fleishman SJ. Emerging themes in the computational design of novel enzymes and protein-protein interfaces. FEBS Lett 2013;587:1147–54.
[205] Kiss G, Çelebi-Ölçüm N, Moretti R, Baker D, Houk KN. Computational enzyme design. Angew Chem Int Ed Engl 2013;52:5700–25.
[206] Lees JPB, Manlandro CM, Picton LK, Tan AZE, Casares S, Flanagan JM, et al. A designed point mutant in Fis1 disrupts dimerization and mitochondrial fission. J Mol Biol 2012;423:143–58.
[207] Eiben CB, Siegel JB, Bale JB, Cooper S, Khatib F, Shen BW, et al. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nat Biotechnol 2012;30:190–2.
[208] Murphy PM, Bolduc JM, Gallaher JL, Stoddard BL, Baker D. Alteration of enzyme specificity by computational loop remodeling and design. Proc Natl Acad Sci USa 2009;106:9215–20.
[209] Hu X, Wang H, Ke H, Kuhlman B. High-resolution design of a protein loop. Proc Natl Acad Sci USa 2007;104:17668–73.
[210] Fernandez-Fuentes N, Dybas JM, Fiser A. Structural characteristics of novel protein folds. PLoS Comput Biol 2010;6:e1000750.
[211] Choi Y, Deane CM. FREAD revisited: Accurate loop structure prediction using a database search algorithm. Proteins 2010;78:1431–40.
[212] Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–91.
[213] Claren J, Malisi C, Höcker B, Sterner R. Establishing wild-type levels of catalytic activity on natural and artificial (beta alpha)8-barrel protein scaffolds. Proc Natl Acad Sci USa 2009;106:3704–9.
[214] Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, et al. De novo computational design of retro-aldol enzymes. Science 2008;319:1387–91.
[215] Wang L, Althoff EA, Bolduc J, Jiang L, Moody J, Lassila JK, et al. Structural analyses of covalent enzyme-substrate analog complexes reveal strengths and limitations of de novo enzyme design. J Mol Biol 2012;415:615–25.
[216] Saab-Rincón G, Olvera L, Olvera M, Rudiño-Piñera E, Benites E, Soberón X, et al. Evolutionary walk between (β/α)(8) barrels: catalytic migration from triosephosphate isomerase to thiamin phosphate synthase. J Mol Biol 2012;416:255–70.
[217] Azoitei ML, Correia BE, Ban Y-EA, Carrico C, Kalyuzhniy O, Chen L, et al. Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. Science 2011;334:373–6.
[218] Priyadarshi A, Roy A, Kim K-S, Kim EE, Hwang KY. Structural insights into mouse anti-apoptotic Bcl-xl reveal affinity for Beclin 1 and gossypol. Biochem Biophys Res Commun 2010;394:515–21.
[219] Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M. Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. Science 2000;289:1546–50.
[220] Hennig M, Darimont BD, Jansonius JN, Kirschner K. The catalytic mechanism of indole-3-glycerol phosphate synthase: crystal structures of complexes of the enzyme from Sulfolobus solfataricus with substrate analogue, substrate, and product. J Mol Biol 2002;319:757–66.
[221] Borchert TV, Abagyan R, Kishan KV, Zeelen JP, Wierenga RK. The crystal structure of an engineered monomeric triosephosphate isomerase, monoTIM: the correct modelling of an eight-residue loop. Structure 1993;1:205–13.
[222] Larsson AM, Bergfors T, Dultz E, Irwin DC, Roos A, Driguez H, et al. Crystal structure of Thermobifida fusca endoglucanase Cel6A in complex with substrate and inhibitor: the role of tyrosine Y73 in substrate ring distortion. Biochemistry 2005;44:12915–22.
[223] Friedberg I. Automated protein function prediction--the genomic challenge. Brief Bioinformatics 2006;7:225–42.
[224] Duan Z-H, Hughes B, Reichel L, Perez DM, Shi T. The relationship between protein sequences and their gene ontology functions. BMC Bioinformatics 2006;7 Suppl 4:S11.
[225] Devos D, Valencia A. Practical limits of function prediction. Proteins 2000;41:98–107.
[226] Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 2003;333:863–82.
[227] Hegyi H, Gerstein M. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. Genome Res 2001;11:1632–40.
[228] Deng M, Chen T, Sun F. An integrated probabilistic model for functional prediction of proteins. J Comput Biol 2004;11:463–75.
[229] Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 2004;32:5539–45.
[230] Valencia A. Automatic annotation of protein function. Curr Opin Struct Biol 2005;15:267–74.

[231] Feliu JX, Benito A, Oliva B, Avilés FX, Villaverde A. Conformational flexibility in a highly mobile protein loop of foot-and-mouth disease virus: distinct structural requirements for integrin and antibody binding. J Mol Biol 1998;283:331–8.

[232] Alanazi AM, Neidle EL, Momany C. The DNA-binding domain of BenM reveals the structural basis for the recognition of a T-N11-A sequence motif by LysR-type transcriptional regulators. Acta Crystallogr D Biol Crystallogr 2013;69:1995–2007.

[233] Espadaler J, Eswar N, Querol E, Avilés FX, Sali A, Marti-Renom MA, et al. Prediction of enzyme function by combining sequence similarity and protein interactions. BMC Bioinformatics 2008;9:249.

[234] Regad L, Martin J, Camproux A-C. Dissecting protein loops with a statistical scalpel suggests a functional implication of some structural motifs. BMC Bioinformatics 2011;12:247.

[235] Hall M, Frank E, Holmes G, Pfahringer B. The WEKA data mining software: an update. Acm Sigkdd … 2009.

[236] Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 2008;36:3420–35.

[237] Quinlan JR. C4. 5: programs for machine learning 1993.

[238] Ince DC, Hatton L, Graham-Cumming J. The case for open computer programs. Nature 2012;482:485–8.

[239] Corpas M, Fatumo SA. How not to be a bioinformatician. Source Code for Biology … 2012.

[240] Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol 2013;9:e1003285.

[241] Schultheiss SJ, Münch MC, Andreeva GD, Rätsch G. Persistence and availability of Web services in computational biology. PLoS ONE 2011.

[242] Kozma D, Simon I, Tusnády GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res 2013;41:D524–9.

[243] Ahmed J, Worth CL, Thaben P, Matzig C, Blasse C, Dunkel M, et al. FragmentStore--a comprehensive database of fragments linking metabolites, toxic molecules and drugs. Nucleic Acids Res 2011;39:D1049–54.

[244] McCammon JA. Theory of biomolecular recognition. Curr Opin Struct Biol 1998.

[245] Kastritis PL, Rodrigues JPGLM, Folkers GE, Boelens R, Bonvin AMJJ. Proteins Feel More Than They See: Fine-Tuning of Binding Affinity by Properties of the Non-Interacting Surface. J Mol Biol 2014.

[246] Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A. PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. Nucleic Acids Res 2014;42:W285–9.

[247] Garcia-Garcia J, Schleker S, Klein-Seetharaman J, Oliva B. BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. Nucleic Acids Res 2012;40:W147–51.

[248] Engin HB, Gursoy A, Nussinov R, Keskin O. Network-based strategies can help mono- and poly-pharmacology drug discovery: a systems biology view. Curr Pharm Des 2014;20:1201–7.

[249] Aragues R, Sali A, Bonet J, Marti-Renom MA, Oliva B. Characterization of protein hubs by inferring interacting motifs from protein interactions. PLoS Comput Biol 2007;3:1761–71.

[250] Kim PM, Lu LJ, Xia Y, Gerstein MB. Relating three-dimensional structures to protein networks provides evolutionary insights. Science 2006;314:1938–41.

[251] Jonassen I, Collins JF, Higgins DG. Finding flexible patterns in unaligned protein sequences. Protein Sci 1995;4:1587–95.

[252] Boxem M, Maliga Z, Klitgord N, Li N, Lemmens I, Mana M, et al. A protein domain-based interactome network for C. elegans early embryogenesis. Cell 2008;134:534–45.