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

Contents

Index of figures XIII

List of tables XV

CHAPTER 1. INTRODUCTION **3**

1.1. Protein are essential molecules 3

1.1.1. Protein structure 4

1.1.2. Protein function 15

1.1.3. Protein-Ligand Interactions 16

1.2. Drug discovery 22

1.2.1. subsection 22

1.3. Drug discovery 22

1.3.1. In-silico methods in drug-discovery 22

1.4. Mycobacterium tuberculosis 23

1.4.1. Tuberculosis treatments and PPcs 23

1.5. Drug resistance in cancer 23

1.5.1. Cancer Treatment and drugs 23

CHAPTER 2. OBJECTIVES 25

CHAPTER 3. NANNOLYZE 27

CHAPTER 4. PREDICTING TARGETS IN MTB 29

CHAPTER 5. DRUG RESISTANCE IN CANCER 31

List of Figures

1.1. Hierarchical distribution of layers in protein structure 5

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] 6

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. 10

1.4. Workflow in comparative protein structure modeling. The figure has been extracted from [28] 12

1.5. Homology threshold curve as a function of alignment length. Data extracted from [31] 14

1.6. Schematic representation of three popular protein-ligand binding theories. 18

5.1. Example 31

List of Tables

1.1. List of the public Protein Modeling Tools 15

1.2. Prova de taula 22

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`

CHAPTER 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 carry many unanswered questions.

---

[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.

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*. Attached 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 organ-

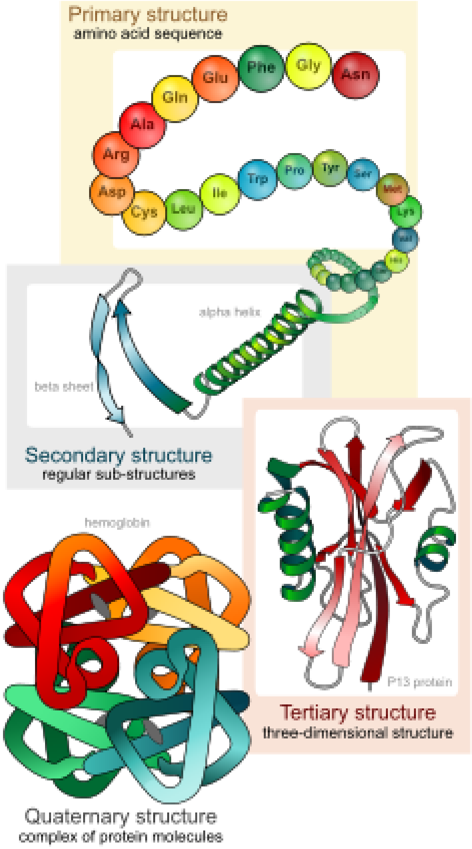

Primary structure
amino acid sequence
Glu
Phe
Gly
Asn
Gln
Ala
Arg
Asp
Cys
Leu
Ile
Trp
Pro
Tyr
Ser
Met
Lys
alpha helix
beta sheet
Secondary structure
regular sub-structures
hemoglobin
P13 protein
Tertiary structure
three-dimensional structure
Quaternary structure
complex of protein molecules

Figure 1.1: Hierarchical distribution of layers in protein structure

ism. 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 evolution: 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* → *structure* → *function* paradigm that prevails in the protein evolution field.

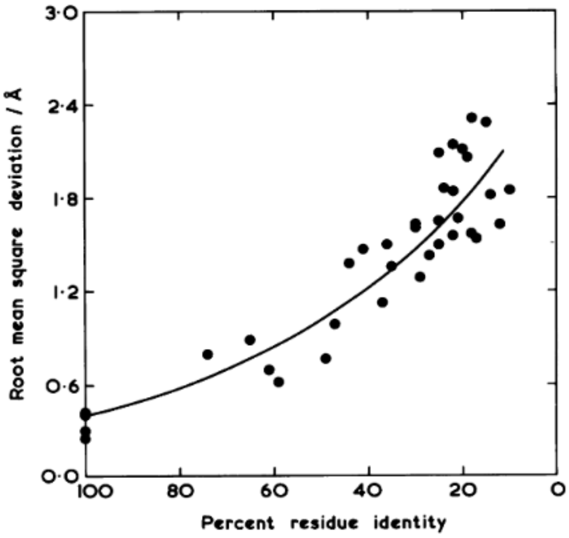


Figure 1.2: The original plot of the relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins. Figure extracted from [5]

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 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* methods1.1.1, 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) (Figure 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/`; March 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**. The NMR spectroscopy technique has been used for years to determine the structure of proteins. Currently, almost 10% of the structures deposited in PDB have been determined by this method (Figure 1.3b). In NMR spectroscopy, the protein is purified, placed in a strong magnetic field, and eventually probed with radio waves. The observed set of atomic resonances is then analyzed to retrieve a list of atomic nuclei that are close in the space. Similarly to X-ray crystallography, this set of restrains is subsequently used to build the structural model of the protein that contains the 3D conformation of each atom in the space [17].

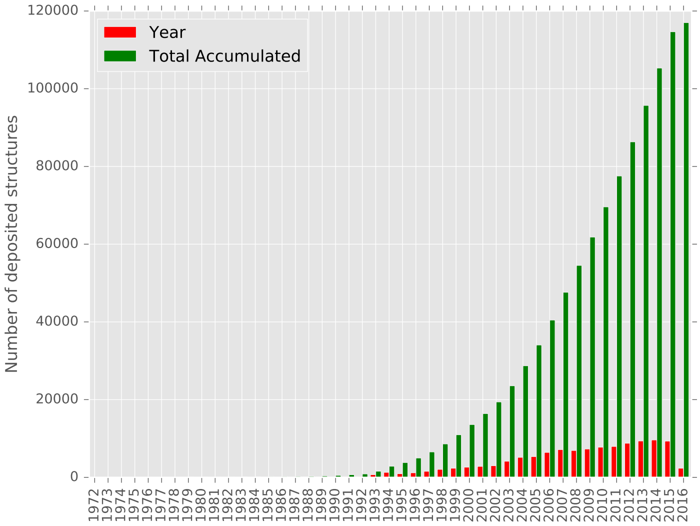
    NMR spectroscopy has a major advantage over X-ray crystallography: it provides information on proteins in solution. Therefore, this method is the main method for studying the atomic structure of highly flexible proteins. A standard NMR structure includes an ensemble of protein structures, all of them being consistent with the experimentally observed set of restraints. The ensemble of structures are very similar in those regions with strong restrains, less constrained regions of the proteins, on the other hand, show less agreement in the generated models. These lack of restriction areas are presumably the flexible regions of the protein since they do not provide a strong signal in the experiment.

    A big limitation in comparison with X-ray crystallography is its applicability: this technique is usually limited to proteins smaller than 35 kDa. Moreover, NMR can only be applied to soluble proteins that do not aggregate and are stable during the NMR experiment [17, 18]. NMR is also inherently insensitive and milligram amount of proteins are required [18]. All these limitations have hampered the broader use of this technique narrowing down the cases where this method is fruitful.

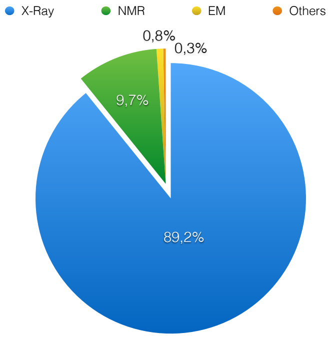
(c) **Electron microscopy** methods. EM methods are emerging as a very versatile tool for determining the structure of large macromolecular complexes. To date, less than 1% of proteins in PDB have been determined by EM methods (Figure 1.3b). However, in recent years there has been dramatic increase in the number of complexes determined by EM technologies. The *revolution* in the structural biology field is perfectly manifested by the cryo-electron microscopy (cryoEM) method: in 2015 alone, cryoEM was used to map the structures of more than 100 different molecules [19]. In cryoEM a beam of electrons is fired at a frozen protein solution. The emerging scattered electrons pass through a lens to create a magnified image on the detector, and the structure can then be deduced afterwards.

The utility of cryoEM and others EM tools lies on the fact that it allows the observation of molecules that have not been fixed in any way, showing them in their native environment. This is the opposite of the crystallization step in X-ray crystallography, which many times hampers the success of the whole procedure. CryoEM have been traditionally used in large molecules such as ribosomes[20] or the V-ATPase[21], but they have also shown their potential in small membrane proteins[22] and medically important proteins[23].

However, there are still a room for further improvement in EM technologies. Despite of recent advances in the resolution, most of the cryoEM structures have lower resolution than the X-ray ones. Furthermore, there are numerous technical unsolved problems that need to be addressed to make easier its standardization and systematical application. Finally, the high prize of cryoEM experiments are many times slowing Its spread and therefore, there is a need to reduce cost in order to make it globally accessible.



(a)



(b)

Figure 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.

Protein Structure Prediction

Despite of the advances in methods for protein structure determination, most of the known proteins lack of structure in the PDB. There are 550,740 annotated and curated protein sequences in UniProtKB (`http://www.uniprot.org`, April 2016). However, only 4% of them (23,195 different protein sequences) have a link to a PDB structure. Therefore, there is a gap between the number of known protein sequences and the number of determined structures: the *sequence-structure gap*. Computational methods for structure determination are helping to bridge this gap. The prediction of the 3D structure of a protein from its amino acid sequence has always been one of the most desirable goals in computational biology. It would save a lot of resources, and it would set a milestone in the structural biology field. Unfortunately, we are still far from being able to predict the structure of any protein from its primary sequence. Overall, four different approaches are commonly used. The first, and most extensively used, is the *homology* or *comparative* modeling, that uses similar experimentally determined structures to model the structure of the protein of interest 1.1.1. Second, *fold recognition* and *threading* methods are used to model protein structures with low similarity to known protein structures [24, 25]. Third, *de novo* or *ab initio* methods make their predictions by combining the principles of physics that rule protein folding, with information derived from known structures but without relying in any type of similarity or evolutionary relationship to known folds [26]. Finally, the *integrative* or *hybrid* methods combines different computational and/or experimental sources to perform the structure prediction [27].

Homology modeling

This type of protein structure prediction methods exploits the evolutionary relationship between the *target* protein (i.e. protein being modeled) and the *template(s)* with known experimental structure. They are based on the biological premise that evolutionary related sequences tend to have similar 3D structures. Figure 1.4 shows the regular steps in comparative protein structure modeling:

1. **Identification of suitable template structures related to the target protein**. This step consist on a search for similar sequences with known 3D structure. This task is easy when a close homologue of the target protein has been solved. Initiatives such as the PSI[11] are helping in this issue increasing the number of modellable proteins. However, there are still many

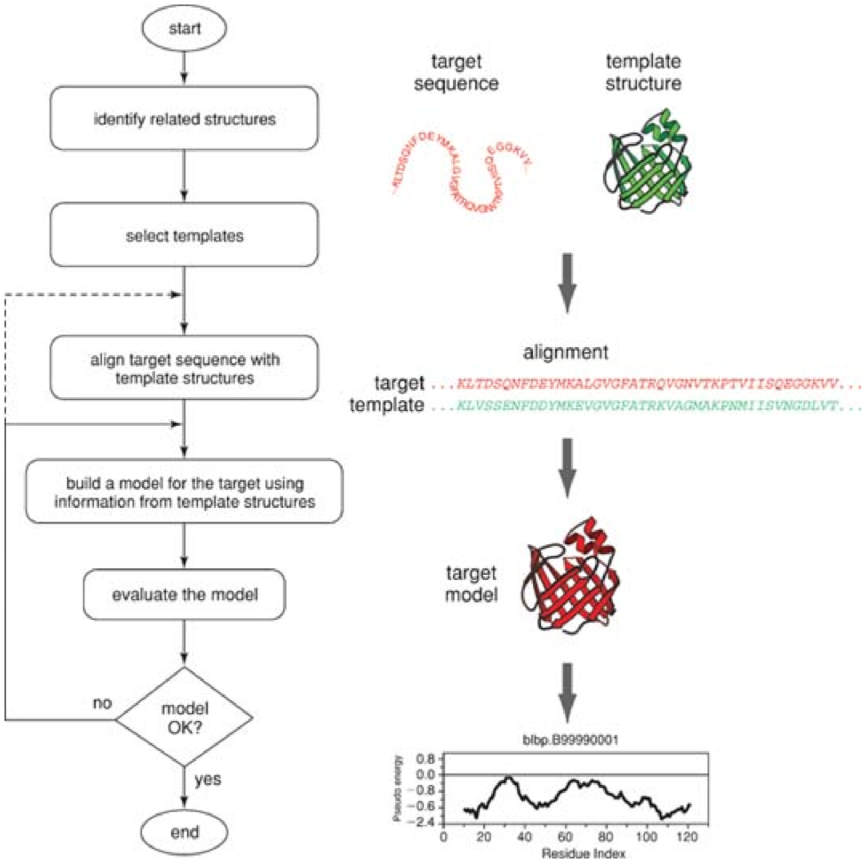

start
identify related structures
select templates
align target sequence with template structures
build a model for the target using information from template structures
evaluate the model
no
model OK?
yes
end
target sequence
template structure
alignment
target ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
template ...KLVSSENFDDYMKEVGVGFATRKVAGMAKPNMIISVNGDLVT...
target model
blbp.B99990001
Pseudo energy
0.8
0.0
-0.8
-0.6
-2.4
0
20
40
60
80
100
120
Residue Index

Figure 1.4: Workflow in comparative protein structure modeling. The figure has been extracted from [28]

proteins with lack of homologous proteins in PDB. In these cases, alternative methods such as *ab initio* modeling should be used.

2. **Alignment between the target and the template(s) sequence(s)**. The sequence identity of the target-template alignment is the most frequently used measure for similarity. Consequently, the sequence identity is also a good predictor of the final 3D model quality. The overall accuracy of models calculated form alignments with sequence identity higher than 40% is usually good (i.e. RMSD[2] lower than 2.0Å). In the 30%-40% identity range, errors can be more severe and are often locate in loops and highly flexible

---

[2] Root Mean Square Deviation is the measure of the average distance between the atoms of two superimposed proteins. Equation $RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$ where δ_i is the distance between the N_i pair of equivalent atoms (usually the Cα).

regions. Below the 30% of sequence similarity, often referred to as *twilight region*, serious errors occurs including the basic fold being mis-predicted [29, 30]. Figure 1.5 shows an empirical threshold for homology modeling extracted from [31]. The region below the curve covers those cases where the alignment does not carry enough information to model the 3D structure, while area above the threshold curve, include those cases where homology modeling is appropriate.

3. **Modeling of the structurally conserved regions and prediction of the structurally variable regions.**. There are different algorithms to assign the spatial coordinates of the target protein using the template(s)-target alignment information. Highly conserved regions are generally well modeled, while those regions with insertions or gaps are more prone to include errors and sub-optimal atomic orientations.

4. **Refinement of the initial model.** In this step the model is refined to idealize bond geometry and to remove errors that may have been introduced in the modeling step. The refinement process pursues the free energy minimization of the generated 3D protein model. Many different algorithms have been applied to perform the minimization step: molecular mechanics force fields [32], molecular dynamics [33], Monte Carlo [34] and Genetic Algorithms[35] are just several examples of the multiple approaches applied in this issue.

5. **Evaluation of the model(s).** Model evaluation seeks for the identification of the different errors that might have occurred during the modeling process. Multiple methods have been developed to assess the quality of a 3D model. In fact, 3D model assessment has a very such productive field for many years that it was included in the seventh edition of the CASP experiment [36]. The general question of how accurate is a model can be reformulated in several specific questions:

    I *Is the selected fold correct?* The fold assessment consist of deciding whether a given protein model has the right fold. Residue-based combined accessible surface and distance-dependent scoring functions have shown the best performance in this task [37].

    II *How do we select the best model among the set of decoys or alternative solutions?* Several models can be generated by making changes in the template-target alignment, by selecting different template(s) structure(s) or by using different seeds in the refinement non-deterministic algorithms. Atom-based distance-depend scoring functions have proved to be useful for this particular task in some cases[38]. However, there is not a gold standard for ranking the generated 3D models. Thus, the model selection eventually relies on the expertise of the person running the modeling.

III *Which is the overall accuracy of the model? Which is the accuracy of the model in a particular region of the model?* These questions can be addressed by defining scores that correlates with the RMSD after superimposing a model and its native structure. Numerous scoring functions have been implemented to address this issue. Some of them, the *physics-based scoring functions*, attempt to approximately calculate the atomic interaction energies in the system. These scoring functions usually encode a set of parameters that describes the energy of a system of particles. Examples of these scoring functions are AMBER[39], CHARMM[40] or MM-PDBSA[41]. Differently the *knowledge-based potentials* or *potentials mean force* are scoring functions derived from an analysis of empirical information. Although the obtained scores are often considered approximations of the free energy, this physical interpretation is incorrect [42]. Nevertheless, since they frequently correlate with the actual free energy differences, they have been broadly used with significant success [43, 44, 45].

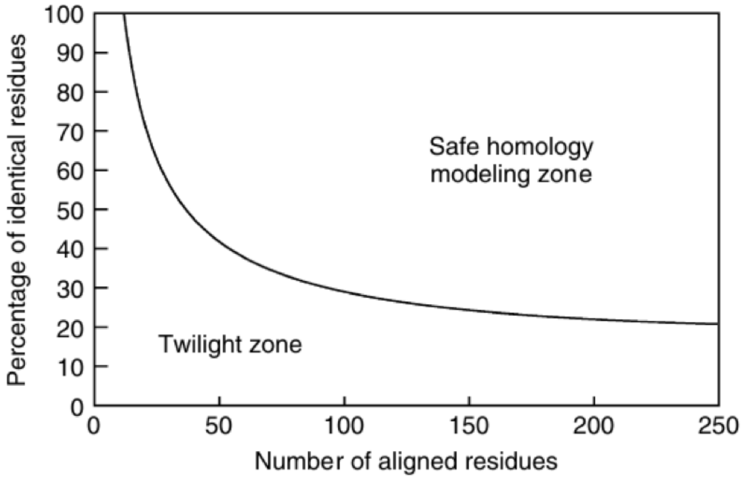


Figure 1.5: Homology threshold curve as a function of alignment length. Data extracted from [31]

The application of comparative modeling is limited by several aspects. The first one, is the availability of a suitable template. Despite of the efforts made to determine at least one structure per known fold [11], divergences between the template and the target hampers the modeling of a correct 3D structure. In fact, models based on aligment below 30% have been shown to be unsatisfactory 1.5. The lack of template problem is even more noticeable in membrane proteins. The limited number available known membrane proteins structures makes its modeling an extremely difficult task. However, the high value of these proteins in diverse therapeutic areas [46, 47], is fostering the development of specific membrane protein modeling methods [48]. Another aspect that restricts the success of homology modeling is the innate flexibility of proteins. Highly flexible regions are more difficult to model and consequently are more prone to errors than more rigid parts of the structure. Despite of these limitations, homology modeling has been successfully applied to many proteins and its currently the main approach to computationally model the 3D structure of proteins[3].

There are many computational methods for predicting the 3D structure of proteins. Table **??** shows some of the most famous ones alongside its websites and references. Each of these methods have their own strengths and weakness and none of them clearly outperforms the others for every case. In my case, I have used Modeller because it is the one we are more familiar with.

Modeling Tool	**Website**	**Reference(s)**
Modeller	`https://salilab.org/modeller/`	[28, 51]
SwissModel	`http://swissmodel.expasy.org`	[52]
HHPred	`http://toolkit.tuebingen.mpg.de/hhpred`	[53]
I-Tasser	`http://zhanglab.ccmb.med.umich.edu/I-TASSER/`	[54, 55, 56]
Rosetta	`http://robetta.bakerlab.org/`	[57]
RaptorX	`http://raptorx.uchicago.edu/`	[58]
3DJIGSAW	`http://bmm.crick.ac.uk/~3djigsaw`	[59]
WhatIf	`http://swift.cmbi.ru.nl/whatif/`	[60]

Table 1.1: List of the public Protein Modeling Tools

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 [61, 62, 63]. Domains on average contain 80-250 residues [64]. Estimates of the number of domains per protein say that more than 70% of procaryotik proteins and 80% of eukaryotic proteins include more than one domain [65, 66]. Among this multi-domain proteins, 95% of them contains only two to five protein domains [65]. 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 [67, 68]. 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 [69] or CATH [70] are purely based on the structure, while others such as Pfam [71] or INTERPRO [72] include information about the function in their classification.

---

[3] For a comprehensive review of homology modeling methods, applications and limitationsplease consider [49, 50]

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*[4]. The next section is focused on protein-compound interaction and it presents the basis for all the work developed during the thesis.

1.1.3. Protein-Ligand Interactions

The roles played by the protein ligands are diverse. Catalysis of enzyme substrates, regulation of the protein activity, cellular communication or defense from external attackers are just few some examples of the multiple functions that small-molecule ligands perform in living organisms. All these functions are performed by small-molecules that selectively bind their *target* proteins. However, given the vast amount of proteins and small-molecule ligands in the cytoplasm, how do the small-molecule ligands select their protein targets? There has been several protein-ligand binding theories. In the *Lock and key Theory* [73], Emil Fischer proposed a system where the binding sites of enzymes are rigid and pre-adjusted geometrically to the natural substrate 1.6a. This theory became widely accepted for years. Nevertheless, during subsequent years, evidence started to accumulate suggesting that the binding sites of proteins do not match perfectly their ligands, but rather the binding process triggers some conformational changes in the enzyme. Therefore the obsolete Lock and key model was replaced by the *Induced fit theory* [74]. The induced fit theory proposes that initially enzymes do not perfectly match their substrate geometrically. However, the binding-process triggers a set of conformational changes in the protein binding site that improves the match 1.6b. More recently, another theory called the *Monod-Wyman-Changeux model or MWC model* came up[75]. This theory contends that proteins are able to shift spontaneously between multiple conformations called *substates*[76, 77]. This model could also explain *allostery*, a phenomenon in which the binding of the molecule to the catalytic site is affected the binding of other ligand to a different site. This theory has undergone some changes and the current accepted theory posits that ligands bind preferentially to one of the conformation sampled spontaneously by the protein, and therefore stabilizes it. It means that, by changing the protein's energy landscape, ligands change a less favorable conformation into the most favorable one. This model does not necessarily refute the induce fit theory since in many cases, the restrains applied by the ligand on the binding site is expected to induce some conformational changes that will further stabilize the interaction [78, 79].

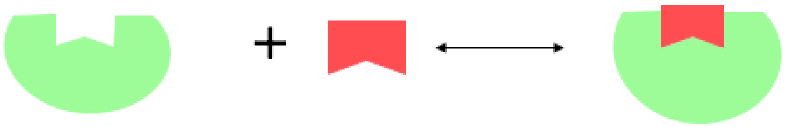
---

[4] 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

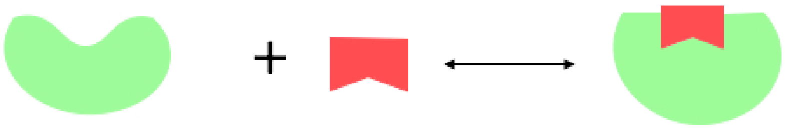
Protein-ligand binding energetics

The high variety of functions that ligands perform by binding to proteins its also reflected in the diversity in its binding affinity. Binding energies usually range from -2.5kcal/mol to -22kcal/mol [80]. The binding strength displayed by proteins matches the biological goal of the binding. For instance, ligands involved in protein communication tend to bind weakly to enable a quick state switch. Cofactors binding to enzymes, on the other hand, tend to bind strongly to their targets. The negative sign of the values reflects that is a favorable binding that releases energy: *the binding free energy*. This energy can be measured experimentally, thorough the equilibrium constant of the binding, or it can be calculated computationally. Equations 1.1 and 1.2 shows, under thermodynamic equilibrium conditions, the relationship between the Gibbs free energy or binding affinity and the equilibrium constant of the binding. R is the ideal gas constant [ref], T is the temperature , $[C]$ the complex concentration and $[P]$ and $[L]$ the protein and ligand concentration respectively.

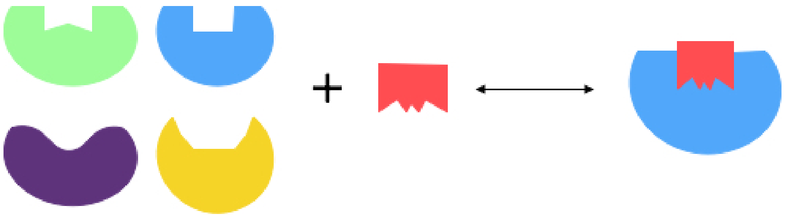
$$\triangle G = -RTlnK_{bind} \tag{1.1}$$



(a) Fischer's *lock and key model*. The protein is represented in green and the ligand in red. The ligand's binding site of the protein matches the ligand perfecltly.



(b) Koshland's *induced fit model*. The protein is represented in green and the ligand in red. The overall shape of the ligand matches the binding site. The ligand bindings causes some conformational changes that improves the interaction.



(c) *MWC model's* representation. The protein changes its conformation constantly (one color per conformation), with at least one these conformation matching the ligand. Its binding triggers some conformational changes that improves the protein's energy landscape.

Figure 1.6: Schematic representation of three popular protein-ligand binding theories.

$$Kbind = \frac{[C]}{[P] * [L]} \tag{1.2}$$

These equations show that the binding free energy can be measured experimentally. However, in many cases the experimental measurement are unfeasible or very difficult due to technical problems. Additionally, the expenses associated with these experiments often restricts its broader application. In such cases, computational methods to calculate the free binding energy are needed. The calculation of binding free energy have acquired a remarkable importance in the drug discovery field where the calculation of ligand-target affinity is crucial for the preclinical phase 1.2. Unfortunately, its calculation its a extremely challenging task. The main points that should be addressed to accurately calculate the binding free energy are:

1. **The free energy of binding 1.1 is the difference of two large energies**. The energy of the complex (E_{pl}) and the energy of the unbound partners ($E_p + E_l$) 1.3:
$$\triangle G_bind = E_pl - (E_p + E_l) \tag{1.3}$$
2. **There are two opposite and complex energies driving the process**. The binding enthalpy ($\triangle H_{bind}$) and the loss of entropy of both the ligand and the protein ($\triangle S_{bind}$):
$$\triangle G_bind = \triangle H_{bind} - T \triangle S_{bind} \tag{1.4}$$
3. **Non explicit representation of the energetic interactions of the system**. Small-molecule binding events on a protein cavity implies the displacement of solvent molecules (usually water molecules). The energy generated by the this exclusion of water molecules is the main driving force in ligand-protein binding [81]. Unfortunately, explicit representation and simulation of all the forces involved this event is computationally unfeasible. A popular approach to model is to use implicit solvent theories [82, 83, 84], where the water molecules are represented as a continuous medium instead of individual explicit molecules. The implicit solvation model is justified in liquids, where the potential of mean force are applied to approximate the behavior of many highly dynamic solvent molecules. However, there could be other medias with specific solvation or dielectric properties that are continuous, but not necessarily uniform, since their properties can be described by different analytic functions [85]. Within the most famous implicit models we

> can find those based on those based on Boltzmann theory (PB) [86] and those based on Generalized-Born (GB) approximation [87].
>
> Hydrogen bonds and salt bridges between the ligand and the protein can also be a source of free energy gain upon ligand binding. This energy gain comes from the displacement of water molecules bound to the protein. The net gain of energy upon hydrogen bond is around 1-2 kcal/mol. Some scoring functions treat all hydrogen bonds equally, while others, distinguish between neutral and charged ones. Other energies that could be modeled and that contribute to the binding affinity calculations are those generated by interactions with metal ions [88]. However, because there can be a covalent component in this type of interactions, its overall binding energy contribution is difficult to model. Finally, nonspecific van der Waals and hydrophobic interactions are also included in some methods as additional energy contributors to the overall free energy of binding [89].

One of the main applications of binding free energy calculation is predicting whether a ligand is binding a particular protein target. In other words, given the predicted binding free energy determine whether a specific compound targets a specific protein binding site. In the next section we will explore further these and other approaches aiming at protein-compound interaction prediction.

Protein-ligand prediction

[5]The importance of ligand-protein interaction prediction is reflected by the large number of available methods that use multiple different approaches [90, 91]. We can distinguish between *free structure* methods (i.e. methods that do not rely on the protein structure to perform its predictions) and *structure-based* methods.

Free structure methods do not require the protein structure to perform their predictions. They usually use prior knowledge on protein compound interactions, to further extend the interactions to new and unseen compounds. The development of these methods can be split in two phases. The first step consist on the creation of a predictive model that uses a collection(s) of protein-compound interactions to learn hidden relationships between compound and their protein targets. In the second step, these predictive models are used to extrapolate this knowledge to new and unseen compounds (or targets). The extrapolation relies on different types of compound or protein similarity. Knowledge-based free structure methods have been assisted by the advent of new and powerful *high-thorughpout screening methods* (HTS) that allowed the creation of large computational compound-protein databases such as ChEMBL [92],Therapeutic Target Database (TTD) [93], Binding MOAD [94], BindingDB [95], PubChem [96, 97] or ZINC [98]. The recent growth of these collections is accordingly improving these method's accuracy and coverage. Moreover, since they do not rely on protein structure they can be theoretically applied to any protein or any compound. Nevertheless, free structure methods do not provide detailed information about the ligand-protein relationship. Information such as binding localization, type of interaction (i.e. allosteric, on-target or off-target) or predicted free energy of binding; that is absolutely essential in the drug discovery process. Consequently ,free structure methods are mostly employed in very early stages of the drug discovery pipeline.

---

[5] In chapter [REF] we present nAnnolyze, a method for protein-ligand interaction prediction. In the introduction of the presented manuscript, there is a discussion of the current state-of-the-art methods in protein-ligand interaction prediction. Therefore, this section is focused in explaining the classification, underlying basics and (dis)advantages of the different approaches.

Structure based target prediction methods use the protein structure to determine whether a small-molecule interact with a protein target. Docking methods have traditionally dominated the structure-based target prediction field. Virtual docking consist on predicting the preferred orientation of one molecule (the ligand) to a second (the protein) one. The process of finding the best orientation of molecule (i.e. its *pose*) to its protein target is not simple, since several entropic, enthalpic and environmental factors have an impact on the interactions between them 1.1.3. The underlying idea of this approach is to generate a comprehensive set of ligand-protein conformations, and then to rank them according to a specific scoring function [99]. The importance of virtual docking methods is not only reflected by the large number of published methods (AutoDock [100, 101], DOCK [102], FLEXX [103], GOLD [104] and GLIDE [88] among others), but also by their success in drug discovery applications [105, 106, 107][6]. However, virtual docking methods also have some limitations. The most apparent one is that they do rely on protein structure. As explained above 1.1.1 the coverage of the human structural proteome is below 40%. Thus, some of the most interesting targets in drug discovery lack of experimentally determined 3D structure. In addition to these structurally inherent problems and despite of some massive applications [109], virtual docking methods are still computationally very expensive. Additionally, they need the prior knowledge of the binding localization in the protein surface which many times is unknown before the screening. To overcome the computational limitations, new structure-based methods use the so-called *comparative docking* approaches that solely rely on structural comparisons, both of compounds and protein targets, to infer new interactions [REFS]. Chapter [REF nAnnolyze] presents nAnnolyze, a method to predict ligand-target interactions using a comparative docking approach. In this chapter Its further discussed the applications and limitations of these type of approaches.

---

[6] For a comprehensive review of virtual docking methods an applications please consider [108]

Resumen Prediction?

1.2. Drug discovery

subsectionTarget prediction in drug discovery process

0	0
0	0

Table 1.2: 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

CHAPTER 2

OBJECTIVES

CHAPTER 3

NANNOLYZE

CHAPTER 4

PREDICTING TARGETS IN MTB

CHAPTER 5

DRUG RESISTANCE IN CANCER

Figure 5.1: Example

Bibliography

[1] A Kessel and N Ben-Tal. *Introduction to Proteins: Structure, Function, and Motion*. Chapman & Hall/CRC Mathematical and Computational Biology. CRC Press, 2010. ISBN: 9781439810729 (cit. on p. 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 (cit. on p. 4).

[3] Wolfgang Kabsch and Christian Sander. «Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features». In: *Biopolymers* 22.12 (1983), pp. 2577–2637. ISSN: 1097-0282. DOI: `10.1002/bip.360221211` (cit. on p. 4).

[4] K A Dill. «Dominant forces in protein folding.» In: *Biochemistry* 29.31 (1990), pp. 7133–7155. ISSN: 0006-2960. DOI: `10.1021/bi00483a001` (cit. on p. 4).

[5] Cyrus Chothial and Arthur M Lesk. «proteins». In: 5.4 (1986), pp. 823–826 (cit. on pp. 5, 6).

[6] Buyong Ma et al. «Protein – protein interactions : Structurally conserved residues distinguish between binding sites and exposed protein surfaces». In: Track II (2003) (cit. on p. 5).

[7] Rajkumar Sasidharan and Cyrus Chothia. «The selection of acceptable protein mutations». In: 2007 (2007) (cit. on p. 5).

[8] Annabel E Todd, Christine A Orengo, and Janet M Thornton. «Evolution of Function in Protein Superfamilies , from a Structural Perspective». In: (2001). DOI: `10.1006/jmbi.2001.4513` (cit. on p. 5).

[9] J C KENDREW et al. «Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 [angst]. Resolution». In: *Nature* 185.4711 (Feb. 1960), pp. 422–427 (cit. on p. 6).

[10] Helen M Berman et al. «The Protein Data Bank». In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242. DOI: `10.1093/nar/28.1.235` (cit. on p. 6).

[11] John C Norvell and Jeremy M Berg. «Update on the Protein Structure Initiative». In: *Structure* 15.12 (Dec. 2007), pp. 1519–1522. ISSN: 0969-2126. DOI: `http://dx.doi.org/10.1016/j.str.2007.11.004` (cit. on pp. 6, 11, 14).

[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». In: (2007), pp. 107–119. DOI: `10.1007/s10969-007-9027-2` (cit. on p. 6).

[13] M S Smyth and J H J Martin. «x Ray crystallography». In: *Molecular Pathology* 53.1 (Feb. 2000), pp. 8–14. DOI: `10.1136/mp.53.1.8` (cit. on p. 7).

[14] Roslyn M Bill et al. «perspective Overcoming barriers to membrane protein structure determination». In: *Nature Biotechnology* 29.4 (2011), pp. 335–340. ISSN: 1087-0156. DOI: `10.1038/nbt.1833` (cit. on p. 7).

[15] Michael B Yaffe. «X-ray crystallography and structural biology». In: *Critical Care Medicine* 33.12 (2005). ISSN: 0090-3493 (cit. on p. 7).

[16] A L Morris et al. «Stereochemical quality of protein structure coordinates.» In: *Proteins* 12.4 (Apr. 1992), pp. 345–64. ISSN: 0887-3585. DOI: `10.1002/prot.340120407` (cit. on p. 8).

[17] Gerhard Wider. «Structure Determination of Biological Macromolecules in Solution Using NMR spectroscopy». In: 1294 (2000), pp. 1278–1294 (cit. on p. 8).

[18] David G Gadian et al. «Mechanisms of Secondary Brain Damage: Current State». In: ed. by Alexander Baethmann, Oliver Kempski, and Ludwig Schürer. Vienna: Springer Vienna, 1993. Chap. NMR Spectr, pp. 1–8. ISBN: 978-3-7091-9266-5. DOI: `10.1007/978-3-7091-9266-5{\_}1` (cit. on p. 8).

[19] Kicking Up et al. «THE REVOLUTION WILL NOT BE CRYSTALLIZED». In: (2015), pp. 7–9 (cit. on p. 8).

[20] Heena Khatter et al. «Structure of the human 80S ribosome». In: *Nature* 520.7549 (Apr. 2015), pp. 640–645. ISSN: 0028-0836 (cit. on p. 9).

[21] Jianhua Zhao, Samir Benlekbir, and John L Rubinstein. «Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase». In: *Nature* 521.7551 (May 2015), pp. 241–245. ISSN: 0028-0836 (cit. on p. 9).

[22] Maofu Liao et al. «Structure of the TRPV1 ion channel determined by electron cryo-microscopy». In: *Nature* 504.7478 (Dec. 2013), pp. 107–112. ISSN: 0028-0836 (cit. on p. 9).

[23] Xiao-chen Bai et al. «An atomic structure of human [ggr]-secretase». In: *Nature* 525.7568 (Sept. 2015), pp. 212–217. ISSN: 0028-0836 (cit. on p. 9).

[24] D T Jones, W R Taylort, and J M Thornton. «A new approach to protein fold recognition». In: *Nature* 358.6381 (July 1992), pp. 86–89 (cit. on p. 11).

[25] J U Bowie, R Luthy, and D Eisenberg. «A method to identify protein sequences that fold into a known three-dimensional structure». In: *Science* 253.5016 (July 1991), pp. 164–170 (cit. on p. 11).

[26] Jooyoung Lee, Sitao Wu, and Yang Zhang. «From Protein Structure to Function with Bioinformatics». In: ed. by Daniel John Rigden. Dordrecht: Springer Netherlands, 2009. Chap. Ab Initio, pp. 3–25. ISBN: 978-1-4020-9058-5. DOI: `10.1007/978-1-4020-9058-5{\_}1` (cit. on p. 11).

[27] Daniel Russel et al. «Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies». In: *PLoS Biology* 10.1 (Jan. 2012), e1001244. ISSN: 1544-9173. DOI: `10.1371/journal.pbio.1001244` (cit. on p. 11).

[28] Narayanan Eswar et al. «Comparative protein structure modeling using MODELLER.» en. In: *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]* Chapter 2 (Dec. 2007), Unit 2.9. ISSN: 1934-3663. DOI: `10.1002/0471140864.ps0209s50` (cit. on pp. 12, 15).

[29] D Baker. «Protein structure prediction and structural genomics». In: *Science* 294 (2001), pp. 93–96. ISSN: 00368075. DOI: `10.1126/science.1065659` (cit. on p. 13).

[30] Su Yun Chung and S Subbiah. «A structural explanation for the twilight zone of protein sequence homology». In: *Structure* 4.10 (Oct. 1996), pp. 1123–1127. ISSN: 09692126. DOI: `10.1016/S0969-2126(96)00119-0` (cit. on p. 13).

[31] C Sander and R Schneider. «Database of homology-derived protein structures and the structural meaning of sequence alignment.» In: *Proteins* 9.1 (Jan. 1991), pp. 56–68. ISSN: 0887-3585. DOI: `10.1002/prot.340090107` (cit. on pp. 13, 14).

[32] Evandro Ferrada and Francisco Melo. «Nonbonded terms extrapolated from nonlocal knowledge-based energy functions improve error detection in near-native protein structure models». In: *Protein Science* 16.7 (2007), pp. 1410–1421. ISSN: 1469-896X. DOI: `10.1110/ps.062735907` (cit. on p. 13).

[33] A Fiser, R K Do, and A Sali. «Modeling of loops in protein structures.» In: *Protein science : a publication of the Protein Society* 9.9 (Sept. 2000), pp. 1753–73. ISSN: 0961-8368. DOI: `10.1110/ps.9.9.1753` (cit. on p. 13).

[34] A Kidera. «Enhanced conformational sampling in Monte Carlo simulations of proteins: application to a constrained peptide.» In: *Proceedings of the National Academy of Sciences of the United States of America* 92.21 (Oct. 1995), pp. 9886–9889. ISSN: 0027-8424 (cit. on p. 13).

[35] D.B. McGarrah and R.S. Judson. «Analysis of the genetic algorithm method of molecular conformation determination». In: *Journal of Computational Chemistry* 14.11 (Nov. 1993), pp. 1385–1395. ISSN: 0192-8651. DOI: `10.1002/jcc.540141115` (cit. on p. 13).

[36] Domenico Cozzetto et al. «Assessment of predictions in the model quality assessment category». In: *Proteins: Structure, Function, and Bioinformatics* 69.S8 (2007), pp. 175–183. ISSN: 1097-0134. DOI: `10.1002/prot.21669` (cit. on p. 13).

[37] Francisco Melo, Roberto Sánchez, and Andrej Sali. «Statistical potentials for fold assessment.» In: *Protein science : a publication of the Protein Society* 11.2 (Mar. 2002), pp. 430–48. ISSN: 0961-8368. DOI: `10.1002/pro.110430` (cit. on p. 13).

[38] R Samudrala and J Moult. «An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction.» In: *Journal of molecular biology* 275.5 (Feb. 1998), pp. 895–916. ISSN: 0022-2836. DOI: `10.1006/jmbi.1997.1479` (cit. on p. 13).

[39] DAVID A CASE et al. «The Amber Biomolecular Simulation Programs». In: *Journal of computational chemistry* 26.16 (Dec. 2005), pp. 1668–1688. ISSN: 0192-8651. DOI: `10.1002/jcc.20290` (cit. on p. 14).

[40] Bernard R. Brooks et al. «CHARMM: A program for macromolecular energy, minimization, and dynamics calculations». In: *Journal of Computational Chemistry* 4.2 (1983), pp. 187–217. ISSN: 0192-8651. DOI: `10.1002/jcc.540040211` (cit. on p. 14).

[41] Federico Fogolari, Alessandro Brigo, and Henriette Molinari. «Protocol for MM/PBSA molecular dynamics simulations of proteins.» In: *Biophysical journal* 85.1 (July 2003), pp. 159–66. ISSN: 0006-3495. DOI: `10.1016/S0006-3495(03)74462-2` (cit. on p. 14).

[42] P D Thomas and K A Dill. «Statistical potentials extracted from protein structures: how accurate are they?» In: *Journal of molecular biology* 257.2 (Mar. 1996), pp. 457–69. ISSN: 0022-2836. DOI: `10.1006/jmbi.1996.0175` (cit. on p. 14).

[43] Min-yi Shen and Andrej Sali. «Statistical potential for assessment and prediction of protein structures». In: *Protein Science : A Publication of the Protein Society* 15.11 (Nov. 2006), pp. 2507–2524. ISSN: 0961-8368. DOI: `10.1110/ps.062416606` (cit. on p. 14).

[44] Hongyi Zhou and Yaoqi Zhou. «Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction.» In: *Protein science : a publication of the Protein Society* 11.11 (Dec. 2002), pp. 2714–26. ISSN: 0961-8368. DOI: `10.1110/ps.0217002` (cit. on p. 14).

[45] Manfred J Sippl. «Knowledge-based potentials for proteins». In: *Current Opinion in Structural Biology* 5.2 (Apr. 1995), pp. 229–235. ISSN: 0959440X. DOI: `10.1016/0959-440X(95)80081-6` (cit. on p. 14).

[46] Changsheng Du and Xin Xie. «G protein-coupled receptors as therapeutic targets for multiple sclerosis». In: *Cell Res* 22.7 (June 2012), pp. 1108–1128. ISSN: 1001-0602 (cit. on p. 15).

[47] Kim R Kampen. «Membrane Proteins: The Key Players of a Cancer Cell». In: *The Journal of Membrane Biology* 242.2 (2011), pp. 69–74. ISSN: 1432-1424. DOI: `10.1007/s00232-011-9381-7` (cit. on p. 15).

[48] Julia Koehler Leman, Martin B Ulmschneider, and Jeffrey J Gray. «Computational modeling of membrane proteins». In: *Proteins* 83.1 (Jan. 2015), pp. 1–24. ISSN: 0887-3585. DOI: `10.1002/prot.24703` (cit. on p. 15).

[49] Marc A Martí-Renom et al. «Comparative Protein Structure Modeling of Genes and Genomes». In: *Annual Review of Biophysics and Biomolecular Structure* 29.1 (June 2000), pp. 291–325. ISSN: 1056-8700. DOI: `10.1146/annurev.biophys.29.1.291` (cit. on p. 15).

[50] Lars Malmström and David R Goodlett. «Protein structure modeling.» In: *Methods in molecular biology (Clifton, N.J.)* 673 (2010), pp. 63–72. ISSN: 1940-6029. DOI: `10.1007/978-1-60761-842-3{\_}5` (cit. on p. 15).

[51] A Sali and T L Blundell. «Comparative protein modelling by satisfaction of spatial restraints.» In: *Journal of molecular biology* 234.3 (Dec. 1993), pp. 779–815. ISSN: 0022-2836. DOI: `10.1006/jmbi.1993.1626` (cit. on p. 15).

[52] Marco Biasini et al. «SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information». In: *Nucleic Acids Research* 42.Web Server issue (July 2014), W252–W258. ISSN: 0305-1048. DOI: `10.1093/nar/gku340` (cit. on p. 15).

[53] Johannes Söding, Andreas Biegert, and Andrei N Lupas. «The HHpred interactive server for protein homology detection and structure prediction». In: *Nucleic Acids Research* 33.Web Server issue (July 2005), W244–W248. ISSN: 0305-1048. DOI: `10.1093/nar/gki408` (cit. on p. 15).

[54] Jianyi Yang et al. «The I-TASSER Suite: protein structure and function prediction». In: *Nat Meth* 12.1 (Jan. 2015), pp. 7–8. ISSN: 1548-7091 (cit. on p. 15).

[55] Ambrish Roy, Alper Kucukural, and Yang Zhang. «I-TASSER: a unified platform for automated protein structure and function prediction». In: *Nature protocols* 5.4 (Apr. 2010), pp. 725–738. ISSN: 1754-2189. DOI: `10.1038/nprot.2010.5` (cit. on p. 15).

[56] Yang Zhang. «I-TASSER server for protein 3D structure prediction». In: *BMC Bioinformatics* 9 (Jan. 2008), p. 40. ISSN: 1471-2105. DOI: `10.1186/1471-2105-9-40` (cit. on p. 15).

[57] David E Kim, Dylan Chivian, and David Baker. «Protein structure prediction and analysis using the Robetta server». In: *Nucleic Acids Research* 32.Web Server issue (July 2004), W526–W531. ISSN: 0305-1048. DOI: `10.1093/nar/gkh468` (cit. on p. 15).

[58] Morten Källberg et al. «Protein Structure Prediction». In: ed. by Daisuke Kihara. New York, NY: Springer New York, 2014. Chap. RaptorX se, pp. 17–27. ISBN: 978-1-4939-0366-5. DOI: `10.1007/978-1-4939-0366-5{\_}2` (cit. on p. 15).

[59] P A Bates et al. «Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM.» In: *Proteins* Suppl 5 (Jan. 2001), pp. 39–46. ISSN: 0887-3585 (cit. on p. 15).

[60] G. Vriend. «WHAT IF: A molecular modeling and drug design program». In: *Journal of Molecular Graphics* 8.1 (Mar. 1990), pp. 52–56. ISSN: 02637855. DOI: `10.1016/0263-7855(90)80070-V` (cit. on p. 15).

[61] Jane S. Richardson. *Advances in Protein Chemistry Volume 34*. Vol. 34. 1981, pp. 167–339. ISBN: 9780120342341. DOI: `10.1016/S0065-3233(08)60520-3` (cit. on p. 16).

[62] Extracellular Proteins That and Modulate Cell-matrix Interactions. «Extracellular Proteins That». In: 266.23 (1991), pp. 15–18 (cit. on p. 16).

[63] D B Wetlaufer. «Nucleation, rapid folding, and globular intrachain regions in proteins.» In: *Proceedings of the National Academy of Sciences of the United States of America* 70.3 (1973), pp. 697–701. ISSN: 0027-8424. DOI: `10.1073/pnas.70.3.697` (cit. on p. 16).

[64] Suhail A Islam et al. «Identification and analysis of domains in proteins presented to identify domains in proteins». In: 8.6 (1995), pp. 513–525 (cit. on p. 16).

[65] Jung-Hoon Han et al. «The folding and evolution of multidomain proteins.» In: *Nature reviews. Molecular cell biology* 8.4 (2007), pp. 319–330. ISSN: 1471-0072. DOI: `10.1038/nrm2144` (cit. on p. 16).

[66] Cyrus Chothia et al. «Evolution of the protein repertoire.» In: *Science (New York, N.Y.)* 300.5626 (2003), pp. 1701–3. ISSN: 1095-9203. DOI: `10.1126/science.1085371` (cit. on p. 16).

[67] Christine Vogel et al. «Structure, function and evolution of multidomain proteins». In: *Current Opinion in Structural Biology* 14.2 (2004), pp. 208–216. ISSN: 0959440X. DOI: `10.1016/j.sbi.2004.03.011` (cit. on p. 16).

[68] G Apic, J Gough, and S a Teichmann. «Domain combinations in archaeal, eubacterial and eukaryotic proteomes.» In: *Journal of molecular biology* 310.2 (2001), pp. 311–325. ISSN: 0022-2836. DOI: `10.1006/jmbi.2001.4776` (cit. on p. 16).

[69] Alexey G. Murzin et al. «SCOP: A structural classification of proteins database for the investigation of sequences and structures». In: *Journal of Molecular Biology* 247.4 (1995), pp. 536–540. ISSN: 00222836. DOI: `10.1016/S0022-2836(05)80134-2` (cit. on p. 16).

[70] Ca Orengo et al. «CATH - a hierarchic classification of protein domain structures». In: *Structure* March (1997), pp. 1093–1109. ISSN: 09692126. DOI: `10.1016/S0969-2126(97)00260-8` (cit. on p. 16).

[71] A Bateman et al. «The Pfam protein families database». In: *Nucleic Acids Research* 28.1 (2002), pp. 276–280. ISSN: 0305-1048 (Print) 0305-1048 (Linking). DOI: `gkd038[pii]` (cit. on p. 16).

[72] Sarah Hunter et al. «InterPro: The integrative protein signature database». In: *Nucleic Acids Research* 37.SUPPL. 1 (2009), pp. 211–215. ISSN: 03051048. DOI: `10.1093/nar/gkn785` (cit. on p. 16).

[73] Friedrich Cramer. «Emil Fischer's Lock and Key Hypothesis after 100 years towards a Supracellular Chemistry». In: *Perspectives in Supramolecular Chemistry*. John Wiley & Sons, Ltd., 1994, pp. 1–23. ISBN: 9780470511411. DOI: `10.1002/9780470511411.ch1` (cit. on p. 16).

[74] D E Koshland. «Enzyme flexibility and enzyme action». In: *Journal of Cellular and Comparative Physiology* 54.S1 (Dec. 1959), pp. 245–258. ISSN: 1553-0809. DOI: `10.1002/jcp.1030540420` (cit. on p. 17).

[75] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. «On the nature of allosteric transitions: A plausible model». In: *Journal of Molecular Biology* 12.1 (May 1965), pp. 88–118. ISSN: 00222836. DOI: `10.1016/S0022-2836(65)80285-6` (cit. on p. 17).

[76] Akio Kitao, Steven Hayward, and Nobuhiro Go. «Energy landscape of a native protein: Jumping among minima model». In: *Proteins: Structure, Function, and Bioinformatics* 33.4 (Dec. 1998), pp. 496–517. ISSN: 1097-0134. DOI: `10.1002/(SICI)1097-0134(19981201)33:4<496::AID-PROT4>3.0.CO;2-1` (cit. on p. 17).

[77] G A Petsko and D Ringe. «Fluctuations in Protein Structure from X-Ray Diffraction». In: *Annual Review of Biophysics and Bioengineering* 13.1 (June 1984), pp. 331–371. ISSN: 0084-6589. DOI: `10.1146/annurev.bb.13.060184.001555` (cit. on p. 17).

[78] J Foote and C Milstein. «Conformational isomerism and the diversity of antibodies.» In: *Proceedings of the National Academy of Sciences of the United States of America* 91.22 (Oct. 1994), pp. 10370–10374. ISSN: 0027-8424 (cit. on p. 17).

[79] Leo C James, Pietro Roversi, and Dan S Tawfik. «Antibody Multispecificity Mediated by Conformational Diversity». In: *Science* 299.5611 (Feb. 2003), pp. 1362–1367 (cit. on p. 17).

[80] Michael F Dunn. «Protein-Ligand Interactions: General Description». In: *eLS*. John Wiley & Sons, Ltd, 2001. ISBN: 9780470015902. DOI: `10.1038/npg.els.0001340` (cit. on p. 17).

[81] Julien Michel, Julian Tirado-Rives, and William L Jorgensen. «Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization». In: *Journal of the American Chemical Society* 131.42 (Oct. 2009), pp. 15403–15411. ISSN: 0002-7863. DOI: `10.1021/ja906058w` (cit. on p. 19).

[82] Krishna Ravindranathan et al. «Improving MM-GB SA Scoring through the Application of the Variable Dielectric Model». In: *Journal of chemical theory and computation* 7.12 (Dec. 2011), pp. 3859–3865. ISSN: 1549-9618. DOI: `10.1021/ct200565u` (cit. on p. 19).

[83] Julien Michel, Marcel L Verdonk, and Jonathan W Essex. «Protein Ligand Binding Affinity Predictions by Implicit Solvent Simulations: A Tool for Lead Optimization». In: *Journal of Medicinal Chemistry* 49.25 (Dec. 2006), pp. 7427–7439. ISSN: 0022-2623. DOI: `10.1021/jm061021s` (cit. on p. 19).

[84] Hao-Yang Liu, Sam Z Grinter, and Xiaoqin Zou. «Multiscale generalized Born modeling of ligand binding energies for virtual database screening». In: *The journal of physical chemistry. B* 113.35 (Sept. 2009), pp. 11793–11799. ISSN: 1520-6106. DOI: `10.1021/jp901212t` (cit. on p. 19).

[85] Yipin Lu et al. «Analysis of Ligand-Bound Water Molecules in High-Resolution Crystal Structures of Protein-Ligand Complexes». In: *Journal of Chemical Information and Modeling* 47.2 (Mar. 2007), pp. 668–675. ISSN: 1549-9596. DOI: `10.1021/ci6003527` (cit. on p. 19).

[86] Kim A Sharp and Barry. Honig. «Calculating total electrostatic energies with the nonlinear Poisson-Boltzmann equation». In: *The Journal of Physical Chemistry* 94.19 (Sept. 1990), pp. 7684–7692. ISSN: 0022-3654. DOI: `10.1021/j100382a068` (cit. on p. 20).

[87] Donald Bashford and David A Case. «GENERALIZED BORN MODELS OF MACROMOLECULAR SOLVATION EFFECTS». In: *Annual Review of Physical Chemistry* 51.1 (Oct. 2000), pp. 129–152. ISSN: 0066-426X. DOI: `10.1146/annurev.physchem.51.1.129` (cit. on p. 20).

[88] Richard A. Friesner et al. «Glide: A New Approach for Rapid, Accurate Docking and Scoring». In: *Journal of Medicinal Chemistry* 47.7 (2004), pp. 1739–1749. ISSN: 00222623. DOI: `10.1021/jm0306430`. arXiv: `arXiv:1011.1669v3` (cit. on pp. 20, 21).

[89] Claudia Steffen et al. «TmoleX a graphical user interface for TURBOMOLE.» In: *Journal of computational chemistry* 31.16 (2010), pp. 2967–2970. ISSN: 1096-987X. DOI: `10.1002/jcc`. arXiv: `NIHMS150003` (cit. on p. 20).

[90] Peter Csermely et al. «Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review». In: *Pharmacology and Therapeutics* 138.3 (2013), pp. 333–408. ISSN: 01637258. DOI: `10.1016/j.pharmthera.2013.01.016`. arXiv: `1210.0330` (cit. on p. 20).

[91] Philip Prathipati and Kenji Mizuguchi. «Systems Biology Approaches to a Rational Drug Discovery Paradigm». In: *Current Topics in Medicinal Chemistry* 16.9 (), pp. 1009–1025 (cit. on p. 20).

[92] A Patrícia Bento et al. «The ChEMBL bioactivity database: an update». In: *Nucleic acids research* 42.Database issue (Jan. 2014), pp. D1083–90. ISSN: 0305-1048. DOI: `10.1093/nar/gkt1031` (cit. on p. 21).

[93] Feng Zhu et al. «Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery». In: *Nucleic Acids Research* 40.Database issue (Jan. 2012), pp. D1128–D1136. ISSN: 0305-1048. DOI: `10.1093/nar/gkr797` (cit. on p. 21).

[94] Liegi Hu et al. «Binding MOAD (Mother Of All Databases).» In: *Proteins* 60.3 (Aug. 2005), pp. 333–40. ISSN: 1097-0134. DOI: `10.1002/prot.20512` (cit. on p. 21).

[95] Tiqing Liu et al. «BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities». In: *Nucleic Acids Research* 35.Database issue (Jan. 2007), pp. D198–D201. ISSN: 0305-1048. DOI: `10.1093/nar/gkl999` (cit. on p. 21).

[96] Sunghwan Kim et al. «PubChem Substance and Compound databases». In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D1202–D1213. DOI: `10.1093/nar/gkv951` (cit. on p. 21).

[97] Yanli Wang et al. «PubChem BioAssay: 2014 update». In: *Nucleic Acids Research* 42.Database issue (Jan. 2014), pp. D1075–D1082. ISSN: 0305-1048. DOI: `10.1093/nar/gkt978` (cit. on p. 21).

[98] John J Irwin et al. «ZINC: A Free Tool to Discover Chemistry for Biology». In: *Journal of Chemical Information and Modeling* 52.7 (July 2012), pp. 1757–1768. ISSN: 1549-9596. DOI: `10.1021/ci3001277` (cit. on p. 21).

[99] Hernán Alonso, Andrey A. Bliznyuk, and Jill E. Gready. «Combining docking and molecular dynamic simulations in drug design». In: *Medicinal Research Reviews* 26.5 (2006), pp. 531–568. ISSN: 01986325. DOI: `10.1002/med.20067` (cit. on p. 21).

[100] Garrett M Morris et al. «AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility». In: *Journal of computational chemistry* 30.16 (Dec. 2009), pp. 2785–2791. ISSN: 0192-8651. DOI: `10.1002/jcc.21256` (cit. on p. 21).

[101] Oleg Trott and Arthur J Olson. «AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading». In: *Journal of computational chemistry* 31.2 (Jan. 2010), pp. 455–461. ISSN: 0192-8651. DOI: `10.1002/jcc.21334` (cit. on p. 21).

[102] Todd J a. Ewing and Irwin D Kuntz. «Critical evaluation of search algorithms for automated molecular docking and database screening». In: *Journal of Computational Chemistry* 18 (1997), pp. 1175–1189. ISSN: 0192-8651. DOI: `10.1002/(SICI)1096-987X(19970715)18:9<1175::AID-JCC6>3.0.CO;2-O` (cit. on p. 21).

[103] Matthias Rarey et al. «A Fast Flexible Docking Method using an Incremental Construction Algorithm». In: *Journal of Molecular Biology* 261.3 (Aug. 1996), pp. 470–489. ISSN: 0022-2836. DOI: `http://dx.doi.org/10.1006/jmbi.1996.0477` (cit. on p. 21).

[104] G Jones et al. «Development and validation of a genetic algorithm for flexible docking.» In: *Journal of molecular biology* 267.3 (Apr. 1997), pp. 727–48. ISSN: 0022-2836. DOI: `10.1006/jmbi.1996.0897` (cit. on p. 21).

[105] Julie R Schames et al. «Discovery of a novel binding trench in HIV integrase.» In: *Journal of medicinal chemistry* 47.8 (Apr. 2004), pp. 1879–81. ISSN: 0022-2623. DOI: `10.1021/jm0341913` (cit. on p. 21).

[106] Istvan J Enyedy et al. «Discovery of Small-Molecule Inhibitors of Bcl-2 through Structure-Based Computer Screening». In: *Journal of Medicinal Chemistry* 44.25 (Dec. 2001), pp. 4313–4324. ISSN: 0022-2623. DOI: `10.1021/jm010016f` (cit. on p. 21).

[107] Eric Vangrevelinghe et al. «Discovery of a Potent and Selective Protein Kinase CK2 Inhibitor by High-Throughput Docking». In: *Journal of Medicinal Chemistry* 46.13 (June 2003), pp. 2656–2662. ISSN: 0022-2623. DOI: `10.1021/jm030827e` (cit. on p. 21).

[108] D Kitchen et al. «Docking and scoring in virtual screening for drug discovery: methods and applications». In: *Nature Reviews Drug Discovery* 3.11 (2004), pp. 935–949. ISSN: 1474-1784. DOI: `10.1038/nrd1549` (cit. on p. 21).

[109] Sara Reardon. «Project ranks billions of drug interactions.» In: *Nature* 503.7477 (2013), pp. 449–50. ISSN: 1476-4687. DOI: `10.1038/503449a` (cit. on p. 21).