El titulo de la tesi: In-silico methods to drug discovery

El subtitulo de la tesi: Cancer

Autor: Francisco Martínez Jiménez

TESI DOCTORAL UPF / ANY L'any de la tesi: 2016

DIRECTOR DE LA TESI
Director: Marc A. Marti-Renom Departament Departament: Biomedicine

A mi madre.

Agradecimientos Agraeixo....

Abstract

This is the abstract of the thesis in English. Please, use less than 150 words.

Resum

Vet aqui el resum de la tesi en catala.

Prefaci

Sumari

Índex de figures

Índex de taules

Summary

sencillo
Consta de

tesi-upf.cls `book`

1. Es redissenya la portada `\maketitle`).
2.
3. Es redefineix `cleardoublepage` para que las paginas en blanco no se numeren

Preambulel

paquets `crop` i `geometry`.

Taules Includes `figure` i un `tabular`

Index

1. lo llamas en preambulo `\usepackage{makeidx} \makeindex` con esto lo imprimes `\printindex` con esto lo creas `makeindex`

CAPÍTOL 1

INTRODUCTION

1.1 Protein are essential molecules

The importance of proteins in biological chemistry is just reflected by their name, derived from the Greek word *proteios*, and that means "of the first rank"[1]. Their presence is so essential that they constitute most of the cell dry mass [1]. They are not only the cell's building blocks, but also they perform nearly all the cell's functions. Some roles of proteins include serving as structural components of cells and tissues (e.g., *keratin* or *collagen*), transmission of information between cells by hormones such as the *insulin* or the *oxytocin*, facilitating the transport and storage of small molecules (e.g., the transport of oxygen by *hemoglobin*) or providing a defense against foreign invaders (e.g., antibodies). Other proteins such as the *actin* and the *myosin* are responsible of muscle contraction and therefore our movement. However, the most fundamental role of proteins is their ability to act as enzymes, which, catalyzes most of the chemical reactions in biological systems. In summary, proteins are crucial macromolecules present in most of the processes carried out by the cell and, in spite of being extensively studied for many years, they still have many unanswered questions.

1.1.1 Protein structure

A protein is a molecule made from a long chain of amino acids linked thorough a covalent peptide bond. Proteins are therefore also known as *polypeptides*. At-

[1]The term protein was coined by Jons Jacob Berzelius in 1838. It was first used by Gerardus Johannes Mulder, advised by Berzelius, in its publication *Bulletin des Sciences Physiques et Naturelles en Néerlande (1838). pg 104. SUR LA COMPOSITION DE QUELQUES SUBSTANCES ANIMALES*, where he observed that all proteins seemed to have the same empirical formula and came out to the erroneous idea that they might be composed of a single type of very large molecule. Berzelius proposed the name because the material seemed to be the primitive substance of animal nutrition that plants prepare for herbivores.

tached to this repetitive chain are those portions of the amino acids that are not involved in the covalent bond, the **side chains**. Side chains confer the different physico-chemical properties of each of the 20 types of amino acids [2]. The composition of the amino acid sequence determines the function and the structure of a protein. That is because the unique sequence creates a specific pattern of attractive and repulsive forces between amino acids along the polypeptide that leads to a folding process resulting in a specific three-dimensional structure. These forces are usually non-covalent interactions between the side chains of the amino acids. Non-covalent interactions are weaker than covalent ones, allowing the folded structure to certain degree of conformation mobility i.e: to be dynamic. This phenomenon is really important to facilitate the interaction with other molecules as we will explore further in 1.1.3.

Protein structures are complex conformation of atoms organized in a hierarchical manner 1.1. The first level of this hierarchy, referred to as the **primary structure**, is the ordered sequence of amino acids of the polypeptide. Certain segments of these chains, tend to form simple shapes such as helices, strands, turns or loops. These folding patterns are referred to as secondary elements and collectively constitute the **secondary structure** of the protein. The two most frequent type of secondary elements are the α-helixes and the β-sheets [3]. The overall chain tends to fold further into a three-dimensional **tertiary structure**. Contrary to the secondary structure, the tertiary structure folding is driven by interactions from amino acids far apart in the primary sequence. The tertiary structure, is generally the most stable form of the protein, that is, the one that minimizes its free energy [4]. Furthermore, the tertiary structure is also the biologically active form of the protein, and its unfolding usually leads towards partial o total inactivation of the protein. Finally, some proteins are composed by multiple folded chains. In such cases, each folded subunit folds independently and then joins the others forming a biologically active complex. This type of organization is considered as the **quaternary structure**.

This traditional paradigm of protein structure has been challenged by some exceptions of proteins that lacks of a fixed or ordered three-dimensional structure. The intrinsically disordered proteins (IDPs) cover a wide spectrum of states from fully unstructured to partially structured including conformations such as *random coils* or *molten globules*. Moreover, some factors may lead to the permanent loss of structure of a protein, and when that occurs, they endanger the entire organism. How problematic protein misfolding can be for the organism is illustrated by examples such as cystic fibrosis, Alzheimer's, Parkinson's and Huntington's diseases.

Figure 1.2 from the seminal paper [5] shows the correlation degree to which protein structures changed as a function of sequence divergence. This work helped to set up the fundaments of what is considered a central paradigm in protein evo-

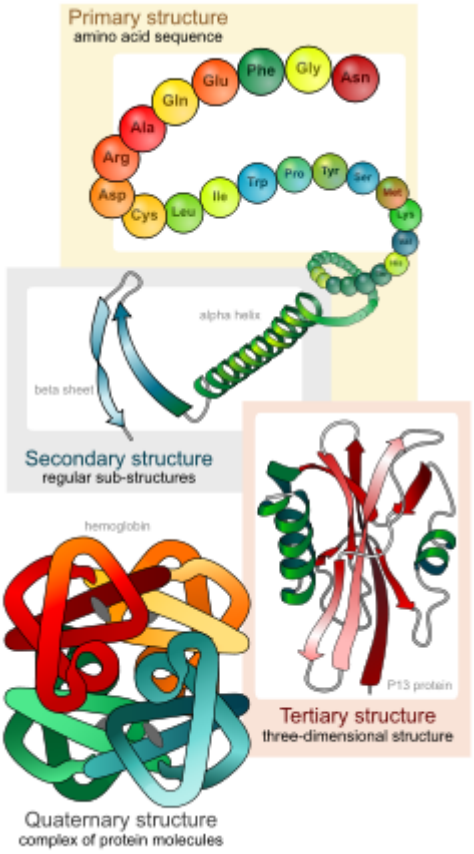


Figura 1.1: Hierarchical distribution of layers in protein structure

lution: protein structure is more conserved than sequence. However, not all the regions in a protein structure are equally conserved. It's been shown that functionally important amino acids, responsible of the interaction with other molecules, are more conserved than the rest of the protein structure [6]. Additionally, the structural core is more conserved than the surface [7]. The high conservation of the core enables the protein to maintain the global shape, while the surface is free to change (i.e.to mutate) some functional features [8]. These evolutionary mechanism are in accordance with the central *sequence* $\rightarrow$ *structure* $\rightarrow$ *function* paradigm that prevails in the protein evolution field.

Protein Structure Determination

Since in 1960, the Brithis biochemist John Kendrew determined the myoblobin structure [9], more than 37,000 different protein structures have been deposited in the Protein Data Bank (PDB) [10]. The PDB is a repository created in the 1970s

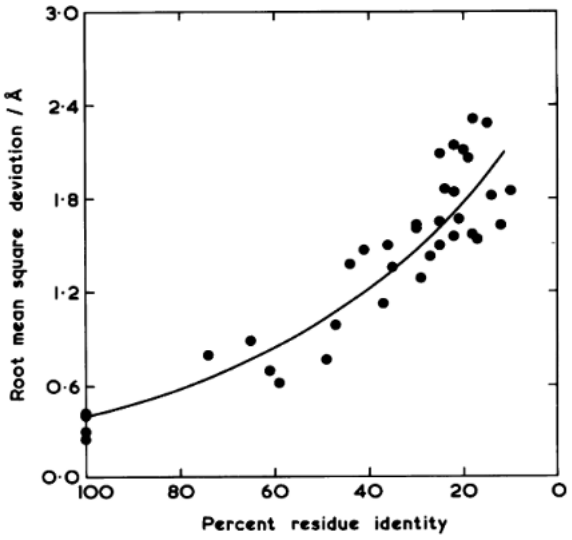


Figura 1.2: The original plot of the relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins. Figure extracted from [5]

with the aim of storing all the 3D protein structures and unifying their format. Figure 1.3a shows the variation of the number of deposited structures over the time. The number of PDB structures has significantly been increased over the last years thanks to initiatives such as the Protein Structure Initiative (PSI) [11] or the structural genomics [12]. The later, was born with the aim of determining the structure of all human proteins. However, soon after, they realized that the goal was unrealistic. Fortunately, the number of folds which represent the complete *fold space* observed in nature is much smaller that the number of proteins. Therefore, the current goal is to determine the structure of a representative set of proteins, that is, at least one protein per fold class. Once It is known the structure of one representative protein, and thanks to the *homology modeling* methods, It is usually feasible inferring the structure of other proteins belonging to the same fold class as we will explore further in the next section 1.1.1.

Several methods are currently used to experimentally determine the 3D structure of a protein. More than 99% of structures deposited in the PDB have been determine by the three main methods: X-ray crystallography **??**, nuclear magnetic resonance spectroscopy (NMR) and electron microscopy (EM)[REF] 1.3b. These methods provide experimental data that helps the scientist to elucidate the final structure of the protein. However, in most cases, the experimental data is not sufficient by itself to build an atomic model from scratch. Additional knowledge about the molecular structure most be added. For example, the preferred

geometry of atoms in a standard protein, the patterns of repulsion and attraction of amino acids, etc. All this information allows the building of the final model that is consistent with both the experimental data and the prior knowledge of the 3D geometry of the molecules. We next briefly explain the three aforementioned methods:

(a) **X-ray crystallography**. Currently, it is the most widely used method in protein structure determination. Almost 90% of the structures deposited in PDB come from X-ray crystallization (Figure 1.3b). In this method, X-rays fired at a crystal of the molecule are diffracted by the electron clouds of the atoms in the crystal, forming an unique pattern that is printed as a picture of the atomic density map. Subsequently, the diffraction pattern is combined with other physio-chemical knowledge of the protein, such as composition or atomic geometrical restrictions, in order to build the final 3D model [13].

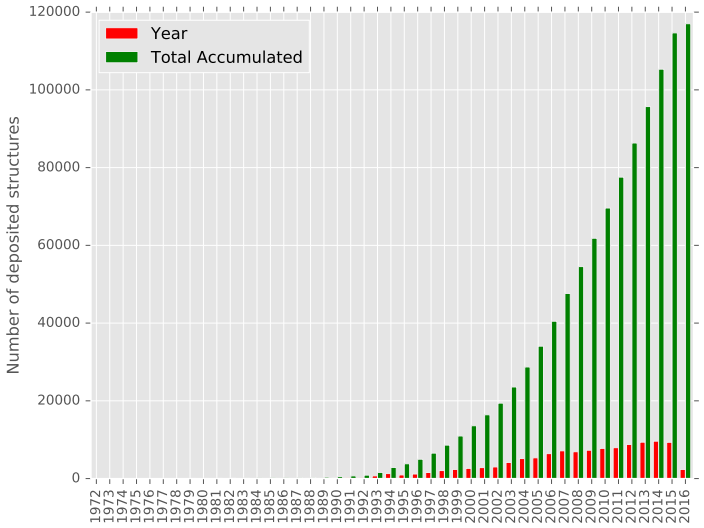
Before the X-ray exposition, it is then necessary a prior step of crystallization of the molecule. Unfortunately, the crystallization step introduces itself a great number of limitations. The flexibility of proteins is one of the these limitations. The flexible nature of proteins makes really difficult the creation of an accurate and homogeneous alignment of multiple molecules used to create the crystal. Another important limitation is the different conditions required for crystallizing each different molecule. These limitation are especially noteworthy in membrane proteins. Despite of nearly 30% of eukaryotic proteins are membrane proteins, only 604 unique membrane protein structures have been solved to date (data extracted from `http://blanco.biomol.uci.edu/mpstruc/`; date 21-03-2016). As a consequence, alternative innovative developments are needed to overcome the numerous obstacles associated with X-ray structure determination of membrane proteins [14].

The accuracy of the final atomic structure relies on the quality of the generated crystals. Two important measures of the accuracy of a crystallographic structure are its atomic *resolution*, which refers to the smallest separation between crystal lattice planes that is resolved in the diffraction pattern [15], and the *R-factor*, which measures how well the refined structure predicts the observed data [16].

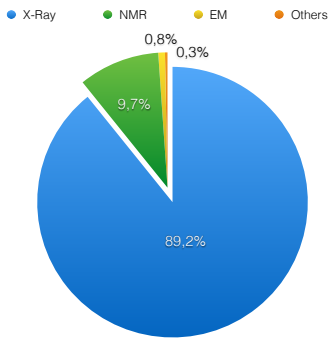
(b) NMR spectroscopy

(c) Electron microscopy

Protein Structure Prediction



(a)



(b)

Figura 1.3: a) Growth of released structures per year. Data extracted from PDB. b) Pie chart with the percetange of structures determined by the different methods. Data extracted from PDB.

Plot proteins predictable with homology vs. only structure determination.

1.1.2 Protein function

The major question in the protein biology field has been to understand the protein sequence, structure, function relationship. It is known that structure of a protein determines its biological function. However, different *regions* of the structure can perform semi-indepent functions from each other. These regions are referred to as **protein domains**. A domain is substructure produced by any part of polypedtide chain that can fold independently into a compact and stable structure [17, 18, 19]. Domains on average contain 80-250 residues [20]. Estimates of the number of domains per protein say that nire than 70% of procaryotik proteins and 80% of eukaryotic proteins include more than one domain [21, 22]. Among this multi-domain proteins, 95% of them contains only two to five protein domains [21]. Domains are not only the basic functional units of proteins, but also the evolutionary units of protein evolution. As proteins have evolved, domains have been modified and combined to build new proteins [23, 24]. Such is the importance of domains in protein evolution, that they have been included in current protein classification methods as one of the major classification parameters. Some of these domain classification methods such as SCOP [25] or CATH [26] are purely based on the structure, while others such as Pfam [27] or INTERPRO [28] include information about the function in their classification.

Domains, and consequently proteins, perform its biological activity by interacting with other molecules. Proteins can interact with other proteins, constructing a protein-protein complex, with ions or with small-molecules. The substance that is bound to the *target* protein is called the **ligand**, while the region of the protein where the ligand is binding is called ligand's *binding site*[2].

1.1.3 Protein-Ligand Interactions

The roles played by the ligands are diverse. Table X shows an example of the different functions that a small-molecule ligands can perform in a protein. Binding constants, allosteric and binding-site, induced fit model. Expandir. Imporatante.

[2]For simplicity, in this manuscript, unless otherwise indicated, the term ligand will only refer to small molecules ligands, while proteins ligands will be explicit named as protein-protein interactions

1.1.4 Protein-ligand prediction

1.2 Drug discovery

0	0
0	0

Taula 1.1: Prova de taula

1.2.1 subsection

Subsection

1.3 Drug discovery

Second

1.3.1 In-silico methods in drug-discovery

Subsection

1.4 Mycobacterium tuberculosis

1.4.1 Tuberculosis treatments and PPcs

1.5 Drug resistance in cancer

1.5.1 Cancer Treatment and drugs

CAPÍTOL 2

OBJECTIVES

CAPÍTOL 3

NANNOLYZE

CAPÍTOL 4

PREDICTING TARGETS IN MTB

CAPÍTOL 5

DRUG RESISTANCE IN CANCER

Figura 5.1: Example

Bibliografia

[1] A Kessel i N Ben-Tal. *Introduction to Proteins: Structure, Function, and Motion*. Chapman & Hall/CRC Mathematical and Computational Biology. CRC Press, 2010. ISBN: 9781439810729 (v. la pàg. 3).

[2] B. Alberts. *Molecular Biology of the Cell: Reference edition*. Molecular Biology of the Cell: Reference Edition v. 1. Garland Science, 2008. ISBN: 9780815341116 (v. la pàg. 4).

[3] Wolfgang Kabsch i Christian Sander. **?**Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features**?** A: *Biopolymers* 22.12 (1983), pàg. 2577 - 2637. ISSN: 1097-0282. DOI: 10.1002/bip.360221211 (v. la pàg. 4).

[4] K A Dill. **?**Dominant forces in protein folding.**?** A: *Biochemistry* 29.31 (1990), pàg. 7133 - 7155. ISSN: 0006-2960. DOI: 10.1021/bi00483a001 (v. la pàg. 4).

[5] Cyrus Chothial i Arthur M Lesk. **?**proteins**?** A: 5.4 (1986), pàg. 823 - 826 (v. les pàg. 4, 6).

[6] Buyong Ma et al. **?**Protein – protein interactions : Structurally conserved residues distinguish between binding sites and exposed protein surfaces**?** A: Track II (2003) (v. la pàg. 5).

[7] Rajkumar Sasidharan i Cyrus Chothia. **?**The selection of acceptable protein mutations**?** A: 2007 (2007) (v. la pàg. 5).

[8] Annabel E Todd, Christine A Orengo i Janet M Thornton. **?**Evolution of Function in Protein Superfamilies , from a Structural Perspective**?** A: (2001). DOI: 10.1006/jmbi.2001.4513 (v. la pàg. 5).

[9] J C KENDREW et al. **?**Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 [angst]. Resolution**?** A: *Nature* 185.4711 (feb. de 1960), pàg. 422 - 427 (v. la pàg. 5).

[10] Helen M Berman et al. **?**The Protein Data Bank**?** A: *Nucleic Acids Research* 28.1 (gen. de 2000), pàg. 235 - 242. DOI: 10.1093/nar/28.1.235 (v. la pàg. 5).

[11] John C Norvell i Jeremy M Berg. **?**Update on the Protein Structure Initiative**?** A: *Structure* 15.12 (des. de 2007), pàg. 1519 - 1522. ISSN: 0969-2126. DOI: http://dx.doi.org/10.1016/j.str.2007.11.004 (v. la pàg. 6).

[12] Opher Gileadi et al. **?**The scientific impact of the Structural Genomics Consortium : a protein family and ligand-centered approach to medically-relevant human proteins**?** A: (2007), pàg. 107 - 119. DOI: 10.1007/s10969-007-9027-2 (v. la pàg. 6).

[13] M S Smyth i J H J Martin. **?**x Ray crystallography**?** A: *Molecular Pathology* 53.1 (feb. de 2000), pàg. 8 - 14. DOI: 10.1136/mp.53.1.8 (v. la pàg. 7).

[14] Roslyn M Bill et al. **?**perspective Overcoming barriers to membrane protein structure determination**?** A: *Nature Biotechnology* 29.4 (2011), pàg. 335 - 340. ISSN: 1087-0156. DOI: 10.1038/nbt.1833 (v. la pàg. 7).

[15] Michael B Yaffe. **?**X-ray crystallography and structural biology**?** A: *Critical Care Medicine* 33.12 (2005). ISSN: 0090-3493 (v. la pàg. 7).

[16] A L Morris et al. **?**Stereochemical quality of protein structure coordinates.**?** A: *Proteins* 12.4 (abr. de 1992), pàg. 345 - 64. ISSN: 0887-3585. DOI: 10.1002/prot.340120407 (v. la pàg. 7).

[17] Jane S. Richardson. *Advances in Protein Chemistry Volume 34*. Vol. 34. 1981, pàg. 167 - 339. ISBN: 9780120342341. DOI: 10.1016/S0065-3233(08)60520-3 (v. la pàg. 9).

[18] Extracellular Proteins That i Modulate Cell-matrix Interactions. **?**Extracellular Proteins That**?** A: 266.23 (1991), pàg. 15 - 18 (v. la pàg. 9).

[19] D B Wetlaufer. **?**Nucleation, rapid folding, and globular intrachain regions in proteins.**?** A: *Proceedings of the National Academy of Sciences of the United States of America* 70.3 (1973), pàg. 697 - 701. ISSN: 0027-8424. DOI: 10.1073/pnas.70.3.697 (v. la pàg. 9).

[20] Suhail A Islam et al. **?**Identification and analysis of domains in proteins presented to identify domains in proteins**?** A: 8.6 (1995), pàg. 513 - 525 (v. la pàg. 9).

[21] Jung-Hoon Han et al. **?**The folding and evolution of multidomain proteins.**?** A: *Nature reviews. Molecular cell biology* 8.4 (2007), pàg. 319 - 330. ISSN: 1471-0072. DOI: 10.1038/nrm2144 (v. la pàg. 9).

[22] Cyrus Chothia et al. **?**Evolution of the protein repertoire.**?** A: *Science (New York, N.Y.)* 300.5626 (2003), pàg. 1701 - 3. ISSN: 1095-9203. DOI: 10.1126/science.1085371 (v. la pàg. 9).

[23] Christine Vogel et al. **?**Structure, function and evolution of multidomain proteins**?** A: *Current Opinion in Structural Biology* 14.2 (2004), pàg. 208 - 216. ISSN: 0959440X. DOI: 10.1016/j.sbi.2004.03.011 (v. la pàg. 9).

[24] G Apic, J Gough i S a Teichmann. **?**Domain combinations in archaeal, eubacterial and eukaryotic proteomes.**?** A: *Journal of molecular biology* 310.2 (2001), pàg. 311 - 325. ISSN: 0022-2836. DOI: 10.1006/jmbi.2001.4776 (v. la pàg. 9).

[25] Alexey G. Murzin et al. **?**SCOP: A structural classification of proteins database for the investigation of sequences and structures**?** A: *Journal of Molecular Biology* 247.4 (1995), pàg. 536 - 540. ISSN: 00222836. DOI: 10.1016/S0022-2836(05)80134-2 (v. la pàg. 9).

[26] Ca Orengo et al. **?**CATH - a hierarchic classification of protein domain structures**?** A: *Structure* March (1997), pàg. 1093 - 1109. ISSN: 09692126. DOI: 10.1016/S0969-2126(97)00260-8 (v. la pàg. 9).

[27] A Bateman et al. **?**The Pfam protein families database**?** A: *Nucleic Acids Research* 28.1 (2002), pàg. 276 - 280. ISSN: 0305-1048 (Print) 0305-1048 (Linking). DOI: gkd038[pii] (v. la pàg. 9).

[28] Sarah Hunter et al. **?**InterPro: The integrative protein signature database**?** A: *Nucleic Acids Research* 37.SUPPL. 1 (2009), pàg. 211 - 215. ISSN: 03051048. DOI: 10.1093/nar/gkn785 (v. la pàg. 9).