Development of chemogenomic approaches for prediction of protein - ligand interactions

Brice Hoffmann

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

École doctorale n°432 : Sciences des Métiers de l'ingénieur

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

l'École nationale supérieure des mines de Paris

Spécialité « Bio-informatique »

présentée et soutenue publiquement par

Brice HOFFMANN

le 16 decembre 2011

Development of chemogenomic approaches for prediction of protein-ligand interactions

Directeur de thèse : **Véronique STOVEN**

Jury

M. Daniel ABERGEL,	Directeur de Recherche, Département de Chimie, Ecole Normale Supérieure	Examinateur
Mme Hélène DEBAT,	Maître de Conférences, Institut de Génétique, Université Versailles Saint-Quentin	Examinatrice
M. Dragos HORVATH,	Directeurs de Recherche, Laboratoire d'Infochimie, Université de Strasbourg	Rapporteur
M. Olivier LEQUIN,	Professeur, Laboratoire des Biomolécules, Université Pierre et Marie Curie	Rapporteur
M. Michael NILGES,	Professeur, Unité de Bioinformatique Structurale, Institut Pasteur	Examinateur
Mme Véronique STOVEN,	Professeur, Centre de Bioinformatique, Mines ParisTech	Examinatrice

MINES ParisTech
Centre de Bio-informatique
35 rue St Honoré 77300 FONTAINEBLEAU

T H È S E

Résumé en français

L'enjeu de cette thèse porte sur le développement de méthodes de bio-informatique appliquées à la prédiction des interactions protéine - ligand. Le chapitre 1 est une courte introduction qui rappelle le rôle clé que jouent ces interactions dans le fonctionnement de la cellule. Il présente également le plan de ce manuscrit.

Dans le chapitre 2, après avoir rappelé quelles sont les molécules clés du monde vivant ainsi que les principes de l'interaction entre les protéines et les ligands, est présenté l'état de l'art de l'étude de ces interactions, à la fois expérimentale et *in silico* avec les méthodes basées sur l'étude des ligands (ligand-based) et celles basées sur l'étude des protéines (protein-based). Enfin la méthode d'apprentissage statistique des machines à vecteurs de support (SVM) appliquée au criblage virtuel est présentée, car cet algorithme a été employé au cours de cette thèse pour développer une méthode de prédiction appartenant aux approches dites de chémogénomique.

Dans un premier temps (chapitre 3), la thèse se concentre sur l'élaboration de cette méthode de chémogénomique appliquée aux protéines de la famille des GPCRs. Cette méthode est basée sur l'utilisation de machines à vecteurs de support pour prédire l'interaction entre des GPCRs et leurs ligands. Cette approche suppose l'emploi de descripteurs pour encoder les protéines et les ligands. Plusieurs types de descripteurs ont été employés, afin de comparer leur pertinence dans le cadre de la chémogénomique. Pour les ligands, des descripteurs correspondant à un encodage de la structure 2D ou de la structure 3D ont été testés. Dans l'approche 2D, une molécule est décrites par un vecteur binaire dont les éléments sont déterminés en fonction d'un graphe qui décrit sa structure chimique. La similarité entre deux molécules est alors évaluée par un coéfficient de Tanimito. Dans le cas de l'approche 3D, les molécules sont décrites par l'ensemble des triplets d'atomes qui la composent, et des distances séparant ces atomes. La similarité entre deux molécules est évaluées en comparant les ensembles de leurs triplets respectifs, en

utilisant un noyau appelé noyau pharmacophore 3D. Les protéines sont encodées également de plusieurs façons: d'une part par leur position dans la hiérarchie des GPCR telle que cette hiérarchie est définie dans la base de données GLIDA, d'autre part par une courte séquence d'acides aminés correspondant aux acide aminées composant la poche de fixation pour le ligand. Les similarités entre protéines sont alors évaluées par plusieurs méthodes à noyaux. Tout d'abord, deux noyaux relativement basiques, n'employant pas ces encodages pour les protéines, sont employées: le noyau Dirac, dans lequel la similarité entre deux protéines différentes est égale à zéro, et le noyau Multitask, dans lequel toutes les protéines sont "également différentes". Le premier correspond en réalité à une approche classique par protéine: aucune information de ligand n'est partagée entre les protéines. Le second noyau correspond à une sorte de base-line pour les méthodes de chémogénomique, dans laquelle l'information concernant les ligands est partagée de manière uniforme entre toutes les protéines. Deux noyaux s'appuyant sur l'encodage des protéines sont ensuite définis et employés: le noyau hiérarchique qui évalue la similarité entre protéines en fonction de leur distance dans la hiérarchie des GPCR, et le noyau "binding pocket" qui évalue la similarité des acides aminés formant les sites de fixation des ligands. Ce dernier consiste à aligner structuralement les protéines de la famille des GPCRs dont la structure a été déterminée expérimentalement (deux au moment de la publication) afin de répertorier les acides aminés impliqués dans la fixation du ligand. Les séquences des autres GPCRs ont ensuite été alignés à ces deux protéines et les acides aminés correspondant au site de fixation ont été concaténés dans un vecteur qui a permis de les comparer. L'espace chémogénomique est encodé par le produit tensoriel des espaces des protéines et des ligands, et les distances entre les paires (protéine, ligand) dans cet espace est évalué par le produit des noyaux calculés sur les protéines et sur les ligands. Toutes les combinaisons entre les noyaux pour protéines et pour ligands ont été testées. La base de données d'interactions entre les GPCRs et leurs ligands utilisée est la GLIDA. Dans cette base, 4051 interactions ont été retenues pour évaluer la méthode. Deux types d'expériences ont été effectuées. La première consiste, pour chaque GPCR, à diviser les interactions connues en cinq parties. Le modèle est entraîné à l'aide de quatre des parties ainsi que des

données de l'ensemble des autres GPCRs puis il est testé sur la cinquième partie. Dans la deuxième expérience, pour chaque GPCR l'ensemble de ses ligands connus ont été ignorés et le modèle a été entraîné en utilisant seulement les données d'interactions des autres GPCRs. Cette expérience revient à évaluer les performances de la méthode dans le cas important mais difficile de GPCR orphelines, pour lesquelles aucun ligand n'est connu. Les résultats ont montré d'une part que la méthode 2D pour les ligands obtient systématiquement de meilleurs résultats que la méthode 3D. D'autre part, la méthode utilisant la hiérarchie a obtenu de meilleurs résultats pour la première expérience alors que la méthode utilisant les vecteurs décrivant le site de fixation a obtenu les meilleurs résultats dans la deuxième expérience. L'ensemble de ces résultats montre par ailleurs que toutes les méthodes de chemogénomique, y compris les plus naïves, présentent de meilleurs performances de prédiction des interactions que les méthodes classiques qui effectuent les prédictions par protéine, sans prendre en compte l'information concernant les interactions connues pour d'autres protéines de la famille.

L'une des limites de la méthode de chémogénomique présentée au chapitre 3 est qu'elle n'est applicable que pour des protéines apparentées, comme les protéines de la famille des GPCR. Nous avons souhaité étendre l'application des méthodes de chémogénomique à des protéines ne présentant aucune similarité de sequence ou de structure. L'idée sous-jacente est qu'il serait intéressant de pouvoir partager l'information sur les interactions protéine-ligand entre n'importe quelles protéines, afin d'accroitre la taille de la base de connaissance utilisable pour prédire de nouvelles interactions. Dans le chapitre 4, pour s'affranchir de l'approche par famille, les méthodes proposées seront applicables pour prédire les interactions protéine-ligand par une approche de chémogénomique, pour les protéines de structure 3D connue. Ici, les protéines sont encodées par le nuage de points correspondant aux atomes qui constituent sa poche de fixation pour le ligand. La similarité entre deux protéines est alors évaluée par la similarité entre les nuages des atomes de leurs poches de fixation pour les ligands. Cette méthode implique un alignement en 3D des atomes formant les deux poches, par rotation et translation. Le meilleur alignement est obtenu en favorisant le regroupement d'atomes des deux poches ayant des propriétés similaires dans des régions proches de l'espace. Cet alignement permet ensuite de mesurer la similarité entre les poches, et définit la similarité entre les

protéines. Pour une poche donnée, la prédiction des ligands est effectuée en fonction des ligands connus pour les poches les plus similaires, par une métode de ”plus proches voisins”. Plusieurs jeux de données ont été utilisés pour l'évaluation des performances. Un premier jeu, issu de la littérature, est constitué d'un ensemble de 100 protéines de familles différentes et dont les sites de fixation sont associés à un ligand, parmi une liste de 10 ligands de taille différente. Une version étendue de ce jeu de données a été créé et comporte 972 poches fixant l'un des 10 ligands. Un troisième jeu comprenant également 100 sites de fixations et 10 ligands de taille similaire a également été constitué et utilisé. La méthode développée ici a été comparée à plusieurs méthodes issues de la littérature. Pour cela deux critères ont été retenus : le score AUC (Air Under the ROC Curve) et l'erreur de classification. Les résultats obtenus ont montrés que la méthode présentée ici obtient les meilleurs résultats sur les deux jeux de données comprenant 100 poches, à la fois en terme d'AUC et d'erreur de classification. D'autre part la version étendue du premier jeu de données nous a permis de montrer que la méthode améliore ses prédictions lorsque la quantité de données augmente. Nous avons montré que l'erreur de classification est un meilleur critère d'évaluation des performances de prédiction que l'AUC qui est classiquement utilisé. Cette méthode possède l'avantage de pouvoir comparer le site de fixation de n'importe quel couple de protéines, quelque soit leurs familles et leurs similarités, à condition de posséder leurs structures 3D.

Enfin, le chapitre 5 discute les principales difficultés rencontrées dans les méthodes de chémogénomique, comme l'encodage des espaces des ligands, des protéines, et des paires (protéine, ligand), ainsi que la constitution des bases de données qui est un élément crucial dans les méthodes d'apprentisage. Le chapitre 3 propose une méthode de chémoénomique par famille de protéines, et le chapitre 4 propose une méthode de mesure de similarité pour protéines de structure connue, permettant la prédiction des interactions protéine-ligand par une méthode de plus proche voisin. Cependant, si cette dernière constitue bien une méthode de chemogénomique, elle ne permet pas l'emploi des SVM car la mesure de similarité définie sur les poches ne possède pas les propriétés d'un noyau. Le chapitre 5 indique donc des pistes d'exploration possible qui permettraient de ”transformer” cette mesure de similarité en noyau, afin de disposer d'une méthode de chémogénomique bénéficiant des performances et des caractéristiques des méthodes SVM à noyaux. Enfin, nous

évoquons comment les méthodes proposées dans cette thèse sont complémentaires d'autres approches de biologie structurale comme la modélisation par homologie ou le docking.

Contents

1

Introduction

Identification of ligands for proteins is a major field of research, both at the fundamental level and for many industrial applications. For example, it can help to decipher the function of a protein known to be involved in a disease, and therefore lead to a better understanding of the molecular disorder associated to this pathology. It can also help to discover new drugs for diseases with uncovered needs.

Historically, experimental methods have been developed to identify protein-ligands interactions, but since the last two decades, they have been supplemented with computational methods. These methods allow very fast and cheap screening of millions of molecules, in order to reduce the number of actual experimental assays to be undertaken.

In chapter 2, we present the state-of-the art in experimental and *in silico* approaches to study protein-ligand interactions. We will first remind the main molecules found in leaving cells, and briefly review how protein-ligand interactions can be studied experimentally. We will give a short overview of experimental High Throughput Screening (HTS) methods. We will also recall the principles of the two main in *silico* strategies, namely "ligand-based" and "structure-based" methods to predict protein- ligand interactions. Finally, we will shortly and intuitively present statistical learning methods that can, very generally, be used to predict properties of objects. We will show how these methods can be applied to the question of predicting protein-ligand interactions.

The main contributions of this thesis belong to two different but however related fields: encoding of proteins, and chemogenomics approaches for prediction of protein-ligand interactions.

Indeed, although encoding of small organic molecules has a long history in the field of cheminformatics, in a form that can be used as input in any computational method, encoding of proteins has been less studied. It has often been restricted to describing the protein by its primary amino-acid sequence. This description is not fully suited to the problem of prediction of protein-ligand interactions, because this description does not point at the protein function in a direct manner. The question of protein encoding is closely related to that of predicting protein-ligand interactions, since the encoded protein is used as input of computational method, and sice the more relevant the encoding with respect to the functional properties of the protein, the better the prediction.

In chapter 3, we propose to encode proteins by a short list of aminoacids expected to be involved in the ligand-binding site of the protein. This method can be applied for proteins within a given family, as long as at least one 3D structure is available in this family. As an example of application, we show that this approach can be used to encode GPCR, a large family of protein receptors of great interest for pharmaceutical industry.

Then, we show the recently introduced chemogenomic approaches, using this encoding of proteins, outperform other state-of-the art ligand prediction methods for GPCR.

In chapter 4, for proteins of known 3D structures, we propose a representation based on the cloud of atoms belonging to the ligand-binding pocket. We have developed a method that allows to compare proteins based on the similarity of their binding pockets (as described by clouds of atoms). We show that this similarity measure can in turn be used to predict ligands for a new pocket, based on known ligands for similar pockets. This method allows to ”learn” from any known protein-ligand complex, whatever its protein family, in order to predict new ligands for any pocket or to propose ligands for ”orphan” pockets, as long as the 3D structures are available.

In chapter 5, we will briefly show how the results obtained in chapter 4 could be generalized and included in a chemogenomic framework similar to that presented in chapter 3.

2

Background

2.1 Molecules of life

Understanding life requires knowing the players of the molecular mechanisms that regulate key cellular functions. The main players are the DNA, RNA and proteins. A link exists between these two types of macromolecules: in fact, DNA is the hereditary material of the cell, containing genes. The sequence of a gene encoding a particular protein. Each protein is the product of gene expression, i.e. results from transcription of DNA into RNA, and translation into proteins.

Like ourselves, the individual cells that form our bodies can grow, reproduce, process information, respond to stimuli, and carry out an amazing array of chemical reactions. These abilities define life. We and other multicellular organisms contain billions or trillions of cells organized into complex structures, but many organisms consist of a single cell. Even simple unicellular organisms exhibit all the properties of life, indicating that the cell is the fundamental unit of life. We face an explosion of new data about the components of cells, what structures they contain, how they touch and influence each other. Still, an immense amount remains to be learned, particularly about how information flows through cells and how they decide on the most appropriate ways to respond.

Molecular cell biologists explore how all the remarkable properties of the cell arise from underlying molecular events: the assembly of large molecules, binding of large molecules to each other, catalytic effects that promote particular chemical reactions, and the deployment of information carried by giant molecules.

2.2 Proteins, machinery of life.

Proteins are an important class of biological molecules. They provide most of the cellular functions. Proteins consist of 20 different natural amino acids, also known as residues. A protein is generally capable of one or more specific tasks. These functions are possible through the structure of the protein. It is therefore their structure that allows proteins to fulfill their function, key residues occupying relative positions in space that allow molecular recognition of the biological partners. Therefore, we understand the importance of studying the three dimensional structure of proteins.

Indeed, proteins fold to form a stable structure, driven by a number of non-covalent interactions such as hydrogen bonding, ionic interactions, ' Van der Waals' forces and hydrophobic packing. These forces give to the protein the cohesion that is necessary to maintain its structure. Determining the 3D structure of proteins is the subject of structural biology, which uses techniques such as X-ray crystallography, nuclear magnetic resonance (RMN) spectroscopy, or electron microscopy.

All freely available 3D structures of proteins are deposited in the Protein Data Bank (PDB). This database grows exponentially, because of the improvement of the technology and the number of researchers involved in this field worldwide. Currently, there are more than 70,000 crystallographic or NMR structures of proteins or nuclear acids available in PDB.

Two main classes of proteins are important for the pharmaceutical industry : receptors and enzymes. In the case of receptors, more than 50% of currently marketed drugs have as main target a protein belonging to this family. For enzymes, although the rate is much lower, this family of proteins will be increasingly targeted by new drugs (1).

2.2.1 Enzymes

An enzyme is a protein (if we exclude the special case of RNA ribozymes) which lowers the activation energy of a reaction and speeds up millions of times chemical reactions of metabolism occurring in the cellular or extracellular environment without changing the balance formed. Enzymes act at low concentrations and they are found intact at the end of reaction: they are biological catalysts (or biocatalysts). For example, glucose oxidase is an enzyme that catalyzes the oxidation of glucose into gluconic acid.

An enzyme, like any protein, is synthesized by living cells from the information encoded in DNA or RNA in the case of some viruses. There are over 3500 different enzymes listed.

2.2.2 Receptors

A receptor is a protein from the cell, cytoplasm or nucleus membrane that binds to a specific factor (a ligand such as a neurotransmitter, a hormone or other substances), inducing a cellular response to this ligand. The behavioral changes of the receptor protein induced by the ligand leads to physiological changes that constitute the "biological effects" of the ligand. There are different types of receptors depending on their ligands and their functions:

- Some receptor proteins are proteins of the outer part of the plasma membrane
- Many receptors for hormones and neurotransmitters are transmembrane proteins embedded in the lipid bilayer of cell membranes. These receptors are coupled to either G proteins or holders of an enzymatic activity, or ion channel allowing the activation of metabolic pathways of signal transduction in response to ligand binding.
- The other major class of receptors consists of intracellular proteins such as steroid hormone receptors. These receptors can sometimes enter the nucleus of the cell to modulate the expression of specific genes in response to activation by the ligand.

2.3 Small molecules

Much of the cell's content is a watery soup flavored with small molecules (e.g., simple sugars, amino acids, vitamins) and ions (e.g., sodium, chloride, calcium ions). The locations and concentrations of small molecules and ions within the cell are controlled by numerous proteins inserted in cellular membranes. These pumps, transporters, and ion channels move nearly all small molecules and ions into or out of the cell and its organelles.

One of the best-known small molecules is adenosine triphosphate (ATP) (Figure 2.1), which stores readily available chemical energy in two of its phosphate chemical bonds. When cells split apart these energy-rich bonds in ATP, the released energy can be harvessed to power an energy-requiring process like muscle contraction or protein biosynthesis. To obtain energy for making ATP, cells break down food molecules. For instance, when sugar is degraded to carbon dioxide and water, the energy stored in the original chemical bonds is released and much of it can be "captured" in ATP. Bacterial, plant, and animal cells can all make ATP by this process. In addition, plants and a few other organisms can harvest energy from sunlight to form ATP using photosynthesis.

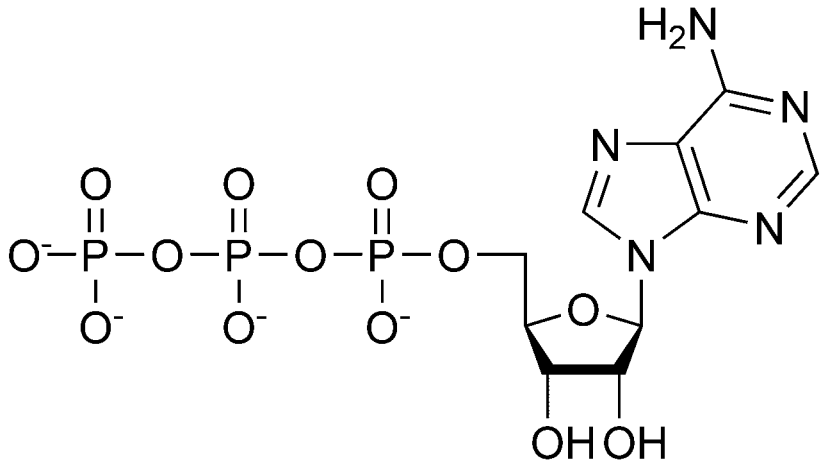


Figure 2.1: representation of an ATP molecule

Other small molecules act as signals both within and between cells. Such signals direct numerous cellular activities. For example, The powerful effect on our bodies of a frightening event comes from the instantaneous flooding of the body with epinephrine, a small-molecule hormone that mobilizes the ”fight or flight” response. The movements needed to fight or flee are triggered by nerve impulses that flow from the brain to our muscles with the aid of neurotransmitters.

Certain small molecules (monomers) in the cellular soup can be joined to form polymers through repetition of a single type of chemical-linkage reaction. Cells produce three types of large polymers, commonly called macromolecules: polysaccharides, proteins, and nucleic acids. Sugars, for example, are the monomers used to form polysaccharides. These macromolecules are critical structural components of plant cell walls and insect skeletons. A typical polysaccharide is a linear or branched chain of repeating identical sugar units. Such a chain carries information: the number of units. However, if the units are not identical, then the order and type of units carry additional information. Some polysaccharides exhibit the greater informational complexity associated with a linear code made up of different units assembled in a particular order. This property, however, is most typical of the two other types of biological macromolecules: proteins and nucleic acids.

2.4 Interactions between proteins and small molecules

A ligand can be defined as a molecule binding to a biological receptor, which is most often a protein (Figure 2.2). This binding generally triggers an effect: modulation of enzyme activity if the receptor is an enzyme, cellular response in the case of a membrane receptor, a cytoplasmic receptor or a nuclear receptor of a cell. Ligands include all small molecules of low molecular weight, such as a metabolite, peptide, substrate, inhibitor or a small drug molecule and excludes macromolecules such as proteins, lipids or nucleic acids sequences such as DNA and RNA, which we would rather call biological partners.

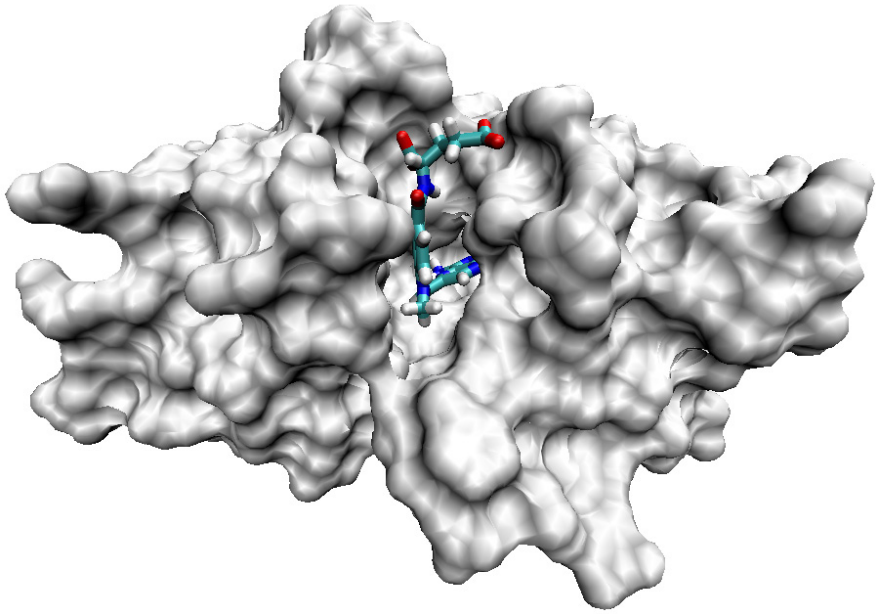


Figure 2.2: 3D structure of a protein (DHFR) with its ligand (methotrexate)

Ligand binding to the protein occurs by intermolecular forces such as ionic bonds, hydrogen bonds and Van der Waals forces, and is usually reversible.

The interaction between proteins and small molecules is related to the presence, in the protein structure, of a specific site called the active site. Broadly, it has the shape of a cavity into which the substrates bond. Once bound, small molecules will react and turn into a product in the case of an enzyme, or cause the activation or inhibition of a signaling pathway in the case of a receptor.

The first model of ligands binding to proteins was the ”lock and key” hypothesis (Figure 2.3). In this model, the affinity of a ligand for a protein is determined by the complementarity between the shapes and the physico-chemical properties of the ligand and of the binding site.

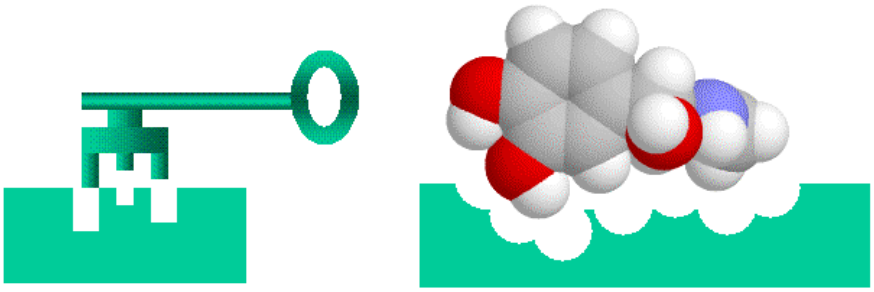


Figure 2.3: The ”lock and key hypothesis”

But this model is not sufficient to describe the interaction. Indeed, proteins and ligands can be very flexible. The model is rather a key capable of deforming the lock when it fits. This phenomenon is called induced fit. The theory suggests that the protein changes its shape to bind the molecule with the proper alignment. The overall effect would be a tighter binding for the molecule and the binding site. Once bound, the molecule intereacts with the protein the same as with the lock and key theory (2).

2.5 Experimental methods to study protein-ligand interactions

The study of protein-ligand interaction is crucial to understand the function of a protein. In fundamental studies, it allows to describe the biological pathways in which the protein is involved. It is also a key step towards the design of modulator molecules that can interfere (positively or negatively) with the function of a protein target, in the context of drug discovery. Indeed, research and discovery of new molecules with activity against proteins has long been the goal of the pharmaceutical industry. There are many experimental techniques to study the interactions between a molecule and a protein. In the following, I will review the most popular methods, distinguishing those that can provide structural information about the complex and the binding mode of the small molecule to the protein, and those that cannot (non structural approaches).

2.5.1 Non structural approaches

2.5.1.1 Spectroscopic methods (UV, fluorescence)

Spectroscopic methods are often used to highlight molecular interactions. The principle is simple: it is based on the change of the absorbance of a molecule after interaction or rupture of molecular interaction. For example, these methods are often used to monitor the evolution of an enzymatic reaction. The simplicity of implementation, the high sensitivity and the low cost in organic materials are the main advantages of these techniques. However, they can be applied only if one of the molecules involved in a complex have spectral properties and if the specific absorbance of the targeted group is affected by the interaction. In some cases, it is possible to attach fluorescent groups to non fluorescent ligands, although this chemical modification might modify the stability of the complex. Many biological assays involving therapeutic targets have been developed that rely on such spectroscopic methods, as reviewed in (3) for fluorescence and (4) for UV spectroscopy. These methods can be used to measure the complex association constant, i.e. to characterize the strength of the interaction.

2.5.1.2 Isothermal Titration Calorimetry (ITC)

ITC is an analysis technique based on the measurement of heat changes induced during a titration. In practice, a macromolecule located in the measuring cell of a calorimeter is gradually saturated at constant temperature by the injection of a ligand using a syringe. For each addition of ligand, there is a thermal exchange that is characteristic of the macromolecule-ligand interactions. The amount of heat measured during the titration allows to obtain thermodynamic parameters of interaction such as free energy changes (ΔG), enthalpy (ΔH) and entropy (ΔS). ITC therefore allows to highlight an interaction, to determine the dissociation constant Kd and the stoichiometry of the system. Furthermore, the thermodynamic data derived by microcalorimetry allow to know precisely the enthalpic and entropic contributions in the interaction energy: we can thus specify the nature of the forces contributing to the formation of complexes (hydrophobic or electrostatic) (5). ITC can also be useful to study ionization phenomena, or to highlight and quantify a competition between two ligands. A drawback of this technique is its quite low sensitivity, requiring large amounts of proteins (in the milligram range) to study the complex, which can be a strong limitation for proteins available only in small quantities. In addition, the experiments are relatively lengthy and difficult to perform. Howeveer, the ITC technique has been used in drug discovery applications thanks to a new

miniaturized, ultrasensitive microcalorimeter. This new microcalorimetry system reduces the quantity of protein (or other macromolecule sample) required to obtain a complete thermodynamic profile by up to 7-fold. The reduction in required sample quantities allows ITC to be effectively utilized at earlier stages of the drug discovery and development process (6).

2.5.1.3 Surface Plasmon Resonance (SPR)

The SPR technology is used to study molecular interactions in real time without labeling one of the two interactants. Without going into many details, the device detects changes in mass at the surface of a sensor chip on which one of the two interactants (for example the ligand) is immobilized, covalently or not. The other interactant (for example the protein) is injected through a microfluidic system in a continuous flow of buffer to the surface of the sensor chip. If the protein and the ligand interact and bind to each other, the device detects an apparent variation in mass for the immobilized molecule. Equilibrium binding constants, kinetic rate constants and thermodynamic parameters are obtained from such study that helps to understand the mechanism of the binding reactions. This information can be directly used to improve binding properties of a drug candidate (see (7) for review).

2.5.2 Structural approaches

Structure based drug design is another method for identifying new drugs and seems to be the most rational way of identifying potential agents. In this approach, the three-dimensional structure of a drug target and its interaction with potential drug molecules is used to guide drug discovery. This structural information can be obtained by various methods, such as X-ray crystallography, NMR, and virtual approaches such as computational chemistry. In the next paragraphs, we will shortly review the two former, which rely or experimental data. Then, virtual approaches will be reviewed in more details, because this techniques are related to the work presented in the following chapters of this manuscript.

2.5.2.1 X-ray diffraction

Crystallography uses X-ray diffraction by the electronic cloud of a molecule to deduce the positions of the atoms constituting the compound to be analyzed. In the case of biomolecules, it is possible to crystallize a protein in complex with its ligand, and to deduce the three dimensional structure of the complex. Such a structure allows to discover the active site of the protein, and

to assess ligand binding modes. This information can be used to conceive optimized molecules that improve their interactions with its target. Crystallography is the technique of choice used in the stages of drug design where an active molecule has been selected and needs to be optimized to ensure high affinity and selectivity. For instance, the development of successful HIV-1 protease (see (8) for review), reverse transcriptase (see (9) for review), or integrase (see (10) for review) inhibitors was achieved through structure-based drug design using the crystal structures of the corresponding enzymes. In the field of antibiotics, the translational apparatus of the bacterial cell remains one of the principal targets of antibiotics for the clinical treatment of infection worldwide. The high-resolution crystal structures of the bacterial ribosome identifying the sites of antibiotic binding are now available, which is central to progress in this area. Experimental assays, coupled with structural studies, have the potential not only to accelerate the discovery of novel and effective antimicrobial agents, but also to refine our understanding of the mechanisms of translation (see (11) for review).

2.5.2.2 Nuclear magnetic resonance (NMR)

Nuclear magnetic resonance is a spectroscopic technique based on the interaction between a magnetic field and the spins of atomic nuclei. The study of protein / ligand interactions at the atomic level by NMR has been made possible through the development of experiments based on observation of the resonance signals of the protein or of the ligand, in presence of each other (12; 13). NMR can provide much information to characterize the interaction between a protein and a ligand. It allows to identify the site of interaction in the protein, i.e. the aminoacids that are in contact with the ligand. It can also provide information about the conformation of the ligand in the active site. NMR can play a critical role in structure determination of many important protein targets such as GPCRs, when they fail to form the single crystals required for X-ray diffraction. NMR can provide valuable dynamic information on proteins and their drug complexes that cannot be easily obtained with X-ray crystallography. These advances suggest that the future discovery and design of drugs might increasingly rely on protocols using NMR approaches (14). However, this technique consumes large amounts of biological material, milligrams per analysis, because of its low sensitivity. In addition, analysis of the spectra is easier in the fast exchange regime, which corresponds to dissociation constant of the protein-ligand complex in the range of millimolar or micromolar, which is not the required range in the context of drug discovery.

2.6 Experimental high throughput screening (HTS)

In many cases, the search of new drug candidates consists in the identification of molecules that strongly and specifically bind to a therapeutic protein target. Therefore, all methods allowing detection and characterization of protein-ligand interactions are of interest in the context of drug research. Some of the above described experimental methods can be used in screening tests on large scale, using robots, leading to the so called high throughput screening (HTS) approaches (15). In the following, I will briefly review HTS.

The screening of chemical compounds for pharmacological activity has been ongoing in various forms for at least 20 years. The screening paradigm says that when a compound interacts with a target in a productive way, that compound then passes the first milestone on the way to becoming a drug. Compounds that fail this initial screen go back into the library, perhaps to be screened later against other targets.

Screening methodologies have improved with time, both in terms of throughput and the amount of information to be derived from the screen. Advances in assay and instrument technologies have provided the means necessary to address these evolving needs.

Using robotics, data processing and control software, liquid handling devices, and sensitive detectors, HTS allows a researcher to conduct biochemical or pharmacological tests. Through this process, one can rapidly identify active compounds which modulate a particular biomolecular pathway. The results of these experiments provide starting points for drug design and for understanding the interaction or the role of a particular biochemical process in biology.

Many pharmaceutical companies are screening 100,000 to 300,000 or more compounds per screen to produce approximately 100 to 300 hits. On average, one or two of these become lead compound series. Larger screens of up to 1,000,000 compounds in several months may be required to generate something closer to five leads. Improvements in lead generation can also come from optimizing library diversity. Since its first advent in the early to mid 1990s, the field of HTS has seen not only a continuous change in technology and processes, but also an adaptation to various needs in lead discovery. HTS has now evolved into a mature discipline that is a crucial source of chemical starting points for drug discovery. Whereas in previous years much emphasis has been put on a steady increase in screening capacity ('quantitative increase') via automation and miniaturization, the past years have seen a much greater emphasis on content and quality ('qualitative increase'). Today, many experts in the field see HTS at a crossroad with the need to decide on either higher throughput/more experimentation or a greater focus

on assays of greater physiological relevance, both of which may lead to higher productivity in pharmaceutical R&D. There will be much more emphasis on rigorous assay and chemical characterization, particularly considering that novel and more difficult target classes will be pursued. In recent years, we have witnessed a clear trend in the drug discovery community toward rigorous hit validation by the use of orthogonal readout technologies, label free and biophysical methodologies. We also see a trend toward the use of focused screening and iterative screening approaches. Hit finding strategy also tends to be much more project-related and better integrated into the broader drug discovery efforts. Recently, fragment-based methods have emerged as a new strategy for drug discovery (16). The main advantages are that useful starting points for lead identification for most targets can be identified from a relatively small (typically 1000-member) library of low molecular weight compounds. The main constraints are the need for a method that can reliably detect weak binding and strategies for evolving the fragments into larger lead compounds. The approach has been validated recently, as series of compounds from various programs have entered clinical trials.

However, HTS workflow is hampered by several drawbacks. One can only sample a tiny proportion of the drug-like chemical space. The screening procedure relies on expensive robotic equipment, and although many progresses in miniaturization have been made that allow reduction of the experimental volumes, the tests require to consume expensive biological and chemical consumables. The rates of false-positive and false-negative are relatively high owing to the contribution of various factors such as nonspecific hydrophobic binding, poor solubility leading to protein or substrate precipitation, aggregation, presence of reactive functional groups, low purity, incorrect structural assignment or compound concentration, interference of the compounds with the assay or its read-out etc... (17).

In parallel of the HTS approaches, virtual screening strategies have developed rapidly both in academic and industrial research. In particular *in silico* methods remain an attractive option for prioritizing structures for focussed screening (18). However, it may also be interesting to perform experimental and virtual screens in parallel, on the same chemical databanks, since comparative studies have shown that these two approaches can lead to identification of different active compounds (19). Today, most pharmaceutical companies have substantial groups devoted to virtual screening approaches (20). In the following, I will briefly review the main topics in the domain of virtual screening.

2.7 Virtual screening

The main aim of these approaches is to quickly and cheaply select a restricted number of molecules expected to bind to the protein target, from large chemical libraries. This smaller set of molecules (typically 5 percent of the original chemical library) are then experimentally tested. Many methods are available, with different ranges of applications, but they can be classified into two classes: structure-based approaches, and ligand-based approaches. In both cases, they require the choice of a molecule library that will be used in the virtual screen.

2.7.1 The molecule library

A molecule library is a set of molecules, whose sizes may vary from a few hundreds to hundreds of thousands of grams per mol, that have been synthesized in large quantities and stored in order to be rapidly available for large scale screening. Chemical libraries can be confidential, like those owns by pharmaceutical industries, composed of synthesized or extracted natural molecules. They can also be commercial libraries. Chemdiv, Chembridge or Asinex are examples of some of the most commonly used molecule libraries. The choice of a molecule library is critical in a virtual screen (also in an experimental screen). The content, design and scale of a molecule library needs to be directly related to the purpose of the screen. Because the goal of the screening evolves during the drug discovery process, the design of the libraries to be screened must be related to the advancement of the project. In order to maximize the probability of identifying structurally different hits, early screening assays must involve diverse libraries giving a broad coverage of the chemical space (21). On the contrary, in the later ”hits to lead” step, targeted libraries made of molecules structurally similar to the identified hits must be designed (22). Because of the development of virtual approaches, molecular libraries are now also provided as virtual libraries, in which molecules are represented in different formats such as smiles, sdf or mol2, and that encode for their chemical structures. Virtual screening methods will use these molecular descriptions, or other descriptions derived from these standard formats, in order to encode molecules and to manipulate them through various algorithms.

2.7.2 Structure-based methods

Structure-based methods, also called docking, cover a range of approaches that exploit the 3D structure of the protein of interest, to predict its potential ligands. The growing numbers of

genomic targets of therapeutic interest (23) and macromolecules (proteins, nucleic acids) for which a three-dimensional structure (3D) is available (24) makes docking increasingly attractive for the identification of bioactive molecules (25; 26).

The role of molecular docking is to predict the active conformation and relative orientation of each molecule of the chemical library within the binding site of the protein of interest. In other words, docking tries to propose a model for the protein-ligand complex, and to evaluate the stability of this complex. Very generally, the search of possible positions of the ligand in the protein structure focuses on a defined protein pocket that has been experimentally determined (for example by directed mutagenesis). These methods use the principle of steric complementarity (Dock, Fred) or of molecular interactions (AutoDock, FlexX, Glide, Gold, ICM, LigandFit, Surflex), to place a ligand in the protein pocket. In most cases, the protein is considered as rigid, although some programs handle flexibility for aminoacid side chains or from small local rearrangements of the backbone. In contrast, ligand flexibility is fully taken into account. Three principles are generally used for the treatment of the flexibility of the ligand:

- a set of conformations of the ligand is calculated beforehand and they are docked to the rigid way the site (eg Fred),
- the ligand is incrementally constructed fragment after fragmentation (eg Dock, FlexX, Glide, Surflex)
- A more or less complete conformational analysis is conducted on the ligand to generate conformations that are most favorable to docking. (eg ICM, Gold, LigandFit).

Typically, several poses for the ligand are generated and ranked by decreasing probability according to a scoring function that tries to estimate the protein-ligand interaction energy. In a docking screen against a protein target, all molecules of a chemical library will be ranked according to their scores, and the 5-10 % molecules of the initial set with the best scores will be viewed as the best molecules for experimental evaluation. In other words, docking can be used as a tool to reduce the size of the molecule library to be experimentally screened, based on the assumption that the best ranked molecules are enriched in true ligands. Docking is an active field of research, because although many programs are available today, there is still a need to improve the methods used to take protein and/or ligand flexibility into account, or to improve the scoring functions. Today, no program has been identified as ”the best” program, and one needs to evaluate the best docking conditions for each project in an ”ad hoc” manner.

2.7.3 Ligand-based approaches

Ligand-based methods exploit prior knowledge of ligands, and non ligands, for a protein of interest, to predict new or better ligands for this protein. Therefore, they cannot be used in early studies, when no ligands have been identified. However, these methods are quite powerful in later stages of drug development, when optimization of lead compounds is searched. Most ligand-based screening methods rely on the comparison of molecules, with the underlying assumption that similar molecules will have similar behaviors: a molecule that is similar to a known ligand will be predicted to bind to the protein target, which is not expected for random compounds. However, the comparison of molecules is not trivial: it relies on how molecules are encoded, and on the method used to measure the similarity. A large variety of tools have been developed to perform these tasks.

2.7.3.1 Descriptors

Descriptors can be used to encode molecules. They are usually calculated from the molecular structure (atoms, bonds, configuration, conformation) or molecular properties (physical, chemical, biological) (27; 28). The descriptors may include atoms and bonds, or the presence or absence of fragments, or other 1D or 2D features. Descriptors may also relate to the 3D arrangement of atoms, when 3D conformation information is taken into account. Ideally, the descriptors should be readily calculable and easily interpretable by computers and by users. They should represent the actual chemical system and take the structure of chemical space into account (29).

1D descriptors are derived from the empirical formula of the molecule (eg C6H6O for phenol). They correspond typically to global molecular properties, such as the molecular weight, hydrophobicity, or general physicochemical properties of the molecule, such as the number of atoms of particular types or hydrogen bond donors and acceptors, solubility (logP). These descriptors bear little information about the structure of the molecule and are essentially used to derive filters such as the Lipinski's rule of five (30), or in combination with other descriptors. It must be noted that these descriptors cannot usually distinguish isomers. The 1D encoding of molecules often take the form of a vector whose elements correspond to these molecular properties used as descriptors (Figure 2.4A). In such vectorial description, similarity between molecules can then be measured according to a Tanimoto coefficient calculated from the elements of these two vectors.

The Tanimoto coefficient is defined as the ratio :

$$T(A,B) = \frac{\sum_{i=1}^{M} A(i)B(i)}{\sum_{i=1}^{M} A(i) + \sum_{i=1}^{M} B(i) - \sum_{i=1}^{M} A(i)B(i)}$$

with A (i) and B (i) are equal to 1 if the i-th descriptor is found in molecule A and B respectively, and 0 if it is absent. M is the total number of descriptors taken into consideration.

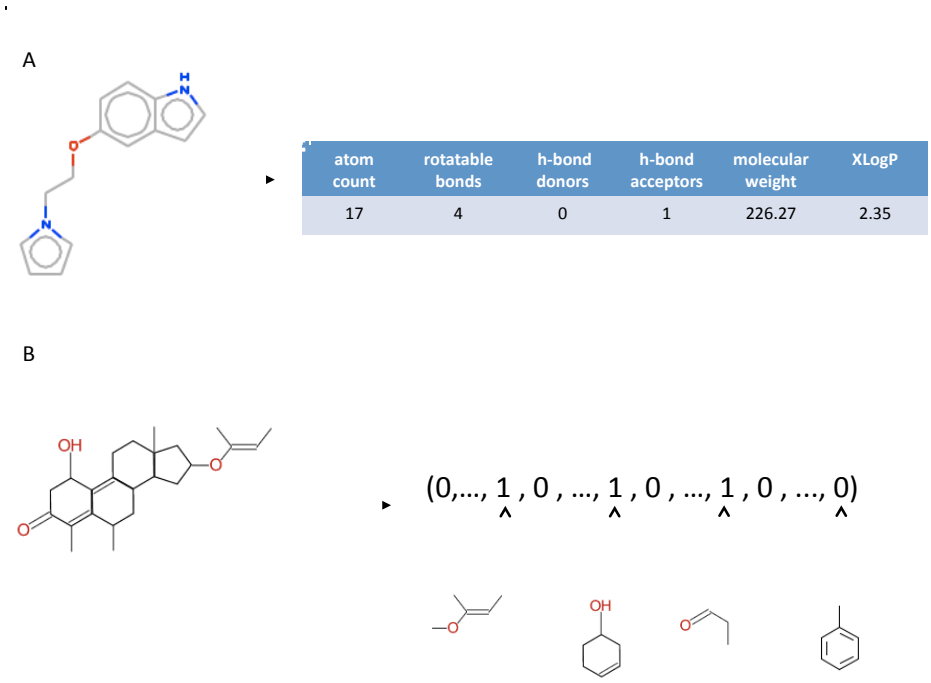
2D descriptors are derived from the 2D structure of the molecule, which can also be viewed as a graph. A first class of 2D descriptors consists of general topological indices, related to the notion of graph invariant in graph theory. Seminal examples include the Wiener and Randic connectivity indices, defined respectively from the length of the shortest path between pairs of atoms, and their degrees in the molecular graph (31). In a related approach, topological autocorrelation vectors measure the autocorrelation of atomic physicochemical properties, such as partial charges or polarity, from pairs of atoms separated by a given topological distance, expressed as the length of the shortest path connecting atoms in the molecular graph (32).

A second class of descriptors represents a molecule by a vector indexed by a set of structural features, and relies on the extraction of substructures from the molecular graph. This process defines a molecular fingerprint, and in practice, two different approaches can be adopted.

The first approach considers a limited set of informative predefined substructures to characterize the molecule. Each substructure is mapped to a bit of the fingerprint, which either accounts for the presence or absence of the substructure in the molecule. A typical implementation is a bistring indexed by 166 predefined substructures known as the MDL MACCS keys (33). This type of encoding is illustrated in (Figure 2.4B). Among the advantages offered by the structural keys is the expressiveness of the features, and the interpretability retained in the representation, because of the one-to-one correspondence between the bins of the vector and the structural features. However, choosing the features to be included in the substructure representation may be challenging in practice. While chemical intuition can be helpful for that purpose (34), this task is more generally related to the problem of graph mining that consists in the automatic identification of interesting structural features within a set of graphs. For chemical applications, such interesting patterns are typically defined as non correlated structures frequently appearing in active compounds, and rarely in inactive compounds (35; 36; 37).

In the alternative approach, molecules are represented by simple structural features called linear molecular fragments, defined as successions of covalently bonded atoms. In this case, typical fingerprints, such as the Daylight fingerprints, characterize a molecule by its exhaustive list of fragments made of up to seven or eight atoms.

In summary, 2D descriptors are calculated from the 2D formula of the molecule, and they provide information about its size, its overall shape and its ramifications. In the case classical of 1D or 2D descriptors encoded in vector representation of molecules, an explicit ”chemical space” is defined in which each molecule is represented by a finite-dimensional vector. These vector representations can be used as such to define similarity measures between molecules such as Tanimoto coefficients.



atom count	rotatable bonds	h-bond donors	h-bond acceptors	molecular weight	XLogP
17	4	0	1	226.27	2.35

Figure 2.4: Example of descriptors that may be used to encode molecules. (A) Example of 1D descriptors based on physicochemical properties. (B) example of 2D descriptors encoding for presence or absence of predefined chemical fragments.

3D descriptors are derived from the 3D structure of the molecules. A first class of three dimensional descriptors requires a preliminary step of molecular alignment, consisting in placing the molecules in a common orientation in the 3D space through operations of rotations and translations. The quality of the alignment is quantified by a scoring function, and the molecules are said to be aligned when it is maximized. Typical scoring functions consider the number of

identical atoms superimposed under a skeleton representation (38), or the overlap of the electron clouds surrounding the molecules (39). In order to handle conformational analysis, the alignment can be flexible, in which case additional degrees of freedom are introduced to handle rotational bonds, or rigid and based on the optimal alignment of pairs of multi-conformers. Aligning molecules can be a quite complex process, and we refer to Lemmen and Lengauer (40) for a review of the existing techniques. Once the molecules are aligned, 3D descriptors can for instance be defined by sampling molecular surfaces according to rays emanating from the center of mass of the aligned molecules (41; 42), or, in the Comparative Molecular Field Analysis (CoMFA) methodology, by measuring the interaction between the molecules and an atomic probe (e.g., a charged or lipophilic atom) at each point of a discrete box enclosing the molecules (43)

An opposite approach consists in extracting descriptors independent of the molecular orientation. Apart from global shape descriptors, such as the Van der Waals volume of the molecule or molecular surfaces areas, most alignment independent descriptors are based on distances between atoms. For example, an early study proposed to characterize a molecule by its matrix of inter-atomic distances (44). While the authors propose several methods to compare such matrices, this approach is not convenient because it does not lead to a fixed size representation of the molecules. Standard vectorial representations can be derived by considering pairs of atoms of the molecule. Topological autocorrelation vectors can for instance be extended to 3D autocorrelation vectors, computing the autocorrelation of atomic properties from pairs of atoms within a specified Euclidean distance range, instead of a given topological distance on the molecular graph (45). Other representations are based on counting the number of times pairs of atoms of particular types are found within predefined distance ranges in the 3D structure of the molecule (46; 47; 48). Considering molecular features based on triplets or larger sets of atoms leads to the notion of pharmacophore. A pharmacophore is usually defined as a three-dimensional arrangement of atoms - or groups of atoms - responsible for the biological activity of a drug molecule (49). Typical pharmacophoric features of interest are atoms having particular properties (e.g., positive and negative charges or high hydrophobicity), hydrogen donors and acceptors and aromatic rings centroids (50). In this context, pharmacophore fingerprints were proposed as the three-dimensional counterpart of molecular fragment fingerprints. Pharmacophore fingerprints represent a molecule by a bitstring encoding its pharmacophoric content, usually defined as the exhaustive list of triplets of pharmacophoric features found

within a set of predefined distances ranges in its 3D structure (51; 52). Strictly speaking, pharmacophore fingerprints encode putative pharmacophores of the molecules, and because the number of potential pharmacophores can be very large, they are usually compressed (53; 54). In chapter 3, we will present and use 2D and 3D pharmacophore representations for molecules that were developed in our laboratory (55) and that we used in this thesis in a chemogenomics framework.

A vast amount of descriptors has therefore been proposed in the literature. The above presentation if far from being exhaustive, and we refer interested readers to the textbooks for a detailed presentation (31; 56). Choosing "good" descriptors for the task to be performed remains nevertheless an open question. For instance, even though the molecular mechanisms responsible for the binding of a ligand to a target are known to strongly depend on their 3D complementarity, different studies account for the superiority of 2D fingerprints over pharmacophore fingerprints in this context (34; 52; 55). This observation suggests that 2D fingerprints might encode to some extent three-dimensional information (27), and in many cases, they actually constitute the "gold-standard" representation of chemical structures. Another explanation is that 3D approaches require to know the 3D geometry of the molecule in its "active" conformation, which is not always available. In such cases, the choice of other conformations such as the most free-state conformation might degrade the performance of 3D approaches.

2.7.3.2 Principle of ligand-based approaches

Many "rational" drug design efforts are based on a principle which states that structurally similar compounds are more likely to exhibit similar properties. Indeed, the observation that common substructural fragments lead to similar biological activities can be quantified from database analysis. A variety of methods, known collectively as Quantitative Structure Activity Relationship (QSAR) have been developed, essentially for the search for similarities between molecules in large databases of existing molecules whose properties are known. The discovery of such a relationship allow to predict the physical and chemical properties of biologically active compounds, and to develop new theories or to understand the phenomena observed. Once a QSAR model has been built to encode this relationship between the chemical space and a given biological activity, this can guide the synthesis of new molecules, limiting the number of compounds to synthesize and test.

The relationship between the structures of molecules and their properties or activities are usually established using methods of statistical learning. The usual techniques are based on

the characterization of molecules through a set of 1D, 2D or 3D descriptors. A model is established, that relates the descriptors that encode the molecule, to its biological activity, based on a learning dataset of molecules for which this activity is known. It is then possible to use this model to predict the activity of a new molecule. Numerous studies show that it is impossible to predict accurately the affinity of chemically diverse ligands (57). It is reasonable to hope to discriminate affinity of ligands in the range of nanomolar, micromolar and millimolar.

Decades of research in the fields of statistics and machine learning have provided a profusion of methods for that purpose. Their detailed presentation is far beyond the scope of this section, and we invite interested readers to refer to the classical textbooks (58; 59) for a thorough introduction. In this section we just give general methodological and historical considerations about their application in chemoinformatics.

Models can be grouped into two main categories depending on the nature of the property to be predicted. Models predicting quantitative properties, such as for instance the degree of binding to a target, are known as regression models. On the other hand, classification models predict qualitative properties. In SAR analysis, most of the properties considered are in essence quantitative, but the prediction problem is often cast into the binary classification framework by the introduction of a threshold above which the molecules are said to be globally active, and under which globally inactive. In the following, the term classification implicitly stands for such binary classification.

In order to build the model, the pool of molecules with known activity is usually split into a training set and a test set. The training set is used to learn the model. The learning problem consists in constructing a model that is able to predict the biological property on the molecules of the training set, but without over-learning on it. This overfitting phenomenon can for instance be controlled using cross-validation techniques, that quantify the ability of the model to predict a subset of the training set that was left out during the learning phase. The test set is used to evaluate the generalization properties of the learned model, corresponding to its ability to make correct prediction on a set of unseen molecules. Different criteria can be used for this evaluation. In regression, it is typically quantified by the correlation between the predicted and the true activity values. In the classification framework, a standard criterion is the accuracy of the classifier, expressed as the fraction of correctly classified compounds. However, if one of the two classes is over-represented in the training set, and/or the cost of misclassification are different, it might be safer to consider the true and false positive and negative rates of classification. The true positive (resp. negative) rate account for the fraction of compounds

of the positive (resp. negative) class that are correctly predicted, and the false positive (resp. negative) rate accounts for the fraction of compounds of the negative (resp. positive) class that are misclassified. In virtual screening applications for instance, where we typically do not want to misclassify a potentially active compound, models with low false negative rates are favored, even it they come at the expense of an increased false positive rate.

Because they usually require a limited set of uncorrelated variables as input, applying these models to chemoinformatics requires to summarize the information about the molecules into a limited set of features, which may not a trivial task due to the vast amount of possible molecular descriptors. A popular way to address this problem in chemoinformatics is to rely on principal component analysis (PCA), that defines a limited set of uncorrelated variables from linear combinations of the initial pool of features, in a way to account for most of their informative content. Alternatively, feature selection methods can be used to identify among an initial pool of features a subset of features relevant with the property to be predicted. Because molecular descriptors are sometimes costly to define, a potential advantage of feature selection methods, over PCA-based approaches, is the fact that they reduce the number of descriptors to be computed for the prediction of new compounds.

Let us now introduce different methods that have been applied to model SAR. The first SAR model was developed in 1964 by Hansch and coworkers who applied a multiple linear regression (MLR) analysis to correlate the biological activity of a molecule with a pair of descriptors related to its electronic structure and hydrophobicity (60). MLR models are still widely applied to model SAR. PCA is commonly used as inputs, in the so-called PC- regression models (61). Moreover, genetic algorithms have been introduced to perform feature selection as an alternative to standard forward selection or backward elimination approaches (62). Related linear approaches can be applied to the classification framework with discriminant analysis algorithms (63). However, because this class of models is limited to encode linear relationships, they can be too restrictive to efficiently predict biological properties. While the models can be enriched with the application of nonlinear transformations of the input variables (64), SAR analysis greatly benefited from the development of nonlinear methods, and in particular artificial neural networks (ANN). Early applications of back-propagation ANN accounted for their predictive superiority over standard linear regression techniques. Many studies have demonstrated the strength of ANN to predict biological properties, and they are now a standard tool to model SAR (65; 66).

Despite their predictive efficiency, a major criticism to ANN is their lack of interpretability, which can be of great importance in chemistry in order to understand the biological mechanisms responsible for the activity. An alternative class of models builds a classifier expressed as a set of rules relating the molecular structure and the biological activity. Such models have been derived for instance using decision trees algorithms (67). From the practical viewpoint, another criticism that can be made to ANN is the fact that they require some expertise, concerning for instance the choice of an architecture, in order to be knowledgeably deployed. Moreover, they are known to be prone to overfitting and are hard to reproduce, because of their random initialization and possible convergence to local minima (68). These theoretical issues are to some extent addressed by the support vector machine (SVM) algorithm, known in particular to avoid the problem of local minima, to prevent overfitting, and to offer a better control of the generalization error (69). Moreover, although its good parametrization remains a crucial point, this algorithm requires less amount of expertise to be deployed. The introduction of SVM in SAR analysis was pioneered by Burbidge and co-workers (68). In this study, the SVM algorithm outperforms several ANN architectures for a particular classification task, related to the ability of molecules to inhibit a biological target. Over the last few years, SVM was shown to be a powerful tool for SAR analysis, often outperforming ANN in classification and regression frameworks. We give in the next section a brief introduction to the SVM algorithm, because we used this algorithm in the chemogenomic approach presented in chapter 3.

2.7.4 Introduction to SVM in virtual screening

One important contribution of this thesis is to explore the use of machine learning algorithms within the newly introduced chemogenomic framework, in order to predict protein-ligand interactions. The principle of chemogenomic approaches will be presented in chapter 3. Although this principle is quite simple (i.e. similar proteins are expected to bind similar ligands), to our knowledge, only a very limited number of studies propose computational methods able to handle chemogenomic data and to perform predictions. The main reasons are that these data are not trivial to generate, to manipulate, and to be used as input in computational methods in a relevant manner in order to make predictions.

We have proposed to use kernel methods in the the context of Support Vector Machine (SVM) methods, because, as it will be explained in chapter 3, they allow easy manipulation and calculation in the chemogenomic space (i.e. the chemical space of small molecules joined to the biological space of proteins). Presenting the full mathematical framework of SVM and

kernel methods is out of the scope of this thesis. However, in this section, we will briefly introduce kernel methods in the context of Support Vector Machine (SVM), because we have proposed to use this type of algorithm to implement a computational method able to handle the chemogenomic data.

The support vector machine algorithm was initially introduced in 1992 by Vapnik and co-workers in the binary classification framework (70), (69). Over the last decade this method has been gaining considerable attention in the machine learning community, which led to the emergence of a whole family of statistical learning algorithm called kernel methods (71), (72). SVM has been successfully aplied in many real world applications, including, for instance, optical character recognition (73), text-mining (74) or bioinformatics (75), often outperforming state-of-the-art approaches.

In the recent years, support vector machines and kernel methods have gained considerable attention in cheminformatics. They offer generally good performance for problems of supervised classification, and provide a flexible and computationally efficient framework to include relevant information and prior knowledge about the data and about the problems to be handled.

We start this section by a brief introduction to SVM in the binary classification framework, and, in a second step, we highlight the particular role played by kernel functions in the learning algorithm.

Let us introduce SVM on the example of predicting ligands for a protein, although all the following discussion has a more general application. We consider the case of a protein for which a learning database is available, that consists in two lists of molecules: a list of ligands, and a list of non ligands (Figure 2.5).

Each molecule can be represented by a set of descriptors, as explained in section 2.7.3.1. These molecules can thus be seen as points in a multidimensional space where each dimension corresponds to one of the descriptors. The idea of SVM is to find a linear separation between these two sets of points in the space defined by the descriptors. For example, in the case of two descriptors, Figure 2.6 illustrates a possible hyperplane that separates the active and inactive molecules represented by black and white dots. Once such a linear separation is found, we can predict the activity of a new molecule by mapping it to the space of descriptors and checking on which side of the hyperplane it falls.

A dataset is called linearly separable if it is possible to find a linear separation between the two classes of points (as in Figure 2.6). In such a case, however, there may be many

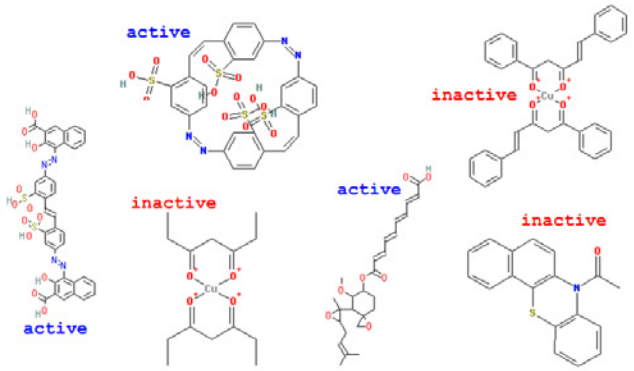


Figure 2.5: active and inactive molecules for a given target

hyperplane that can separate the two sets of points, and the question is how the ”best” separator can be defined (Figure 2.7).

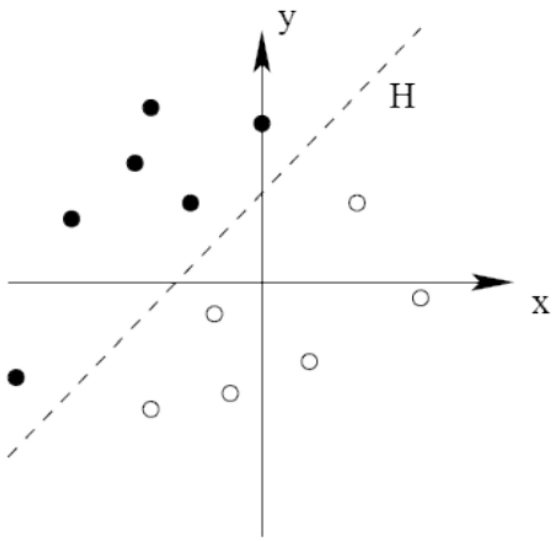


Figure 2.6: Diagram of a hyperplane that separates the two sets of points

SVM implements a particular strategy to define the ”best” separating hyperplane: it chooses the one that correctly separates the data (all points of the same class are on the same side of the separator), while maximizing its distance to the closest points of each class (Figure 2.8).

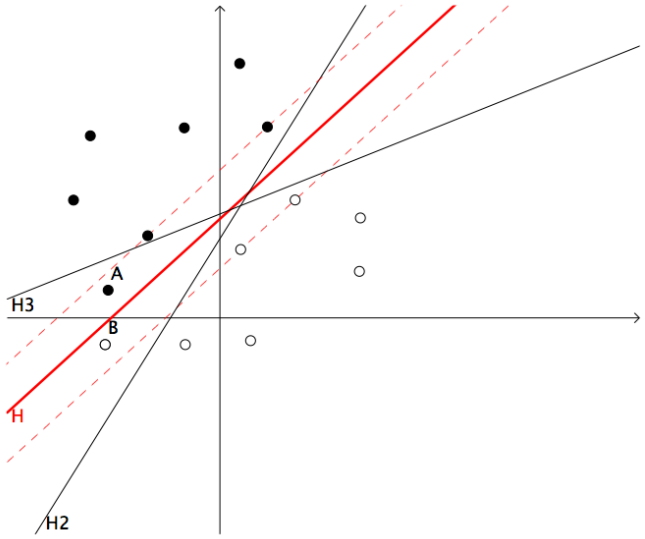


Figure 2.7: The diagram shows that with an optimal hyperplane H (in red), new examples (A and B) are correctly classified although it falls within the margin while with other hyperplanes (H2 and H3) new examples (A and B) may be misclassified.

This distance is called the ”margin”, and SVM are therefore often referred to as large margin classifiers. Intuitively, the margin is related to the confidence we make in the separation. The nearest points to the hyperplane are called support vectors and have given their name to the method.

The above discussion relies on the hypothesis that it is possible to separate the training data using a linear classifier. This is unfortunately not true in most real-life applications, e.g., because of noise in the input data or label errors. SVM implement a slightly different strategy in that case: it searches the hyperplane of largest margin while tolerating a few errors on the training data. Note that larger margin usually leads to more errors, and a trade-off must therefore be set between these two opposite objectives. This trade-off is controlled by a user-defined parameter, usually called C in the case of SVM: larger C correspond to fewer errors but smaller margins, while smaller C correspond to larger margin but more errors. The choice of the optimal value for C is typically determined by cross-validation.

Learning linear classifier is however not well adapted in cases where data are intrinsically non-linearly separable, as illustrated in Figure 2.9. In such cases, it would be beneficial to directly learn a non-linear function to separate positive and negative examples. An intuitive

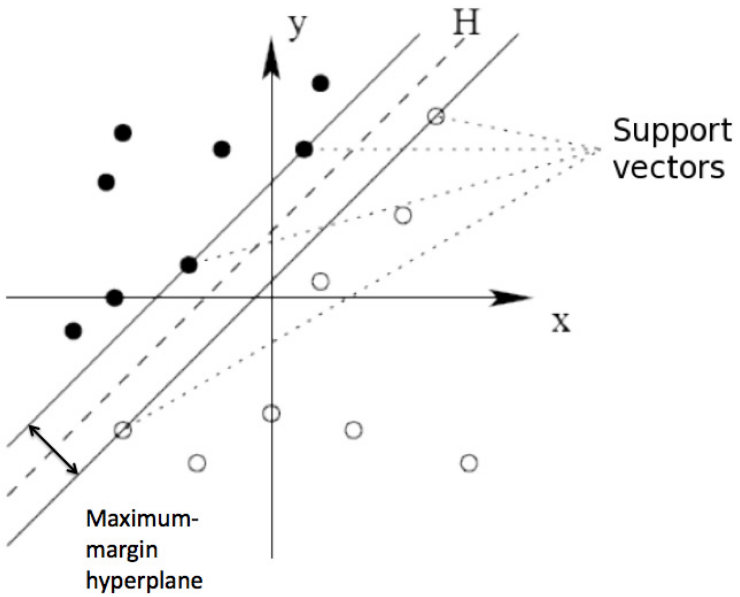


Figure 2.8: Support vectors used to determine the hyperplane

way to tackle this problem is to remember that the positions in space of the data points are tightly linked to the method used to encode the data (for example to encode molecules). In other words, there could exist some other representation of molecules (i.e. in another space called the feature space), in which the images of the data might become linearly separable. SVM implement this idea by finding a new representation of the classification problem which can be solved by a linear classifier, as illustrated in Figure 2.10

In practice, SVM do not explicitly perform a non-linear transformation of the input points to make them linearly separable: instead, they perform this non-linear embedding implicitly, thanks to so-called positive definite kernels. We will not present here the full mathematical basis of kernels and SVM with kernels, but we will only recall, and admit, the most important definitions and properties of these methods, and underline their interest for our prediction problem.

A kernel is a real-valued function $K(x, x')$ that, intuitively, measures the similarity between any two data points x and x'. In our case, x and x' are two molecules. In addition, the kernels that can be used by SVM must fulfill two mathematical conditions: they must be symmetric,

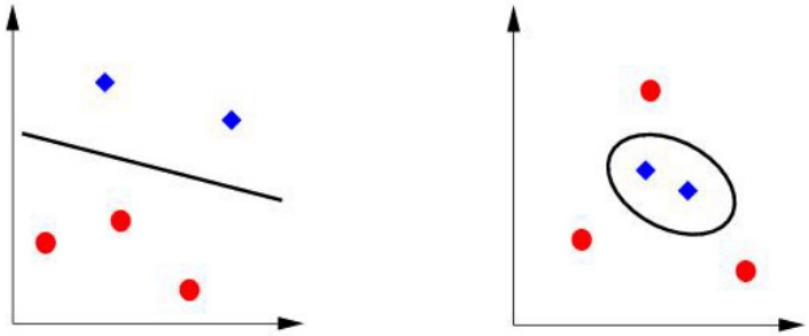


Figure 2.9: The right diagram shows an example of a separable case.The left diagram shows an example of a nonlinearly separable case

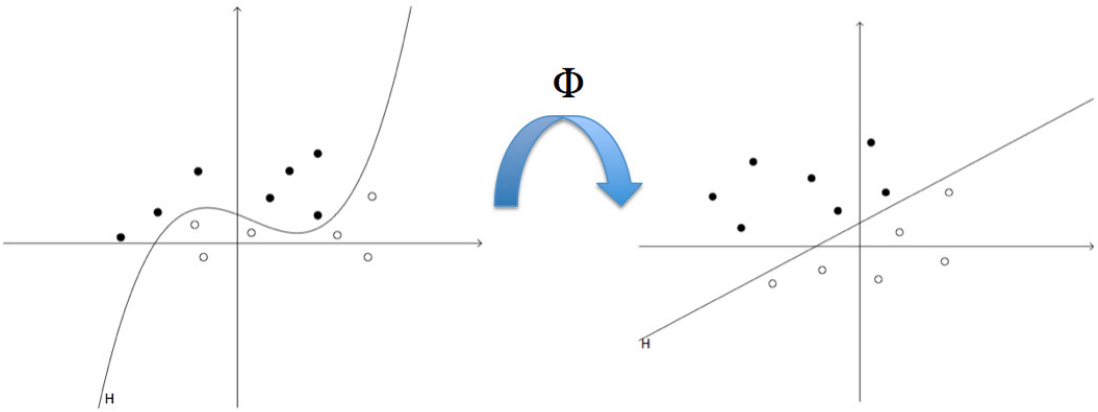


Figure 2.10: Left:nonlinear separation problem in the input space. Right: linear separation problem in the feature space

in the sense that for any two data x and x',

$$K(x, x') = K(x', x),$$

and in addition they must be positive definite in the sense that for any N data $(x_1, x_2, \ldots, x_N)$ and any N real numbers $(a_1, a_2, \ldots, a_N) \in \mathbb{R}^N$ the following holds:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j K(x_i, x_j) \geq 0.$$

A fundamental result states that each positive definite kernel is associated to an embedding of the input data into a feature space. More precisely, the value of the kernel $K(x, x')$ between to data x and x' is exactly the inner product between the two data mapped to the feature space. Note that the embedding may be linear or nonlinear. In other words, defining a positive definite kernel is equivalent to implicitly define a feature space to map data, instead of explicitly computing descriptors in this sapce.

SVM can be used to estimate a linear classifier in the feature space associated to positive definite kernels. The beauty of this approach is that finding the optimal separating hyperplane in the feature space can be done without computing explicitly the descriptors: instead, it suffices to be able to compute inner product between any two points in the feature space to train a SVM. Since by definition the inner product between two data x and x' in the feature space is equal to the kernel $K(x, x')$, SVM classifiers can be learned using only the kernel to represent the input points.

More precisely, for a given training set of points $x_1, \ldots, x_n$, it can be shown that the equation of the SVM hyperplane in the feature space admits a representation of the form:

$$\forall x \in \mathcal{X},\ f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)\,.$$

The $(\alpha_i)_{1,\ldots,n}$ are found by solving a convex optimization problem. This solution hyperplane separates the feature space into two subspaces. Images of positive examples (i.e. ligands for a protein) are situated in the half-space of points for which $f(x)$ is positive while images of negative examples (non-ligands) are situated in the half-space of points for which $f(x)$ is negative. Therefore, to predict the class of a new molecule x, one needs to calculate the value $f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$ and assign the class corresponding to the sign of $f(x)$.

Using SVM to predict the binding of ligands to a target therefore requires to define a kernel (i.e., a similarity measure) between molecules. Relevant kernels should implicitly encode

biochemical characteristics of molecules that are relevant with respect to the protein - ligand interaction mechanism.

In the case where molecules are encoded as vectors using classical molecular descriptors, there exists several families of kernel functions including polynomial, Gaussian, Laplacian and sigmoid, that can be tested for the problem of interest. This is relevant to infer nonlinear functions of existing descriptors.

Alternatively, one may go beyond existing descriptors and directly develop kernel functions to compare molecules. For example, new kernels were developed in the recent years based on direct comparison of the 2D or 3D structures of molecules. Indeed, molecules can be encoded by vectors, but they can also be encoded by other representations such as their atom coordinates in the 3D space. If a similarity measure K between molecules can be defined on such a representation, it can be used directly with a SVM. Examples of such original kernels for molecules developed in our laboratory are given and used in the next chapter.

3

Virtual screening of GPCRs: an *in silico* chemogenomic approach

In this chapter, we will present a chemogenomic approach that allows to predict protein-ligand interaction, by learning from known interactions not only for the protein of interest, but also from other proteins belonging to the same family. We will show that such sharing of information between proteins improves the prediction performances with respect to the classical approach that makes predictions for proteins independently, i.e. one by one. We will take the example of the GPCR family of proteins, because this important family is one of the target of many drugs. We will also show that the chemogenomic approach allows to perform predictions in the difficult case of orphan receptor, a case that cannot be handled by classical approaches.

3.1 Introduction to GPCRs

G-protein-coupled receptors (GPCRs) form one of the largest and most diverse family of proteins. Their main role is the signal transduction from outside to inside the cell, in response to stimuli. In mammals, this family of proteins is particularly important.

In humans, the GPCR superfamily is comprised of an estimated 600-1,000 members and is the largest known class of molecular targets with proven therapeutic value. They are ubiquitous in our body, being involved in regulation of every major mammalian physiological system (76), and play a role in a wide range of disorders including allergies, cardiovascular dysfunction, depression, obesity, cancer, pain, diabetes, and a variety of central nervous system disorders (77; 78; 79).

3. VIRTUAL SCREENING OF GPCRS: AN *IN SILICO* CHEMOGENOMIC APPROACH

Their location on the cell surface gives them an important role in chemical sensing, essential for communication between cells. This location makes them readily accessible to drugs, and 30 GPCRs are the targets for the majority of best-selling drugs, representing about 40% of all prescription pharmaceuticals on the market (80). Besides, the human genome contains several hundreds unique GPCRs which have yet to be assigned a clear cellular function, suggesting that they are likely to remain an important target class for new drugs in the future (81).

The ligands that bind and activate these receptors include light, neurotransmitters, odorants, biogenic amines, lipids, proteins, amino acids, hormones, nucleotides, chemokines and many others. (82) They vary in size from small molecules to peptides and to large proteins.

GPCRs can be grouped into 6 classes based on sequence homology and functional similarity (83):

- Class A (or 1) (Rhodopsin-like)
- Class B (or 2) (Secretin receptor family)
- Class C (or 3) (Metabotropic glutamate/pheromone)
- Class D (or 4) (Fungal mating pheromone receptors)
- Class E (or 5) (Cyclic AMP receptors)
- Class F (or 6) (Frizzled/Smoothened)

Family A is by far the largest group, and includes the receptors for light (rhodopsin) and adrenaline (adrenergic receptors) and, indeed, most other GPCR receptor types, including the olfactory subgroup. The olfactory receptors constitute most of these sequences, but nearly 200 non-olfactory GPCRs that recognize over 80 distinct ligands have also been functionally characterized.

Family B contains the receptors for the gastrointestinal peptide hormone family (secretin, glucagon, vasoactive intestinal peptide (VIP) and growth-hormone-releasing hormone), corticotropin-releasing hormone, calcitonin and parathyroid hormone.

Family C is also relatively small, and contains the metabotropic glutamate receptor family, the GABAB receptor, and the calcium-sensing receptor, as well as some taste receptors. All family C members have a very large extracellular amino terminus that is crucial for ligand binding and activation.

Family D, containing the mating factor receptors, STE2 and STE3, are integral membrane proteins that may be involved in the response to mating factors on the yeast cell membrane (84). The amino acid sequences of both receptors contain high proportions of hydrophobic residues grouped into 7 domains, in a manner reminiscent of the rhodopsins and other receptors believed to interact with G-proteins. (85)

Family E consists of Cyclic AMP receptors and is a distinct family of G-protein coupled receptors. The cyclic AMP receptors coordinate aggregation of individual cells into a multicellular organism, and regulate the expression of a large number of developmentally-regulated genes (86). The amino acid sequence of these receptors contain high proportions of hydrophobic residues grouped into 7 domains, as in rhodopsins. However, while a similar 3D framework has been proposed to account for this, there is no significant sequence similarity between these families: the cAMP receptors thus bear their own unique signature.

Family F contains in particular the frizzled family of GPCR proteins (87). The group frizzled appears to be the most ancestral form of GPCRs. GPCRs from family F serve as receptors in the Wnt signaling pathway. Some Wnt proteins inhibit the degradation of beta-catenin, which can regulate transcription of specific genes. Other Wnt exert their influences in other ways, such as increasing intracellular concentrations of Ca2+ and decreasing intracellular concentrations of cyclic guanosine monophosphate (cGMP) (88).

3.2 GPCRs and signal transduction

On the basis of homology with rhodopsin, GPCR are predicted to be integral membrane proteins sharing a common global topology that consists of seven transmembrane alpha helices, an intracellular C-terminal, an extracellular N-terminal, three intracellular loops and three extracellular loops. The GPCR arranges itself into a tertiary structure resembling a barrel, with the seven transmembrane helices forming a cavity within the plasma membrane which serves a ligand-binding domain that is often covered by EL-2. This gives rise to their other names, the 7-TM receptors or the heptahelical receptors (Figure 3.1).

GPCRs transduce extracellular stimuli to give intracellular signals through interaction of their intracellular domains with heterotrimeric G proteins. (82) The transduction of the signal through the membrane by the receptor is not completely understood. It is known that the inactive G protein is bound to the receptor in its inactive state. Once the ligand is recognized, the receptor shifts conformation and, thus, mechanically activates the G protein, which detaches

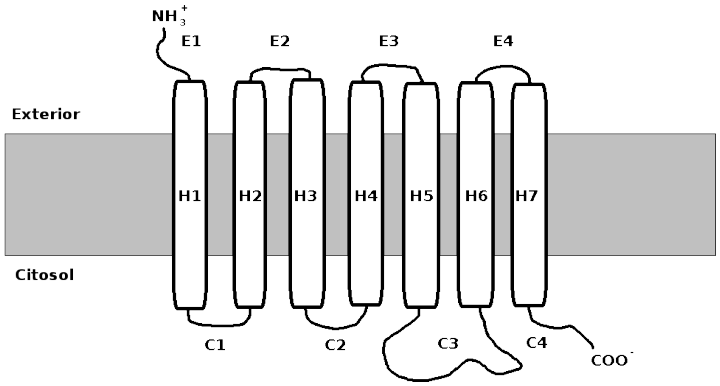


Figure 3.1: Scheme of a gpcr

from the receptor. The receptor can now either activate another G protein or switch back to its inactive state. Traditionally, the GPCRs are expected to exhibit specificity for the transmitters (Goldstein, 1974). This specificity results from evolutionary processes that aim at diversifying the intercellular interactions.

A ligand (i.e. a transmitter molecule) may bind to more than one GPCR, and in such cases, these GPCRs usually share high degree of similarity and belong to the same subfamily. Reciprocally, a given GPCR may bind more than one transmitter molecule, but then, these transmitters usually share structural similarities and are often part of the same synthesis pathway, as in the case of the neuropeptides synthesized from the same precursor (Douglass et al., 1984). A classical example for structurally similar ligands binding to related receptors are the opioid receptors and the natural opioid peptides, respectively (89).

Ligand binding is followed by a change in the conformation of the receptor that may involve disruption of a strong ionic interaction between the third and sixth transmembrane helices (90; 91), which facilitates activation of the G-protein heterotrimer. Depending on the type of G protein to which the receptor is coupled, a variety of downstream signaling pathways can be activated (reviewed by (92; 93)). Signaling is then attenuated (desensitized) by GPCR internalization, which is facilitated by arrestin binding (94). Signaling, desensitization and potential resensitization are regulated by complex interactions of various intracellular domains of the GPCRs with numerous intracellular proteins (95; 96).

3.3 Targeting GPCRs

In silico prediction of interactions between small molecules and GPCRs is not only of particular interest for the drug industry, but also a useful step for the elucidation of many biological process. First, it may help to decipher the function of so-called *orphan* GPCRs (89), for which no natural ligand is known. Second, once a particular GPCR is selected as a target, it may help in the selection of promising molecule candidates to be screened *in vitro* against the target for lead identification.

Virtual screening of GPCRs is however a daunting task, both for receptor-based approaches (i.e. docking) and for ligand-based approaches. The former relies on the prior knowledge of the 3D structure of the protein, in a context where only two GPCR structures (for bovine rhodopsin and for human β_2-adrenergic receptor) were known when the present work was undertaken. Indeed, GPCRs, like other membrane proteins, are notoriously difficult to crystallize. As a result, docking strategies for screening small molecules against GPCRs are often limited by the difficulty to correctly model the 3D structure of the target. To circumvent the lack of experimental structures, various studies have used 3D structural models of GPCRs built by homology modeling using bovine rhodopsin as a template structure. Docking a library of molecules into these modeled structures allowed the recovery of known ligands (97; 98; 99; 100), and even identification of new ligands (101; 102). However, docking methods still suffer from docking and scoring inaccuracies, and homology models are not always reliable-enough to be employed in structure-based virtual screening. Methods have been proposed to enhance the quality of the models for docking studies by global optimization and flexible docking (98), or by using different sets of receptor models (100). Nevertheless, these methods have been applied only to class A receptors and they are expected to show limited performances for GPCRs sharing lower sequence similarity with rhodopsin, especially in the case of receptors belonging to classes B, C and D.

Alternatively, ligand-based strategies, in particular quantitative structure-activity relationship (QSAR), attempt to predict new ligands from previously known ligands, often using statistical or machine learning approaches. Ligand-based approaches are interesting because they do not require the knowledge of the target 3D structure and can benefit from the discovery of new ligands. However, their accuracy is fundamentally limited by the amount of known ligands, and degrades when few ligands are known. Although these methods were successfully used to retrieve strong GPCR binders (103), they are efficient for lead optimization within a previously

identified molecular scaffold, but are not appropriate to identify new families of ligands for a target. At the extreme, they cannot be pursued for the screening of orphan GPCRs.

3.4 Introduction to *in silico* chemogenomic approach

Instead of focusing on each individual target independently from other proteins, a recent trend in the pharmaceutical industry, often referred to as chemogenomics, is to screen molecules against several targets of the same family simultaneously (104; 105).

In this chapter we will discuss this type of approach which consists in making interaction predictions between a ligand and a GPCR using all information available on the entire family of GPCRs, i.e. all known interactions between GPCRs and ligands.

The underlying paradigm in chemogenomics is that "similar receptors bind similar ligands" (106). Or in other words, for a given receptor, known ligands for similar receptors, may serve as a starting point to predict potential ligands.

The systematic screening of interactions performed by chemogenomics between the chemical space of small molecules and the biological space of protein targets, can be thought of as an attempt to fill a large 2D interaction matrix, where columns correspond to targets, rows to small molecules, and the (i, j)-th entry of the matrix indicates whether the j-th molecule can bind the i-th target (Figure 3.2).

While in general the matrix may contain some description of the strength of the interaction, such as the association constant of the complex, we focussed on a simplified description that only differentiates binding from non-binding molecules, which results in a binary matrix of target-molecule pairs. This matrix is already sparsely filled with our current knowledge of protein-ligand interactions, and chemogenomics attempts to fill the holes. While classical docking or ligand-based virtual screening strategies focus on each single column independently from the others in this matrix, i.e. treat each target independently from the others, the chemogenomic approach is motivated by the observation that similar molecules can bind similar proteins. Therefore, information about a known interaction between a ligand and a GPCR could be a useful hint to predict interaction between similar molecules and similar GPCRs. This can be of particular interest when, for example, a particular target has few or no known ligands, but similar proteins have many. In that case, it is tempting to use the information about the known ligands of similar proteins for a ligand-based virtual screening of the target of interest.

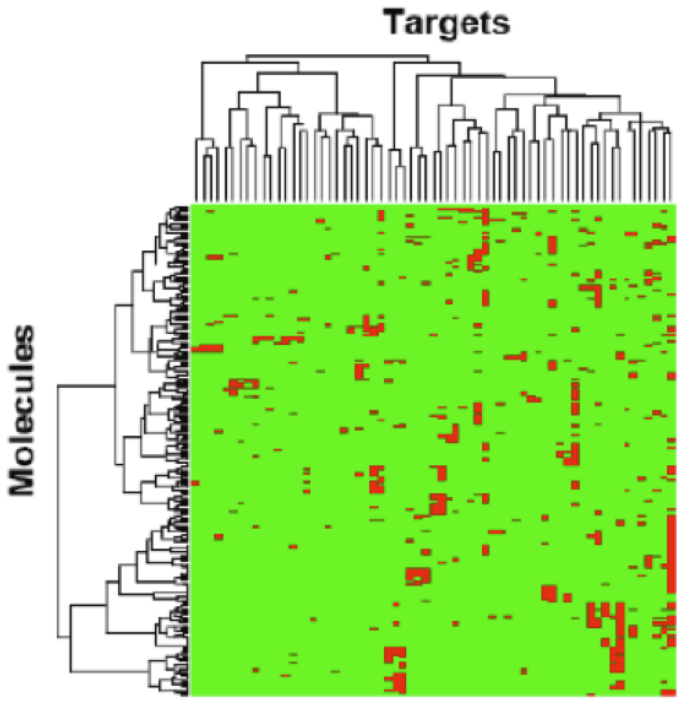


Figure 3.2: Example of matrix of known interactions between molecules and targets. Red squares correspond to known interactions, green squares correspond to unknown interactions

In this context, we can formally define *in silico* chemogenomics as the problem of predicting interactions between a molecule and a ligand (i.e. a hole in the matrix) from the knowledge of all other known interactions or non-interactions (i.e. the known entries of the matrix).

This intuitive concept however raises many questions. How are described molecules and targets ? And how can receptor or ligand similarity be defined ?

Chemogenomics is an interdisciplinary field that attempts to answer these questions and exploit the answers for the accelerated discovery of novel chemical starting points or lead series (104; 107). One of the underlying questions is that chemogenomics require to propose description of the protein space, i.e. a method to encode proteins, and also a method to measure similarities between these objects. Similarly, one needs to encode the chemical space, in order to describe molecules, and define a method to measure similarities between molecules.

As an illustration of protein encoding, a classical and widely used description for proteins is the amino acid sequence, the similarity between proteins being then measured according to their sequence similarity, as evaluated using for example a BLAST score

A variety of methods have been proposed in QSAR approaches to encode a molecule by a multidimensional vector whose elements bear various physical or chemical properties of this molecule, such as its molecular weight, solubility, or presence or absence of various chemical groups, as seen in chapter 2. Classically, similarity between molecules is then measured according to a Tanimoto coefficient calculated from the elements of these two vectors (see section 2.7.3.1)

In chemogenomics, we do not only need to tackle the problem of encoding proteins and molecules, and to define their associated similarity measures. In fact, we consider the space of target - molecule pairs (t, m) and we attempt to predict new (t, m) pairs based on their similarity (in the space of pairs) to other pairs known to interact or not. This raises the question of how the space of (target-molecules) pairs is encoded, and how similarities between pairs can be defined.

A simple idea is to use a vector corresponding to the concatenation of the vectors that represent the protein target and of the vector that represents the molecule.

For example, if $m(a_1, ..., a_n)$ is a vector that describes a ligand and $t(b_1, ..., b_p)$ a vector that describes a protein target, a couple is then defined by $(t, m) = (a_1, ..., a_n, b_1, ..., b_n)$. This corresponds to encoding the chemogenomics space of (t, m) pairs by joining the biological space and the chemical space.

This strategy was pioneered by (108). (109) predicted ligands of orphan GPCR. They merged descriptors of ligands and targets to describe putative ligand-receptor complexes, and used Support Vector Machine (SVM) to discriminate real complexes from ligand-receptors pairs that do not form complexes.

(110) proposed to encode the chemogenomic space by the tensor product of the biological space and the chemical space. In this case, the vector representing a protein-ligand pair corresponds to the tensor product of the vectors encoding the protein and that encoding the ligand.

For example, if $m(a_1, ..., a_n)$ is a vector describing a ligand and $t(b_1, ..., b_p)$ a vector describing a protein, a couple is then defined as $(t, m) = (a_1b_1, ..., a_ib_j, ..., a_nb_p)$ with i and j varying from 1 to n and from 1 to p respectively.

Therefore, the dimension of the chemogenomic space encoded as the tensor product space is $n \times p$, whereas it is $n + p$ in the case of the joined space.

(110) introduced for the first time the use of the tensor product space to encode protein - ligand pairs. They applied this representation to the case of GPCRs and their ligands. They encoded proteins and ligands in their respective biological and chemical spaces. In both spaces, they defined a similarity measure using positive definite kernels (see section 2.7.4 for definition of kernels). In this case, they showed that one could define a similarity measure (or kernel) in the chemogenomic space of (t, m) pairs by the product of the similarity measures in the biological and chemical spaces :

$$K_{pair}\left((t, m), (t', m')\right) = K_{target}(t, t') \times K_{molec}(m, m'). \tag{3.1}$$

where $K_{target}(t, t')$ is the similarity measure between the two proteins and $K_{molec}(m, m')$ is the similarity measure between the two ligands.

In this framework, the computational time required to measure the similarity in the chemogenomic space described as a tensor product space is equal to that required by a description in a joined space (i.e. increasing as $n + p$ and not $n * p$).

However, (110) was not able to show significant benefits of this chemogenomic approach with respect to the individual approach that learns a separate classifier for each GPCR, except in the case of orphan GPCRs for which their approach performed better than the baseline random classifier (as already mentioned, the individual approach cannot make predictions for orphan receptors).

One possible explanation could be that the way proteins and ligands were encoded in this study, and the way similarity measures were defined, were not well suited to the problem. This underlines the importance of how the chemogenomic space is encoded, and how similarity measure is defines in this space because this has a strong impact on the prediction performance.

In a preliminary study (111) performed in our laboratory , we used a chemogenomics approach to predict interaction between various potential targets including GPCRs, enzymes and ion channels. Using descriptors for targets based on their sequences or based on the classification of proteins families, we obtained predictors that were more accurate than state-of-the-art individual methods both for the orphan targets and for the targets for which some ligands were already known.

This illustrates the general importance of choosing relevant descriptors for proteins and ligands, which is true not only in chemogenomic approach, but also in any *in silico* virtual screening projects.

In this chapter, we went one step further in this direction and present an *in silico* chemogenomic approach specifically tailored for the screening of GPCRs although the method could in principle be adapted to other classes of therapeutic targets.

We tested 2D and 3D descriptors to describe molecules, and five ways to describe GPCRs, including a description of their relative positions in current hierarchical classifications of the GPCR superfamily, and information about key residues likely to be in contact with the ligand. We evaluated the performance of all combinations of these descriptions on the data of the GLIDA database (112), which contains 34686 reported interactions between human GPCRs and small molecules, and observed that the choice of the descriptors has a significant impact on the accuracy of the models. However, in all cases, we obtained significant improvements of the prediction accuracy with respect to the individual learning setting.

We will also show that the method can be applied to predict ligands for orphan GPCRs

In the following sections, we will present the mathematical formalism of the methods proposed by (109; 110) for *in silico* chemogenomics with SVM, and briefly introduced above (2). Then we will present the descriptors we propose to use for molecules and GPCRs within this framework.

3.5 Methods

3.5.1 Encoding the chemogenomic space within the SVM framework

We consider the problem of predicting interactions between GPCRs and small molecules. For this purpose, we assume that a list of target/small molecule pairs $\{(t_1, m_1), \ldots, (t_n, m_n)\}$ (t corresponds to protein targets and m to molecules), known to interact or not, is given. This can be viewed as the learning dataset. Such information is often available as a result of systematic screening campaigns in the pharmaceutical industry, or on dedicated databases. In our case, we used the GLIDA database as presented below.

Our goal is then to create a model to predict, for any new candidate pair (t, m), whether the small molecule m is likely to bind the GPCR t.

A general method to create the predictive model is to follow these four steps:

1. Choose n_{tar} descriptors to represent each GPCR target t in the biological space by a n_{tar}-dimensional vector $\Phi_{tar}(t) = (\Phi^1_{tar}(t), \ldots, \Phi^{n_{tar}}_{tar}(t))$;

2. In parallel, choose n_{mol} descriptors to represent each molecule m in the chemical space by a n_{mol}-dimensional vector $\Phi_{mol}(m) = (\Phi^1_{mol}(m), \ldots, \Phi^{n_{mol}}_{mol}(m))$;

3. Derive a vector representation of a candidate target/molecule complex $\Phi_{pair}(t, m)$ from the representations of the target $\Phi_{tar}(t)$ and of the molecule $\Phi_{mol}(m)$;

4. Use a statistical or machine learning method to train a classifier able to discriminate between binding and non-binding pairs, using the training set of binding and non-binding pairs $\{\Phi_{pair}(t_1, m_1), \ldots, \Phi_{pair}(t_n, m_n)\}$.

While the first two steps (selection of descriptors) may be specific to each particular chemogenomic problem (for example, in our case predicting ligands for GPCRs), the last two steps define the strategy we used for *in silico* chemogenomics, whatever the underlying biological question might be, as long as vector representation can be defined.

As mentioned above, (109; 113) proposed to concatenate the vectors $\Phi_{tar}(t)$ and $\Phi_{mol}(m)$ to obtain a $(n_{tar}+n_{mol})$-dimensional vector representation of the ligand-target complex $\Phi_{pair}(t, m)$, which corresponds to working in the joined space.

(110) followed a slightly different strategy for the third step, by forming descriptors for the pair (t, m) as tensor product of small molecule and target descriptors. More precisely, given a

molecule m described by a vector $\Phi_{mol}(m)$ and a GPCR t described by a vector $\Phi_{tar}(t)$, the pair (t, m) is represented by the tensor product:

$$\Phi_{pair}(t, m) = \Phi_{tar}(t) \otimes \Phi_{mol}(m)\,, \tag{3.2}$$

that is, a $(n_{tar} \times n_{mol})$-dimensional vector whose entries are products of the form $\Phi^i_{tar}(t) \times \Phi^j_{mol}(m)$, for $1 \leq i \leq n_{tar}$ and $1 \leq j \leq n_{mol}$.

In principle, both in the case of the tensor product space and of the joined space, a SVM can be trained to estimate a linear function $f(t, m)$ in the space of target/molecule pairs, that takes positive values for interacting pairs and negative values for non-interacting ones. This function is then used to predict whether a new (t, m) pair interact or not.

A potential issue with the tensor product approach, however, is that the size of the vector representation $n_{tar} \times n_{mol}$ for a pair may be prohibitively large for practical computation and manipulation.

For example, using a vector of molecular descriptors of size 1024 for molecules, and representing a protein by the vector of counts of all 20 amino-acids in its sequence results in more than 20k dimensions for the representation of a pair.

As pointed out by (110), a classical property of tensor products is that the inner product between two tensor products $\Phi_{pair}(t, m)$ and $\Phi_{pair}(t', m')$ is the product of the inner product between $\Phi_{tar}(t)$ and $\Phi_{tar}(t')$, on the one hand, and the inner product between $\Phi_{mol}(m)$ and $\Phi_{mol}(m')$, on the other hand. More formally, this property can be written as follows:

$$\left(\Phi_{tar}(t) \otimes \Phi_{mol}(m)\right)^\top . \left(\Phi_{tar}(t') \otimes \Phi_{mol}(m')\right) = \Phi_{tar}(t)^\top . \Phi_{tar}(t') \times \Phi_{mol}(m)^\top . \Phi_{mol}(m')\,, \tag{3.3}$$

where $u^\top v = u_1 v_1 + \ldots + u_d v_d$ denotes the inner product between two d-dimensional vectors u and v.

These kernels then allow to use SVM to train a linear classifier which permits to predict interacting and non-interacting pairs.

In the presented framework, the SVM does not need to compute the $n_{tar} \times n_{mol}$ products to calculate the vector that describes each pair (t, m) in the tensor product space. It only computes the respective inner products in the target and ligand spaces, before taking the product of both numbers, which corresponds to $(n_{tar} + n_{mol} + 1)$ products.

Finally any measure of similarity in the two spaces (proteins and ligands) which leads to a positive definite matrix can be used in the SVM. This offers a way to use SVM in a chemogenomic approach with a versatile and computationally efficient method.

This flexibility to manipulate molecule and target descriptors separately can moreover be combined with other tricks that sometimes allow to compute efficiently the inner products in the target and ligand spaces, respectively. Many such inner products, i.e. kernels, have been developed recently both in computational biology (75) and chemistry (114; 115; 116), and can be easily combined within the chemogenomics framework as follows: if two kernels for molecules and targets are given as:

$$K_{mol}(m, m') = \Phi_{mol}(m)^\top \Phi_{mol}(m'), K_{tar}(t, t') = \Phi_{tar}(t)^\top \Phi_{tar}(t'), \tag{3.4}$$

then we obtain the inner product between tensor products, i.e. the kernel between pairs, by:

$$K\left((t, m), (t', m')\right) = K_{tar}(t, t') \times K_{mol}(m, m'). \tag{3.5}$$

In summary, as soon as two vectors of descriptors and kernels K_{mol} and K_{tar} are chosen, we can solve the *in silico* chemogenomics problem with an SVM using the product kernel between pairs. Ideally, the particular descriptors and kernels used should ideally encode properties related to the ability of similar molecules to bind similar targets.

Note that when the chemogenomic space is described by the joined biological and chemical spaces, protein ligand pairs are encoded by a vector which is a concatenation of the two spaces. In this case, similarity measures can only be defined on combined descriptions of proteins and ligands. If various descriptions and similarity measures have been proposed in these two spaces, their effect on the prediction performance cannot be explored independently. On the contrary, the use of tensor product space allows to combine any pairs of kernel in the chemical and biological spaces.

In the next two subsections, we present different possible choices of descriptors – or kernels – for small molecules and GPCRs, respectively.

3.5.2 Descriptors and similarity measures for small molecules

The problem of explicitly representing and storing small molecules as finite-dimensional vectors has a long history in chemoinformatics, and a multitude of 1D, 2D or 3D molecular descriptors have been proposed (28), as presented in chapter 2.

In this study, we chose to select two existing kernels, that were developed in our laboratory, encoding respectively 2D and 3D structural information of the small molecules. (116) (55)

We used the freely and publicly available ChemCPP[1] software to compute the 2D and 3D pharmacophore kernel, as detailed below.

- *The 2D Tanimoto kernel.*

  Our first set of descriptors characterizes the 2D structure of the molecules. For a small molecule m, we define the vector $\Phi_{mol}(m)$ as the binary vector whose bits indicate the presence or absence of all linear graphs of length u or less, as subgraphs of the 2D structure of m. We chose $u = 8$ in our experiment, i.e. characterize the molecules by the occurrences of linear subgraphs of length 8 or less, a value that we previously observed to give good results in several virtual screening tasks (116).

  This vector description of the 2D structure of the molecules allowed to define a kernel based on Tanimoto coefficient (55).

  We used as a similarity measure corresponding to the Tanimoto kernel:

$$K_{ligand}(m, m') = \frac{\Phi_{lig}(m)^\top \Phi_{lig}(m')}{\Phi_{lig}(m)^\top \Phi_{lig}(m) + \Phi_{lig}(m')^\top \Phi_{lig}(m') - \Phi_{lig}(m)^\top \Phi_{lig}(m')}, \tag{3.6}$$

  which was proven to be a valid inner product, giving very competitive results on a variety of QSAR or toxicity prediction experiments (117).

- *3D pharmacophore kernel.*

  While 2D structures are known to be very competitive in ligand-based virtual screening for identification of molecules presenting some given chemical, physical or biological properties (118), we reasoned that the protein-ligand recognition process takes place in the 3D space.

  Thus, we decided to test descriptors implicitly encoding for the 3D conformation of the molecules. To perform this task, we used a 3D pharmacophore kernel proposed in our group (55).

[1] Available at `http://chemcpp.sourceforge.net`.

This kernel relies on a 3D pharmacophore representation of the molecule. A pharmacophore is usually defined as a three-dimensional arrangement of atoms - or groups of atoms - responsible for the biological activity of a drug molecule (49).

In the present approach, a molecule is defined by a set of triplets of atoms with their coordinates in 3D space. Comparing two molecules can be based on the comparison of triplets of atoms found in their structures (Figure 3.3).

A triplet of atoms constitutes a triangle whose vertices are atoms and edges are interatomic distances. Two triplets from two different molecules can be considered to be similar if the edges and vertices of their corresponding triangles are pairwise similar. In the case of vertices, the similarity might include comparison of the nature of the corresponding atoms (i.e. carbon atoms are similar to any other carbon atom), but it might also include other physicochemical characteristics such as the atom partial charges.

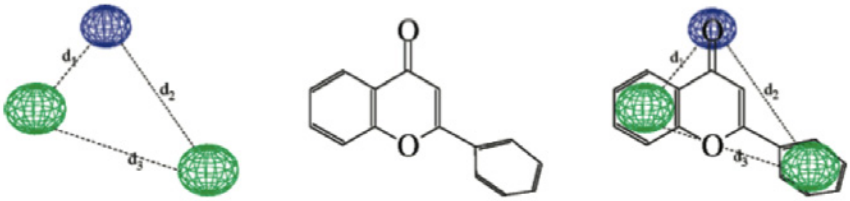


Figure 3.3: Left: A three-point pharmacophore made of one hydrogen-bond acceptor (top-most sphere) and two aromatic rings, with distances d_1, d_2, and d_3 between the features. Middle: The molecule of flavone. Right: Match between flavone and the pharmacophore.

The three-points pharmacophore kernel compares molecules according to their similarity in their 3D pharmacophore representations. Let us present this kernel.

The 3D structure of a molecule is represented as a set of points in $\mathbb{R}^3$. These points correspond to the 3D coordinates of the atoms of the molecule (for a given arbitrary basis of the 3D Euclidean space), and they are labeled with some information related to the atoms. More formally, we define a molecule m as:

$$m = \{(x_i, l_i) \in \mathbb{R}^3 \times \mathscr{L}\}_{i=1,\ldots,|m|} \tag{3.7}$$

where $|m|$ is the number of atoms that composes the molecule and $\mathscr{L}$ denotes the set of atom labels. The label should contain the relevant information to characterize a pharma-

cophore based on atoms. It might for instance be defined by the type of atoms (C, N, O, ...) or various physicochemical atomic properties (e.g., partial charge). The three-point pharmacophores correspond to triplets of distinct atoms of the molecules. The set of pharmacophores of the molecule m can therefore be formally defined as:

$$\mathscr{P}(m) = \{(p_1, p_2, p_3) \in m^3, p_1 \neq p_2, p1 \neq p_3, p_2 \neq p_3\} \tag{3.8}$$

where p_1, p_2, and p_3 represent atoms of the molecule.

More generally, the set of all possible pharmacophores is defined as $\mathscr{P} = (\mathbb{R}^3 \times \mathscr{L})^3$, to ensure the inclusion $\mathscr{P}(m) \subset \mathscr{P}$.

For any positive definite kernel for pharmacophores $K_{\mathscr{P}} : \mathscr{P} \times \mathscr{P} \mapsto \mathbb{R}$, we define a corresponding pharmacophore kernel for any pair of molecules m and m' by:

$$K(m, m') = \sum_{p \in \mathscr{P}(m)} \sum_{p' \in \mathscr{P}(m')} K_{\mathscr{P}}(p, p') \tag{3.9}$$

with the convention that $K(m, m) = 0$ if either $\mathscr{P}(m)$ or $\mathscr{P}(m)$ is empty. In other words, the problem of constructing a pharmacophore kernel for molecules therefore boils down to the simpler problem of defining a kernel between pharmacophores.

A chemically relevant measure of similarity between pharmacophores should obviously quantify at least two features: first, similar pharmacophores should be made of similar atoms, where the notion of similarity can for instance be based on atom types or properties such as partial charges, and second, the atoms should have similar relative positions in the 3D space (i.e. the triangles should have similar edges). It is therefore natural to study kernels for pharmacophores that decompose as follows:

$$K\mathscr{P}(p, p') = K_I(p, p') \times K_S(p, p') \tag{3.10}$$

where K_I is a kernel function that measures similarity of the triplets of atoms based on their labels (atom types or atom charges for example), and K_S is a kernel introduced to quantify their spatial similarity (i.e. similarity of the triangles formed by the triplet of atoms)

For any pair of pharmacophores $p = [(x_1, l_1), (x_2, l_2), (x_3, l_3)]$ and $p' = [(x'_1, l'_1), (x'_2, l'_2), (x'_3, l'_3)]$, this suggests the definition of kernels as follows:

$$K_I(p, p') = \prod_{I=1}^{3} K_{Feat}(l_i, l'_i) \tag{3.11}$$

$$K_S(p, p') = \prod_{I=1}^{3} K_{Dist}(||x_i - x_{i+1}||, ||x'_i - x'_{i+1}||) \tag{3.12}$$

where $||\cdot||$ denotes the Euclidean distance, the index $i+1$ is taken modulo 3, and K_Feat and K_Dist are kernel functions introduced to compare pairs of labels from $\mathscr{L}$ and pairs of distances, respectively.

The kernel we used for K_Dist is the Gaussian radial basis function (RBF) kernel defined by:

$$K_{Dist}^{RBF}(x, y) = exp\left(- \frac{||x - y||^2}{2\sigma^2} \right) \tag{3.13}$$

where $\sigma > 0$ is the bandwidth parameter that is optimized as part of the training of the classifier.

We used atom types for atom labels (C, N, O, ...). Therefore, we used the following Dirac kernel as a natural default choice to compare a pair of atom labels $l, l' \in \mathscr{L}$:

$$K_{Feat}^{Dirac}(l, l') = \left[\begin{array}{ll} 1 & \text{if} \quad l = l' \\ 0 & \text{otherwise} \end{array} \right. \tag{3.14}$$

This approach requires the choice of a 3D conformer for each molecule, in a context where there exists a large number of methods for exploring the conformation space, and where we lack significant data for bound ligands in GPCR structures. Therefore, we chose to build a 3D version of the ligand database in which molecules are represented in the conformation proposed by the Omega program (OpenEye Scientific Software), because it performs rapid systematic conformer search, and has been showed to present good performances for retrieving bioactive conformations (119). For each of the 2446 retained ligands, the conformer was generated using the standard Omega parameters, except for a $1\mathring{A}$ RMSD clustering of the conformers, instead of the 0.8 default value. Partial charges were calculated for all atoms using the Molcharge program (OpenEye

Scientific Software) with standard parameters. This ligand database was then used to calculate a 3D pharmacophore kernel for molecules (55).

3.5.3 Descriptors and similarity measures for GPCRs

SVM and kernel methods are also widely used in bioinformatics (75), and a variety of approaches have been proposed to design kernels between proteins, ranging from kernels based on the amino-acid sequence of a protein (120; 121; 122; 123; 124; 125; 126) to kernels based on the 3D structures of proteins (127; 128; 129). These kernels have been used in conjunction with SVM for various tasks related to structural or functional classification of proteins. While any of these kernels can theoretically be used as a GPCR kernel in (3.5), we investigated in this work a few kernels described below, aimed at illustrating the flexibility of our framework and test various hypothesis.

- **A baseline method: the Dirac kernel.** The Dirac kernel between two targets t, t' is defined as:

$$K_{Dirac}(t,t') = \begin{cases} 1 & \text{if } t = t' \,, \\ 0 & \text{otherwise.} \end{cases} \tag{3.15}$$

  This basic kernel simply represents different targets as orthonormal vectors. In other words, using Dirac kernel for proteins amounts to performing classical learning independently for each target. This kernel will be used in our work as a reference baseline approach to which chemogenomic methods (that, on the contrary, aim at sharing information between targets) will be compared.

  From equation (3.5), we see that orthogonality between two proteins t and t' implies orthogonality between all pairs (l,t) and (l',t') for any two small molecules l and l'. This means that a linear classifier for pairs (l,t) will be trained without sharing any information of known ligands between different targets.

- **The multitask kernel : uniform sharing of ligand information.**

  The multitask kernel between two targets t, t' is defined as:

$$K_{multitask}(t,t') = 1 + K_{Dirac}(t,t') \,.$$

  or :

$$K_{Multitask}(t,t') = \begin{cases} 2 & \text{if } t = t' \,, \\ 1 & \text{otherwise.} \end{cases} \tag{3.16}$$

This kernel was originally proposed in the context of multitask learning (130). Multitask learning is an approach to machine learning, that learns a problem together with other related problems at the same time, using a shared representation. This often leads to a better model for the main task, because it allows the learner to use the commonality among the tasks.

Unlike the dirac kernel, this kernel is never equal to zero even when t and t' are different. It takes the constant value of 1 when t and t' are different, which allows sharing of information uniformly between proteins, regardless of their similarity. In other words, this means that ligand information known for any GPCR will be taken into account uniformly to predict ligands for a given GPCR. This kernel takes the value of 2 for identical proteins. This means that ligand information known for the protein under study will have a double weight for prediction of new ligands for this protein, with respect to ligand information arising from other proteins.

The next two kernels allow to share information between proteins based on otherwise known biological information. With this kernel, the amount of information shared between proteins depends on their similarity according to some relevant biological characteristics. This is not the case of the multitask kernel which propagates information between proteins uniformly.

- **The *hierarchy* kernel.** Proteins have often been classified according to their sequence, or to some physical or biological property (for example, the type of reaction they catalyze, in the case of enzyme), defining families and superfamilies or proteins. For example, the SCOP, or the KEGG databases classify proteins according to their overall fold or enzymatic properties, respectively. These classifications can be used to define hierarchies that, in turn, can be used to define kernels. Let us explain this principle in the case of GPCRs.

  In the GLIDA database, GPCRs are grouped into 7 classes based on sequence homology and functional similarity: the rhodopsin family (class A), the secretin family (class B), the metabotropic family (class C) and some smaller classes containing other GPCRs (see Figure 3.4). The GLIDA database further subdivides each class of targets by type of ligands, for example amine or peptide receptors or more specific families of ligands. This defined a natural hierarchy in the GLIDA database that can be used to compare GPCRs.

- Class A Rhodopsin like
  - Amine
  - Peptide
  - Hormone protein
  - (Rhod)opsin
  - Olfactory
  - Prostanoid
  - Nucleotide-like
  - Cannabinoid
  - Platelet activating factor
  - Gonadotropin-releasing hormone
  - Thyrotropin-releasing hormone and Secretagogue
  - Melatonin
  - Lysosphingolipid and LPA
  - Leukotriene B4 receptor
  - Class A Orphan/other
- Class B Secretin like
  - Calcitonin
  - Corticotropin releasing factor
  - Gastric inhibitory peptide
  - Glucagon
  - Growth hormone-releasing hormone
  - Parathyroid hormone
  - PACAP
  - Secretin
  - Vasoactive intestinal polypeptide
  - Diuretic hormone
  - EMR1
  - Latrophilin
  - Brain-specific angiogenesis inhibitor (BAI)
  - Methuselah-like proteins (MTH)
  - Cadherin EGF LAG (CELSR)
- Class C Metabotropic glutamate / pheromone
  - Metabotropic glutamate
  - Calcium-sensing like
  - Putative pheromone receptors
  - GABA-B
  - Orphan GPRC5
  - Orphan GPCR6
  - Bride of sevenless proteins (BOSS)
  - Taste receptors (T1R)
- Class D Fungal pheromone
  - Fungal pheromone A-Factor like (STE2,STE3)
  - Fungal pheromone B like (BAR,BBR,RCB,PRA)
  - Fungal pheromone M- and P-Factor
  - Fungal pheromone other
- Class E cAMP receptors (Dictyostelium)
- Frizzled/Smoothened family
- Class D Fungal pheromone
  - frizzled
  - Smoothened

Figure 3.4: Hierarchy used in the hierarchy kernel

We used the underlying idea that, the more two proteins are close in the hierarchy, the more common ancestors they have, leading to high similarity that can be taken into account in a hierarchy kernel.

Suppose that the hierarchy contains n nodes (taking leaves into account). A protein t can be encoded as a binary vector $\Phi_h(t)$ of length n. Element i of the vector is equal to 1 if node i is present in the hierarchy of t, and 0 otherwise. This allow to define a scalar product between proteins.

The similarity between two proteins t and t' can be measured by the scalar product between their feature vectors $\Phi_h(t)$ and $\Phi_h(t')$.

Quite intuitively, the hierarchy kernel between two GPCRs was therefore defined as the number of common ancestors in the corresponding hierarchy :

$$K_{hierarchy}(t, t') = \langle \Phi_h(t), \Phi_h(t') \rangle,$$

Note that this kernel is never equal to zero because all proteins shared at least the root of the tree which corresponds to the common protein ancestor in this family. The underlying idea in the hierarchy kernel is that, the more proteins are close in the hierarchy (i.e. the more they share common ancestor nodes), the more they are considered to be similar, the more they are expected to share similar ligands.

- **The binding pocket kernel.** Here, the underlying idea is that the protein-ligand recognition process occurs in 3D space in a pocket involving a limited number of residues. Therefore, proteins displaying similar binding pockets are expected to bind similar ligands. The questions are how to extract and encode the pocket, and how similarity is measured between pockets. We tried to describe the GPCR space using a representation of this pocket. An additional difficulty resides in the fact that although the GPCR sequences are known, the residues forming this pocket are *a priori* unknown. However, mutagenesis data showed that the transmembrane binding site is situated in a similar region for all GPCRs (131), and this information was confirmed by the two available X-ray structures available at the time we started this project. In order to identify residues potentially involved in the binding pocket of GPCRs of unknown structure studied in this work, we proceeded in several steps, somewhat similarly to (132).

(a) The two known structures (at the time of this study), PDB entries 1U19 and 2RH1 (133; 134), were superimposed using the STAMP algorithm (135). Although retinal is an inverse agonist and forms a covalent bond with Rhodopsin, while carazolol is an agonist and binds non-covalently, root mean square deviation between these two complexed structures is only of $1.6\mathring{A}$ in the transmembrane helices (136). In the superimposed structures, the retinal and 3-(isopropylamino)propan-2-ol ligands are localized in the same region of the transmembrane space, which is in agreement with global conservation of binding pockets, as shown on Figure 3.5.

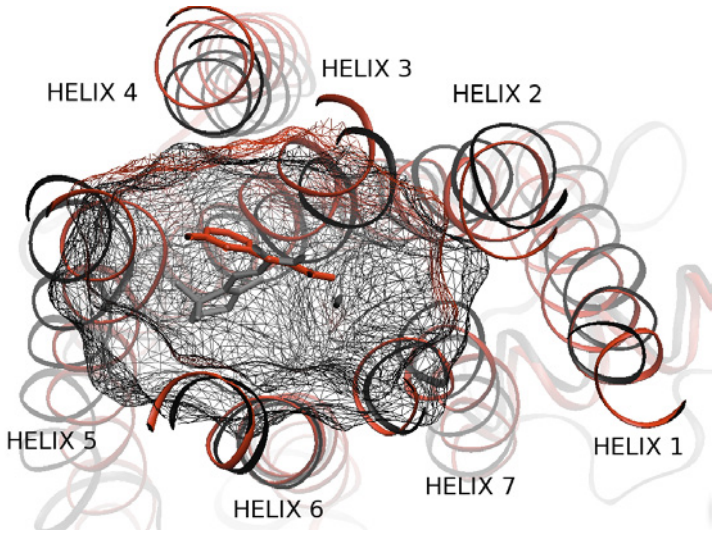


Figure 3.5: Binding pocket. Representation of the binding pocket of β2-adrenergic receptor (in red) and bovine Rhodopsin (in black) viewed from the extracellular surface. On the center of the pocket, 3-(isopropylamino)propan-2-ol and cis-retinal have been represented to show the size and the position of the pocket around each ligand. Figure drawn with VMD (137)

(b) The structural alignment of bovine rhodopsin and of human β_2-adrenergic receptor was used to generate a sequence alignment of these two proteins.

(c) For both structures, in order to identify residues potentially involved in stabilizing interactions with the ligand (i.e. residues defining the pocket), we selected residues that presented at least one atom situated at less than $6\mathring{A}$ from at least one atom of the ligand. Figure 3.6 shows that these two pockets clearly overlap, as expected.

(d) Residues of the two pockets (as defined in c) were labeled in the structural sequence alignment performed in (b). These residues were found to form small sequence clusters

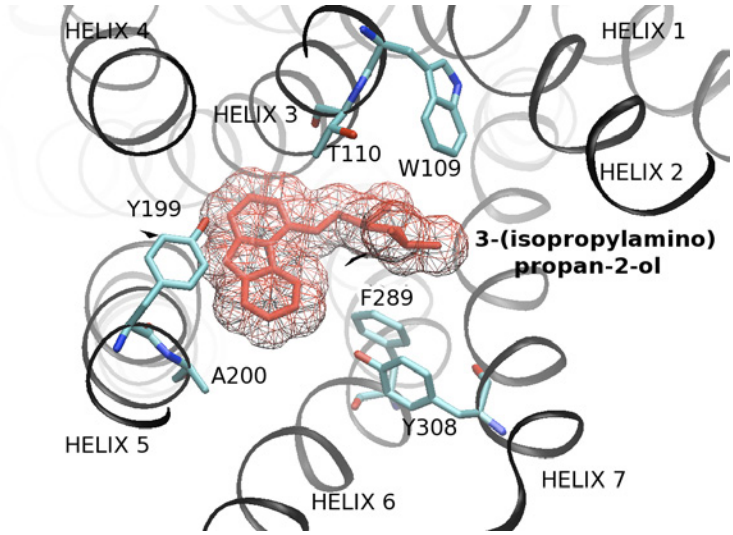


Figure 3.6: 3-(isopropylamino)propan-2-ol and the protein environment of β2-adrenergic receptor as viewed from the extracellular surface. 3-(isopropylamino)propan-2-ol and the protein environment of β1-adrenergic receptor as viewed from the extracellular surface. Amino acid side chains are represented for 6 of the 31 residues (in cyan, blue and red) of the binding pocket motif. Transmembrane helix and 3-(isopropylamino)propan-2-ol are colored in black and red respectively. Figure drawn with VMD (137)

that were in correspondence in this alignment. These clusters were situated mainly in the apical region of transmembrane segments and included a few extracellular residues. This is consistent with the fact that it has been previously demonstrated that extracellular loops can play a role in ligand binding together with transmembrane regions (138).

(e) All studied GPCR sequences, including bovine rhodopsin and human β_2-adrenergic receptor were aligned using CLUSTALW (139) with Blossum matrices (140). Sequences which could not be correctly aligned (e.g. with important gaps in the transmembrane regions) were discarded in order to only keep comparable sequences. We have then checked that conserved residues of the transmembrane helices were correctly aligned, according to (141), and local misalignments were corrected. In addition, the structural alignment of bovine rhodopsin and human β_2-adrenergic receptor, and known conserved positions were used to locally correct misalignments. For each protein, residues in correspondence in this alignment with a residue of the binding pocket (as defined above) of either bovine rhodopsin or human β_2-adrenergic receptor were retained. This lead to a different number of residues per protein, because of sequence variability. For example, in extracellular regions, some residues from bovine rhodopsin or human β_2-adrenergic receptor had a corresponding residue in some sequences but not in others. In order to provide a homogeneous description of the binding pocket for all GPCRs, in the list of residues initially retained for each protein, only residues situated at positions where no gaps were found in any of the GPCRs were kept.

(f) Each protein was then encoded by a vector whose elements corresponded to the aminoacids potentially involved in the binding of the ligand, as defined in (e). This description of each protein by a vector of size 31 filled with amino acid residues, as illustrated in Table 3.1, implicitly codes for a 3D information on the receptor pocket.

These amino acid vectors were then used to build a kernel that allows comparison of binding pockets. In this representation, the inner product between two proteins (i.e. two binding pocket motifs) is simply the number of residues they have in common at the same positions:

$$K_{pb}(x, x') = \sum_{i=1}^{l} \delta(x[i], x'[i]),$$

positions on ß2-adrenergic receptor	82	109	110	113	114	115	116	117	118	121	175	183	195	199	200	203
ß2-adrenergic receptor	M	W	T	D	V	L	C	V	T	I	R	N	T	Y	A	S
5-hydroxytryptamine 5A receptor	V	W	I	D	V	L	C	C	T	I	I	E	S	Y	A	S
Adenosine A2b receptor	V	L	A	V	L	V	L	T	Q	I	I	K	K	M	V	N
Gamma-aminobutyric acid type B receptor	E	D	E	E	A	V	E	G	H	T	L	G	S	F	D	G
Relaxin 3 receptor 2	L	V	L	T	V	L	N	V	Y	I	V	G	L	Y	Q	R

positions on ß2-adrenergic receptor	204	207	208	212	282	286	289	290	293	308	311	312	313	315	316
ß2-adrenergic receptor	S	S	F	L	F	W	F	F	N	Y	L	N	W	G	Y
5-hydroxytryptamine 5A receptor	T	A	F	L	F	W	F	F	E	K	F	L	W	G	Y
Adenosine A2b receptor	F	C	V	L	F	W	V	H	N	M	A	I	L	S	H
Gamma-aminobutyric acid type B receptor	S	A	W	E	F	L	Y	H	R	L	T	V	G	L	V
Relaxin 3 receptor 2	V	A	F	L	F	W	N	H	T	F	T	T	C	A	H

Table 3.1: Aligned receptor pocket residues. Residues of 5-hydroxytryptamine 5A receptor, Adenosine A2b receptor, Gamma-aminobutyric acid type B receptor and Relaxin 3 receptor 2 (shown as examples) aligned with β2-adrenergic receptor binding site amino acids. The binding pocket motif of β2-adrenergic receptor has been used as reference to determine the positions in the protein sequence of the residues involved in the formation of the binding site of the 79 other GPCRs.

where l is the length of the binding pocket motifs (31 in our case), $x[i]$ is the i-th residue in x and $\delta(x[i], x'[i])$ is 1 if $x[i] = x'[i]$, 0 otherwise. This kernel defines what we call the baseline binding pocket kernel. We also tested a polynomial kernel of degree p over the baseline kernel:

$$K_{ppb}(x, x') = \left(K_{pb}(x, x') + 1\right)^p .$$

We only used a degree of $p = 2$. However, it could be interesting in future studies to test wether other values of this parameter could further improve the performances.

To summarize, in the binding pocket kernel, proteins are encoded by subsequences of aminoacids representing the ligand binding pocket and are compared according to their subsequences similarity.

3.5.4 Data description

3.5.4.1 Filtering the GLIDA database

We used the GLIDA GPCR-ligand database (112) which includes 22964 known ligands for 3738 GPCRs from human, rat and mouse.

The ligand database contains highly diverse molecules, from ions and very small molecules up to peptides, and a significant number of duplicates. These redundancies were eliminated. Elimination of duplicates present in the GLIDA database was important here, because it could have led to over-optimistic evaluation in the cross-validation procedure described below. The remaining molecules were filtered in order to satisfy two constraints.

Indeed, since the long term goal is to identify drug candidates targeting GPCRs, it was important to retain only drug-like compounds, i.e. molecules having the adequate physico-chemical characteristics to be potential drugs candidates satisfying ADME criteria (142).

Therefore, to only keep drug-like compounds, we filtered the GLIDA database using the Filter program (OpenEye Scientific Software) with standard parameters, which removes molecules according to calculated properties such as molecular weight, hydrogen bond donor and acceptor count, number of rotatable bonds, ring size and number etc... as discussed in (30; 143; 144; 145).

For example, only molecules of molecular weights ranging from 150 Da to 450 Da were kept (the classically accepted range for drugs), since the aim was to evaluate if statistical learning was possible on drug-like compounds. Another example was the elimination of molecules

with more than 10 rotatable bonds (although most of them being already filtered out on the molecular weight criterion). Indeed, they correspond to very flexible molecules that are not suitable for the use of 3D descriptors. Overall these filters retained 2446 molecules, available under a 2D description file in the GLIDA data bank, and giving 4051 interactions with the human GPCRs. The number of molecules retained is only a small fraction of the GLIDA database, but it corresponds to all drug-like compounds of this database.

We also filtered proteins from the GLIDA database: we loaded the sequences of all GPCRs that are able to bind any of these ligands, which resulted in 80 sequences, all corresponding to human GPCRs. The retained GPCRs were significantly diverse in sequence, most of them sharing 15% to 50% pairwise sequence similarities. Furthermore, they belonged to various families, according to the GLIDA classification. They are found in several subfamilies of class A (rhodopsin-like receptors), classes B (secretin family) and C (metabotropic family). In the GLIDA database, GPCRs are classified in hierarchy (as mentioned above) which was used in the hierarchy kernel.

3.5.4.2 Buliding of the learning dataset

Statistical learning methods required to build a learning dataset containing positive and negative example of the considered property. In our case, this property correspond to the ability of a given protein t and a given ligand $l$to form an interaction.

The GLIDA database provided positive examples corresponding to known interactions involving any GPCR and any of the filtered molecules. For each of these positive interactions, we generated a negative interaction involving the same receptor and one of the ligands of the database not known to interact with this receptor. We are aware that this may have generated a few false negative points in our benchmark. One possible improvement would be to use experimentally tested negative interactions.

However, the mean similarity between the different ligands in the database using the 2D Tanimoto kernel (see section 3.5.2), which is later used in our method, is quite low (0.13). Besides, only 6.7% of the ligands have a mean similarity of more than 0.2 to the other ligands. This suggests that even if false negatives have to be expected, the method used here to generate negative interaction is reasonable.

Let us now present how the performance of the studied protein and ligand kernels were evaluated in a chemogenomic study.

3.6 Results

We ran two different sets of experiments in order to illustrate two important points.

In a first set of experiments, for each GPCR, we divided the known interactions for this GPCR (i.e. the line of the interaction matrix corresponding to this GPCR) into 5 folds (Figure 3.7A). The classifier was trained with four folds of this GPCR and the whole data from the other GPCRs (i.e. all other lines of the interaction matrix).

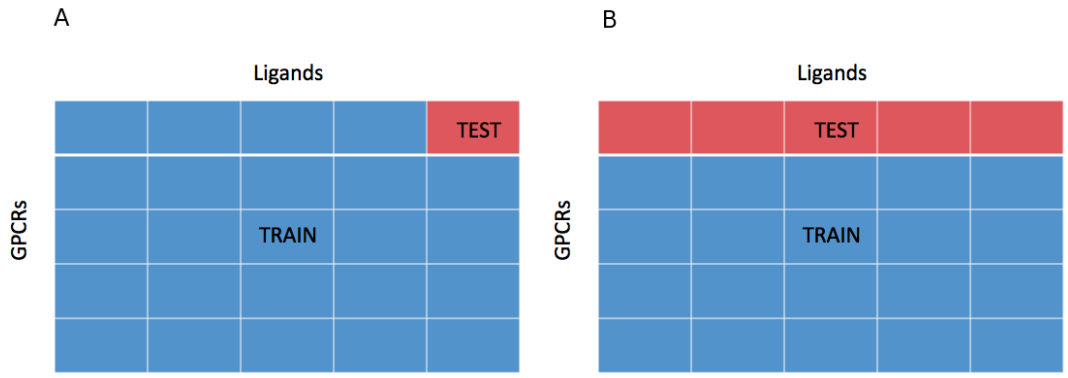


Figure 3.7: representation of the data as an interaction matrix between GPCRs and ligands. (A) corresponds to the classifier trained with four folds of this GPCR and the whole data from the other GPCRs. (B) corresponds to the classifier trained on the whole data from the other GPCRs, and tested on the data of the considered GPCR

.

The prediction accuracy for the GPCR under study was then tested on the remaining fold. The goal of these first experiments was to evaluate if using data from other GPCRs improved the prediction accuracy for a given GPCR, as compared to the classical approach where only the data for the studied GPCR are taken into account.

In a second set of experiments, for each GPCR we ignored all ligand data available for this particular GPCR. We trained a classifier on the whole data from the other GPCRs, and tested on the data of the considered GPCR (Figure 3.7B). The goal was to assess how efficient our chemogenomic approach would be to predict the ligands of orphan GPCRs.

Without going into mathematical details, in both experiments, the C parameter appearing in the SVM method (cf section 2.7.4) was selected by internal cross validation on the training set among values equal to $2^i, i \in \{-8, -7, \ldots, 5, 6\}$ as recommended by (146).

Table 3.2 shows the results of the first experiments with all the ligand and GPCR kernel combinations.

$K_{tar} \backslash K_{lig}$	2D Tanimoto	3D pharmacophore
Dirac	86.2 ± 1.9	84.4 ± 2.0
multitask	88.8 ± 1.9	85.0 ± 2.3
hierarchy	93.1 ± 1.3	88.5 ± 2.0
binding pocket	90.3 ± 1.9	87.1 ± 2.3
poly binding pocket	92.1 ± 1.5	87.4 ± 2.2

Table 3.2: Prediction accuracy for the first experiment with various ligand and target kernels

3.6.1 Performance of protein kernels

For both ligand kernels, we observe an improvement between the individual approach (Dirac kernel, 86.2%) and the baseline multitask approach (multitask kernel, 88.8%). The latter kernel is merely modeling the fact that each GPCR is uniformly similar to all other GPCRs, and twice more similar to itself. It does not use any prior information on the GPCRs, and yet, using it improves the global performance with respect to individual learning. This result illustrates the interest of chemogenomic approaches: sharing information between proteins, even in a very naive manner, improves prediction performance.

Other kernels allowing to share information between GPCR using more biologically relevant information (hierarchy and binding pocket kernels) further improved the prediction accuracy. In particular, the hierarchy kernel add more than 4.5% of precision with respect to naive multitask approach. All the other informative GPCR kernels also improve the performance. The polynomial binding pocket kernel is almost as efficient as the hierarchy kernel, which is an interesting result.

Indeed, one could fear that using the hierarchy kernel, for the construction of which some knowledge about the ligands may have been used, could have introduced bias in the results. Such bias is certainly absent in the binding pocket kernel. The fact that almost the same performance can be reached with kernels based on the mere sequence of GPCRs' pockets is therefore an important result. Figure 3.8 shows three of the GPCR kernels. The baseline multitask is shown as a comparison. Interestingly, many of the subgroups defined in the hierarchy can be

found in the binding pocket kernel, that is, they are retrieved from the simple information of the binding pocket sequence.

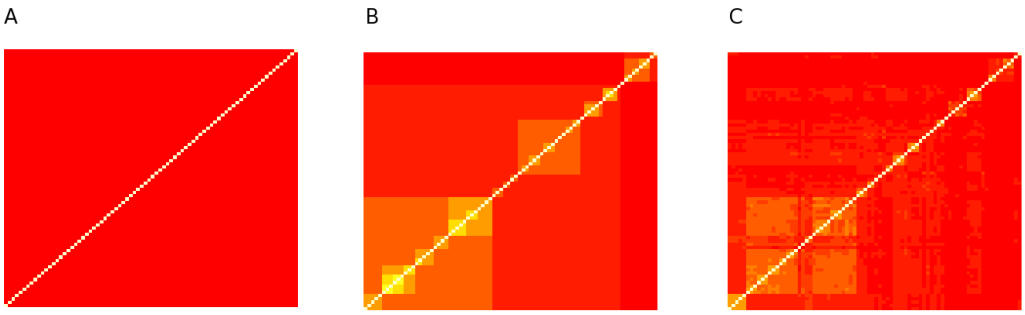


Figure 3.8: GPCR kernel Gram matrices. The above square matrices correspond to the pairwise similarity measures between the GPCR of the studied dataset based on 3 considered kernels: multitask (A), hierarchy (B) and binding pocket (C) kernels.

3.6.2 Performance of ligand kernels

For all protein kernels, Table 3.2 show that the 3D kernel for the ligands, did not perform as well as the 2D kernel. This may be explained by the fact the pharmacophore kernel is not suited to this problem. However, another point should be discussed. In the case of the 3D approach, molecular encoding and similarity measure is highly dependent on the conformers chosen in 3D space. Choosing the relevant conformer for a ligand, i.e; its active conformation in which it can bind to its target, is not a trivial task. The way conformers were chosen probably had an impact on the performance of the 3D approach. This point is discussed below, in section 3.7.

3.6.3 Impact of the number of training point on the prediction performance

Figure 3.9 illustrates how the improvement brought by the chemogenomics approach varies with the number of available training points (i.e the number of known GPCR - ligand couples). As one could have expected, the strongest improvement is observed for the GPCRs with few (less than 20) training points (i.e. less than 10 known ligands since for each known ligand

an artificial non-ligand was generated). When more training points become available, the improvement is less important, and sharing the information across the GPCRs can even degrade the performances.

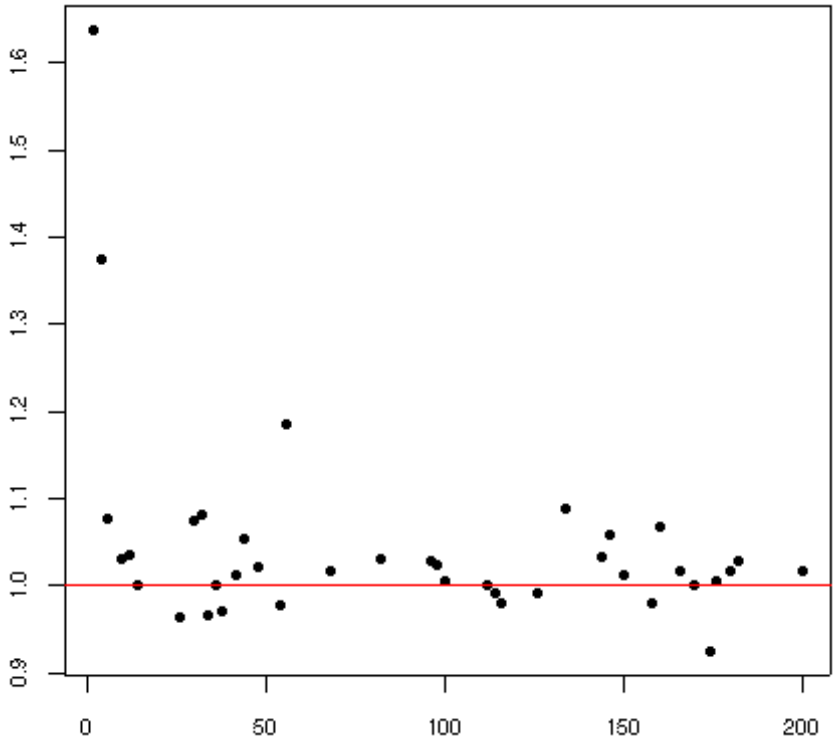


Figure 3.9: Improvement of the chemogenomics approach. Improvement (as a performance ratio) of the hierarchy GPCR kernel against the Dirac GPCR kernel as a function of the number of training samples available. Restricted to $[2 - 200]$ samples for the sake of readability.

This is an important result, first because, as showed on Figure 3.10, many GPCRs have few known ligands (in particular, 11 of them have only two training points), and second because it shows that when enough training points are available, individual learning will probably perform as well as or better than our chemogenomics approach.

3.6.4 Prediction performance of the chemogenomic approach on orphan GPCR

Our second experiment (described above) intends to assess how our chemogenomic approach can perform when predicting ligands for orphan GPCRs, i.e. with no training data available

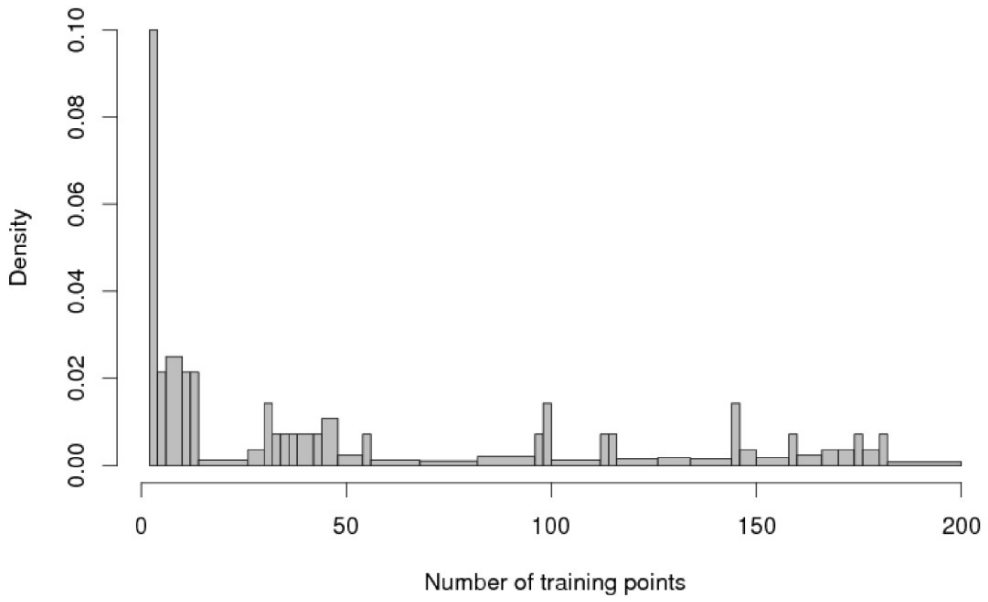


Figure 3.10: Distribution of the number of GPCR proteins having a given number of training points (number of ligands). Restricted to $[2 - 200]$ samples for the sake of readability.

for the GPCR of interest. Table 3.3 shows that the dirac kernel, equivalent to individual learning performs random prediction as one could expect for orphans receptors. Naive multitask approach provides modest improvement of the performance. Note that kernels baring more biological information such as hierarchical and binding pocket kernels achieve 77.4% and 78.1% of precision respectively, that is almost 30% better than the random approach one would get when no data are available. Here again, the fact that the binding pocket kernel which only uses the sequence of the receptor pocket performs as well as the hierarchy-based kernel is encouraging. It suggests that given a orphan receptor for which only its sequence is known, it is possible to make reasonable ligand predictions.

$K_{tar} \backslash K_{lig}$	2D Tanimoto	3D pharmacophore
Dirac	50.0 ± 0.0	50.0 ± 0.0
multitask	56.8 ± 2.5	58.2 ± 2.2
hierarchy	77.4 ± 2.4	76.2 ± 2.2
binding pocket	78.1 ± 2.3	76.6 ± 2.2
poly binding pocket	76.4 ± 2.4	74.9 ± 2.3

Table 3.3: Prediction accuracy for the second experiment with various ligand and target kernels

3.7 Discussion

Our results demonstrate that chemogenomic approaches outperform individual approach, in particular in cases where very limited or no ligand information is available, as shown in Table 3.3 and Figure 3.9. In the case of well studied GPCRs, more classical ligand-based methods (QSAR) may be better suited to predict new strong binders from a large number of known ligands, as shown in Figure 3.9. Consistent with this observation, Tables 3.4 and 3.5 show that in the two types of experiments, the improvement is observed for all subfamilies of GPCRs retained in this study. This is an interesting result since most of published virtual screening studies on GPCRs were applied to class A GPCRs.

Since our chemogenomic approach is a ligand-based approach, it would probably be interesting to use it in combination with docking. Indeed, although prior known ligands can help tuning docking procedures to the receptor under study, it can in principle be used with little or no ligand information. When more experimental 3D structures become available for GPCRs

Family $\backslash K_{tar}$	Dirac	multitask	hierarchy	BP	PBP
Rhodopsin peptide receptors (18)	73.7	80.0	85.8	83.8	83.7
Rhodopsin amine receptors (35)	91.1	92.1	94.0	93.9	94.1
Rhodospin other receptors (17)	83.6	88.0	95.7	95.9	95.9
Metabotropic glutamate family (9)	73.1	93.5	98.9	83.3	93.3
Secretin family (1)	50.0	100.0	100.0	50.0	100.0

Table 3.4: Prediction accuracy by GPCR family for the first experiment. Mean prediction accuracy for each GPCR family for the first experiment with the 2D Tanimoto ligand kernel and various target kernels. The numbers in bracket are the numbers of receptors considered in the experiment for each family. BP is the binding pocket kernel and PBP the poly binding pocket kernel

Family $\backslash K_{tar}$	Dirac	multitask	hierarchy	BP	PBP
Rhodopsin peptide receptors (18)	50.0	50.6	66.7	74.0	65.3
Rhodopsin amine receptors (35)	50.0	56.0	73.7	74.0	73.1
Rhodospin other receptors (17)	50.0	50.2	86.5	87.6	85.5
Metabotropic glutamate family (9)	50.0	79.7	93.9	87.2	91.3
Secretin family (1)	50.0	100.0	100.0	50.0	100.0

Table 3.5: Prediction accuracy by GPCR family for the second experiment. Mean prediction accuracy for each GPCR family for the second experiment with the 2D Tanimoto ligand kernel and various target kernels. The numbers in bracket are the numbers of receptors considered in the experiment for each family. BP is the binding pocket kernel and PBP the poly binding pocket kernel.

in the future, this will help building reliable models for a wider range of GPCRs that would be suitable for docking studies. Joint use of ligand-based chemogenomic and docking would certainly improve predictions.

We chose to use a binary descriptor for the receptor-ligand interaction, while QSAR or docking methods usually try to rank molecules according to their predicted affinity for the receptor. However, affinity prediction is still a subject of research at the level of a single receptor, at least when using methods whose calculation times are compatible with the screening of large molecular databanks. In this context, we feel that in chemogenomic approaches, where information is shared between different proteins, such quantitative prediction is even more challenging. This led us to retain the binary binding and non-binding descriptors, although it would formally have been straightforward to use a regression algorithm instead of a classification one to make quantitative predictions.

It is not always easy to compare the performances of a new method to other existing methods, and particularly in the case of GPCRs. Indeed, at least to our knowledge, there is up to now no public data from previous screening studies available as a benchmark to compare different screening methods on the same data. This urged us to give public access to the ligand and receptor databases used in this study, to the detailed experimental protocol of the study, and to the predictions made by our chemogenomic approach for each GPCR (see additional files sections 4.3 and 4.4).

This provides a benchmark which we hope will contribute to a fair evaluation of different methods and trigger new developments. This benchmark could be used to compare predictions made by other methods.

Our approach boils down to the application of well-known machine learning methods in the constructed chemogenomic space. We used a systematic way to build such a space by combining a given representation of the ligands with a given representation of the GPCRs into a binding-prediction-oriented GPCR-ligand couple representation. This allows to use any ligand or GPCR descriptor or kernel existing in the cheminformatics or bioinformatics literature, or new ones containing other prior information. Our experiments showed that the choice of the descriptors was crucial for the prediction, and more sophisticated features for either the ligands or the GPCRs could probably further improve the performances. Among these features, improvements in the 3D ligand descriptors could probably be obtained. Indeed, 3D pharmacophore kernels did not always reach the performance of 2D kernels for the ligands. This is

apparently in contradiction with the idea that protein-ligand interaction is a process occurring in the 3D space, and with previous work in our group (55).

Different explanations can be proposed. First, it is possible that the bioactive conformation was not correctly predicted for all molecules used in this study. For the two ligands for which it was known, i.e. retinal and 3-(isopropylamino)propan-2-ol from PDB entries 1U19 and 2RH1 respectively, we found that the predicted conformation, using the same method as for all other molecules, was very close to the experimental conformation, with RMSD values of less than $1\mathring{A}$. However, in absence of any other information on other bound ligand conformations, it is not possible to rule out the possibility that for other molecules, the prediction was not correct.

Although more complete conformational space exploration for all ligands was out of the scope of our experiments and would be a study by itself, work in this direction could improve the method. In particular, since 2D ligand-based methods are not easily suitable to make predictions outside of the molecular scaffolds for which information is known, ligand-based methods using 3D description are of particular interest, because they are expected to allow better predictions on molecules presenting diverse molecular patterns. However, conformer generation and selection is a major drawback of using 3D descriptors, especially in the case of large ligands with many free torsion angles. Synergy between our method and docking would provide a means for the choice of a conformer. The principle could be to build homology models for the GPCRs, dock the molecular database in the modeled binding pockets, and derive a 3D database using, for each molecule, the conformer associated to the best docking solution.

Various evidence suggest that, within a common global architecture, a generic binding pocket mainly involving transmembrane regions hosts agonists, antagonists and allosteric modulators. In order to identify this pocket automatically, other studies report the use of sequence alignment and the prediction of transmembrane helices. (131) detected hypervariable positions in transmembrane helices for identification of residues forming the binding pocket, although some positions were more conserved. Indeed, conserved residues are probably important for structural stabilization of the pocket, while variable positions are involved in ligand binding, in order to accommodate the wide spectrum of molecules that are GPCR ligands. Analyzing the positions of variable positions, these authors proposed potential binding pockets for GPCRs, and found that the corresponding residues were frequently in the GRAP mutant database for GPCRs (147). Interestingly, they pointed that residues at hypervariable positions were found within a distance of $6\mathring{A}$ from retinal in the rhodopsin X-Ray structure, which is also a classical distance cutoff above which it is admitted that protein-ligand interactions become negligible.

The simple and automatic method used in the present work for extracting GPCRs potential binding pockets, and are in good agreement with this study. It is also important to note that GPCRs are known to exist in dynamic equilibrium between inactive- and several active-state conformations (148), and different ligands sometimes trigger distinct conformational changes and stabilize different receptor conformations (149).

Taking into account receptor plasticity constitutes in itself a research domain in docking. Its use is of particular interest for screening GPCR homology models since residue positions are not exactly known. Therefore flexible docking procedures have been proposed and applied on GPCR proteins (98; 150). Moreover, a modeling method has been proposed to get insights on transmembrane bundle plasticity (151).

In our case, receptor flexibility might influence the definition of the binding pocket, since it initially relies on the identification of residues in the two reference structures (1U19 and 2RH1) that present at least one atom situated at less than $6\mathring{A}$ of the ligand. Therefore, we made the implicit hypothesis that receptor conformational changes upon ligand binding does not drastically affect this list of residues. When more structures become available in this family of proteins, a better appreciation of such conformational rearrangements will be possible, which could be taken into account in the binding pocket definition and could help to improve the method.

(147) found that hierarchical tree representations of GPCR subfamilies calculated with full-length GPCR sequences or with only binding pocket residues were similar, and that locally, the latter was in better agreement with functional data, although their binding pocket included only 31 residues. This result is also in good agreement with our finding that the hierarchy kernel based (among other information) on full length sequence (from GLIDA) and the kernel based on the binding pocket provided very similar performances. As mentioned in the Results section, it is however important to note that the kernels based on the binding pocket were built without any ligand information that could lead to some bias and artificially better performance.

3.8 Conclusion

We showed how sharing information across the GPCRs by considering a chemogenomic space of the GPCR-ligand interaction pairs could improve the prediction performances, with respect to the single receptor approach. In addition, we showed that using such a representation, it was possible to make reasonable predictions even when all known ligands were ignored for a given

GPCR, that is, to predict ligands for orphan GPCRs. Our results demonstrate that chemogenomic approaches is particularly suited to cases where very limited or no ligand information is available, as shown in Table 3.3.

This chemogenomics approach is related to ligand-based approaches. However, sharing information among different GPCRs allows, in this case, to perform prediction on orphan GPCRs, which is also possible using target-based methods. Nevertheless, the latter are limited by the number of known receptor structures and the difficulty to apply such methods on homology models.

Interesting developments of this method could include application to other important drug target families, like enzymes or ion channels (23), for which most of the descriptors used for the GPCRs in this work could directly be transposed, and other, more specific ones could be designed.

3.9 Additional files

GPCR $\backslash K_{tar}$	Dirac	multitask	hierarchy	BP	PBP
Rhodopsin peptide receptors					
AG2R(5)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
CCKAR(6)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
CML2(1)	50.0 ± 0.0	50.0 ± 35.4	100.0 ± 0.0	50.0 ± 35.4	50.0 ± 35.4
CXCR3(1)	50.0 ± 35.4	0.0 ± 0.0	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4
EDNRA(50)	100.0 ± 0.0	99.0 ± 0.9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
EDNRB(48)	96.9 ± 1.1	91.8 ± 3.4	98.0 ± 1.1	99.0 ± 0.9	99.0 ± 0.9
GASR(2)	100.0 ± 0.0	75.0 ± 21.7	75.0 ± 21.7	75.0 ± 21.7	75.0 ± 21.7
GPR7(1)	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4
LSHR(4)	70.0 ± 11.0	70.0 ± 11.0	70.0 ± 11.0	70.0 ± 11.0	70.0 ± 11.0
NK1R(24)	92.0 ± 4.4	82.0 ± 5.2	86.0 ± 5.4	88.0 ± 3.3	86.0 ± 3.6
NK2R(1)	50.0 ± 35.4	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
NK3R(1)	50.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
OPRD(27)	92.3 ± 1.7	86.7 ± 4.4	90.3 ± 4.9	90.3 ± 2.8	90.3 ± 2.8
OPRK(24)	96.0 ± 3.6	98.0 ± 1.8	98.0 ± 1.8	98.0 ± 1.8	98.0 ± 1.8
OPRM(21)	100.0 ± 0.0	97.5 ± 2.2	97.5 ± 2.2	97.5 ± 2.2	97.5 ± 2.2
OXYR(3)	90.0 ± 8.9	100.0 ± 0.0	90.0 ± 8.9	100.0 ± 0.0	100.0 ± 0.0
SSR1(3)	90.0 ± 8.9	90.0 ± 8.9	90.0 ± 8.9	90.0 ± 8.9	90.0 ± 8.9
CCR3(1)	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4
Rhodopsin amine receptors					
5HT1A(196)	91.6 ± 1.3	90.1 ± 2.2	88.8 ± 0.8	91.8 ± 1.5	90.8 ± 1.7
5HT1B(28)	82.7 ± 3.0	96.0 ± 3.6	98.0 ± 1.8	100.0 ± 0.0	100.0 ± 0.0
5HT1D(172)	93.3 ± 1.0	92.4 ± 0.9	92.7 ± 0.9	94.8 ± 0.7	94.8 ± 0.7
5HT1E(16)	87.5 ± 5.5	90.8 ± 3.4	96.7 ± 3.0	90.8 ± 3.4	90.8 ± 3.4
5HT1F(49)	86.7 ± 1.2	90.9 ± 0.8	88.8 ± 1.7	92.9 ± 1.1	91.7 ± 2.1
5HT2A(79)	94.9 ± 1.4	95.6 ± 1.4	93.0 ± 1.7	94.3 ± 1.7	94.9 ± 1.4
5HT2B(72)	81.2 ± 3.3	78.3 ± 2.9	83.9 ± 1.8	83.2 ± 2.0	83.2 ± 2.0
5HT2C(198)	88.6 ± 1.2	86.8 ± 1.2	89.4 ± 1.4	89.6 ± 0.8	90.1 ± 1.3
5HT4R(87)	92.5 ± 2.0	86.7 ± 2.5	85.7 ± 2.0	87.9 ± 2.1	89.0 ± 2.0
5HT5A(7)	80.0 ± 8.4	75.0 ± 10.0	75.0 ± 10.0	75.0 ± 10.0	75.0 ± 10.0
5HT6R(13)	95.0 ± 4.5	96.7 ± 3.0	91.7 ± 4.7	95.0 ± 4.5	100.0 ± 0.0
5HT7R(15)	90.0 ± 6.0	90.0 ± 3.7	96.7 ± 3.0	93.3 ± 3.7	93.3 ± 3.7
ACM1(527)	96.7 ± 0.6	94.3 ± 0.9	95.5 ± 1.0	96.1 ± 0.7	96.1 ± 0.8
ACM2(24)	82.0 ± 5.2	90.0 ± 2.8	92.0 ± 3.3	94.0 ± 3.6	92.0 ± 3.3
ACM3(58)	93.2 ± 2.6	90.5 ± 0.7	91.3 ± 1.3	96.4 ± 1.5	95.6 ± 1.3
ACM4(21)	90.0 ± 5.5	95.0 ± 2.7	95.0 ± 2.7	92.5 ± 2.7	95.0 ± 2.7
ACM5(16)	94.2 ± 3.2	94.2 ± 3.2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0

GPCR $\backslash K_{tar}$	Dirac	multitask	hierarchy	BP	PBP
ADA1A(80)	93.1 ± 2.1	98.8 ± 0.7	99.4 ± 0.6	98.1 ± 0.7	98.8 ± 0.7
ADA1B(67)	90.5 ± 3.7	95.7 ± 1.9	98.6 ± 0.8	97.0 ± 0.7	97.0 ± 0.7
ADA1D(73)	90.4 ± 2.4	96.0 ± 1.1	98.7 ± 0.7	98.0 ± 0.7	98.0 ± 0.7
ADA2A(234)	95.7 ± 0.5	96.8 ± 0.3	98.5 ± 0.2	98.5 ± 0.2	98.5 ± 0.2
ADA2B(224)	95.1 ± 1.2	95.5 ± 1.3	98.2 ± 0.7	98.2 ± 0.7	98.0 ± 0.7
ADA2C(225)	95.3 ± 0.4	96.4 ± 0.4	97.6 ± 0.4	97.6 ± 0.4	97.8 ± 0.3
ADRB1(50)	98.0 ± 1.1	97.0 ± 1.8	99.0 ± 0.9	99.0 ± 0.9	99.0 ± 0.9
ADRB2(48)	92.8 ± 1.9	95.9 ± 0.9	96.9 ± 1.1	98.0 ± 1.1	98.0 ± 1.1
ADRB3(57)	98.2 ± 1.0	95.5 ± 2.2	97.3 ± 1.6	97.3 ± 1.6	97.3 ± 1.6
DRD1(100)	93.5 ± 1.8	94.5 ± 1.5	95.0 ± 1.4	94.5 ± 1.3	94.5 ± 1.3
DRD2(106)	93.4 ± 0.8	92.9 ± 1.8	92.4 ± 1.6	91.5 ± 1.7	91.9 ± 1.9
DRD3(41)	86.7 ± 2.6	89.2 ± 3.1	89.3 ± 3.8	90.4 ± 3.2	91.5 ± 2.8
DRD4(143)	92.3 ± 0.8	92.7 ± 1.1	93.7 ± 1.3	93.7 ± 1.4	94.1 ± 1.3
DRD5(7)	95.0 ± 4.5	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
HRH1(19)	89.2 ± 4.3	92.5 ± 2.7	86.7 ± 0.7	92.5 ± 2.7	92.5 ± 2.7
HRH2(22)	91.0 ± 3.5	93.5 ± 3.7	96.0 ± 3.6	96.0 ± 3.6	96.0 ± 3.6
HRH3(88)	97.2 ± 0.8	96.1 ± 1.3	97.7 ± 0.9	97.7 ± 0.5	97.7 ± 0.5
HRH4(5)	80.0 ± 11.0	70.0 ± 17.9	100.0 ± 0.0	80.0 ± 11.0	80.0 ± 11.0
Rhodopsin other receptors					
AA1R(56)	96.4 ± 1.5	96.4 ± 0.8	96.4 ± 1.5	97.3 ± 1.0	97.3 ± 1.0
AA2AR(73)	96.0 ± 1.7	97.3 ± 1.1	98.6 ± 0.8	98.0 ± 1.2	98.0 ± 1.2
AA2BR(83)	97.6 ± 1.0	98.2 ± 0.7	99.4 ± 0.6	99.4 ± 0.6	99.4 ± 0.6
AA3R(17)	97.5 ± 2.2	82.5 ± 1.8	94.2 ± 3.2	95.0 ± 4.5	95.0 ± 4.5
CLTR1(18)	89.2 ± 2.5	84.2 ± 4.1	89.2 ± 4.3	91.7 ± 3.1	91.7 ± 3.1
LT4R1(2)	50.0 ± 25.0	50.0 ± 25.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
LT4R2(2)	50.0 ± 25.0	50.0 ± 25.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MTR1A(91)	97.3 ± 1.1	96.8 ± 1.4	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MTR1B(90)	97.8 ± 0.9	97.8 ± 0.9	99.4 ± 0.5	99.4 ± 0.5	99.4 ± 0.5
MTR1L(75)	98.7 ± 0.7	99.3 ± 0.6	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
PAFR(1)	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4
PE2R1(5)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
PE2R2(7)	100.0 ± 0.0	95.0 ± 4.5	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
PE2R3(5)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
PE2R4(5)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
R3R2(1)	50.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
TA2R(63)	100.0 ± 0.0	99.2 ± 0.7	99.2 ± 0.7	100.0 ± 0.0	100.0 ± 0.0

GPCR $\backslash K_{tar}$	Dirac	multitask	hierarchy	BP	PBP
Metabotropic glutamate family					
GABR1(1)	50.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	50.0 ± 35.4	100.0 ± 0.0
GABR2(1)	50.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0	50.0 ± 35.4
MGR1(34)	98.3 ± 1.5	91.4 ± 4.7	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MGR2(6)	95.0 ± 4.5	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MGR3(5)	100.0 ± 0.0	90.0 ± 8.9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MGR5(5)	90.0 ± 8.9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MGR6(5)	100.0 ± 0.0	90.0 ± 8.9	90.0 ± 8.9	100.0 ± 0.0	90.0 ± 8.9
MGR7(6)	95.0 ± 4.5	90.0 ± 8.9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MGR8(3)	80.0 ± 17.9	80.0 ± 17.9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Secretin family					
VIPR1(1)	50.0 ± 35.4	100.0 ± 0.0	100.0 ± 0.0	50.0 ± 35.4	100.0 ± 0.0

Table 3.6: Prediction accuracy by GPCR for the first experiment. Mean prediction accuracy for each GPCR for the first experiment with the 2D Tanimoto ligand kernel and various target kernels. The GPCR identifiers are the GLIDA references. The numbers in bracket are the numbers ligands considered in the experiment for each GPCR. BP is the binding pocket kernel and PBP the poly binding pocket kernel.

Family	Dirac	multitask	hierarchy	BP	PBP
Rhodopsin peptide receptors					
AG2R(5)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
CCKAR(6)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
CML2(1)	50.0 ± 0.0	50.0 ± 35.4	100.0 ± 0.0	50.0 ± 35.4	50.0 ± 35.4
CXCR3(1)	50.0 ± 35.4	0.0 ± 0.0	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4
EDNRA(50)	100.0 ± 0.0	99.0 ± 0.9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
EDNRB(48)	96.9 ± 1.1	91.8 ± 3.4	98.0 ± 1.1	99.0 ± 0.9	99.0 ± 0.9
GASR(2)	100.0 ± 0.0	75.0 ± 21.7	75.0 ± 21.7	75.0 ± 21.7	75.0 ± 21.7
GPR7(1)	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4
LSHR(4)	70.0 ± 11.0	70.0 ± 11.0	70.0 ± 11.0	70.0 ± 11.0	70.0 ± 11.0
NK1R(24)	92.0 ± 4.4	82.0 ± 5.2	86.0 ± 5.4	88.0 ± 3.3	86.0 ± 3.6
NK2R(1)	50.0 ± 35.4	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
NK3R(1)	50.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
OPRD(27)	92.3 ± 1.7	86.7 ± 4.4	90.3 ± 4.9	90.3 ± 2.8	90.3 ± 2.8
OPRK(24)	96.0 ± 3.6	98.0 ± 1.8	98.0 ± 1.8	98.0 ± 1.8	98.0 ± 1.8
OPRM(21)	100.0 ± 0.0	97.5 ± 2.2	97.5 ± 2.2	97.5 ± 2.2	97.5 ± 2.2
OXYR(3)	90.0 ± 8.9	100.0 ± 0.0	90.0 ± 8.9	100.0 ± 0.0	100.0 ± 0.0
SSR1(3)	90.0 ± 8.9	90.0 ± 8.9	90.0 ± 8.9	90.0 ± 8.9	90.0 ± 8.9
CCR3(1)	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4
Rhodopsin amine receptors					
5HT1A(196)	91.6 ± 1.3	90.1 ± 2.2	88.8 ± 0.8	91.8 ± 1.5	90.8 ± 1.7
5HT1B(28)	82.7 ± 3.0	96.0 ± 3.6	98.0 ± 1.8	100.0 ± 0.0	100.0 ± 0.0
5HT1D(172)	93.3 ± 1.0	92.4 ± 0.9	92.7 ± 0.9	94.8 ± 0.7	94.8 ± 0.7
5HT1E(16)	87.5 ± 5.5	90.8 ± 3.4	96.7 ± 3.0	90.8 ± 3.4	90.8 ± 3.4
5HT1F(49)	86.7 ± 1.2	90.9 ± 0.8	88.8 ± 1.7	92.9 ± 1.1	91.7 ± 2.1
5HT2A(79)	94.9 ± 1.4	95.6 ± 1.4	93.0 ± 1.7	94.3 ± 1.7	94.9 ± 1.4
5HT2B(72)	81.2 ± 3.3	78.3 ± 2.9	83.9 ± 1.8	83.2 ± 2.0	83.2 ± 2.0
5HT2C(198)	88.6 ± 1.2	86.8 ± 1.2	89.4 ± 1.4	89.6 ± 0.8	90.1 ± 1.3
5HT4R(87)	92.5 ± 2.0	86.7 ± 2.5	85.7 ± 2.0	87.9 ± 2.1	89.0 ± 2.0
5HT5A(7)	80.0 ± 8.4	75.0 ± 10.0	75.0 ± 10.0	75.0 ± 10.0	75.0 ± 10.0
5HT6R(13)	95.0 ± 4.5	96.7 ± 3.0	91.7 ± 4.7	95.0 ± 4.5	100.0 ± 0.0
5HT7R(15)	90.0 ± 6.0	90.0 ± 3.7	96.7 ± 3.0	93.3 ± 3.7	93.3 ± 3.7
ACM1(527)	96.7 ± 0.6	94.3 ± 0.9	95.5 ± 1.0	96.1 ± 0.7	96.1 ± 0.8
ACM2(24)	82.0 ± 5.2	90.0 ± 2.8	92.0 ± 3.3	94.0 ± 3.6	92.0 ± 3.3
ACM3(58)	93.2 ± 2.6	90.5 ± 0.7	91.3 ± 1.3	96.4 ± 1.5	95.6 ± 1.3
ACM4(21)	90.0 ± 5.5	95.0 ± 2.7	95.0 ± 2.7	92.5 ± 2.7	95.0 ± 2.7
ACM5(16)	94.2 ± 3.2	94.2 ± 3.2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0

Family	Dirac	multitask	hierarchy	BP	PBP
ADA1A(80)	93.1 ± 2.1	98.8 ± 0.7	99.4 ± 0.6	98.1 ± 0.7	98.8 ± 0.7
ADA1B(67)	90.5 ± 3.7	95.7 ± 1.9	98.6 ± 0.8	97.0 ± 0.7	97.0 ± 0.7
ADA1D(73)	90.4 ± 2.4	96.0 ± 1.1	98.7 ± 0.7	98.0 ± 0.7	98.0 ± 0.7
ADA2A(234)	95.7 ± 0.5	96.8 ± 0.3	98.5 ± 0.2	98.5 ± 0.2	98.5 ± 0.2
ADA2B(224)	95.1 ± 1.2	95.5 ± 1.3	98.2 ± 0.7	98.2 ± 0.7	98.0 ± 0.7
ADA2C(225)	95.3 ± 0.4	96.4 ± 0.4	97.6 ± 0.4	97.6 ± 0.4	97.8 ± 0.3
ADRB1(50)	98.0 ± 1.1	97.0 ± 1.8	99.0 ± 0.9	99.0 ± 0.9	99.0 ± 0.9
ADRB2(48)	92.8 ± 1.9	95.9 ± 0.9	96.9 ± 1.1	98.0 ± 1.1	98.0 ± 1.1
ADRB3(57)	98.2 ± 1.0	95.5 ± 2.2	97.3 ± 1.6	97.3 ± 1.6	97.3 ± 1.6
DRD1(100)	93.5 ± 1.8	94.5 ± 1.5	95.0 ± 1.4	94.5 ± 1.3	94.5 ± 1.3
DRD2(106)	93.4 ± 0.8	92.9 ± 1.8	92.4 ± 1.6	91.5 ± 1.7	91.9 ± 1.9
DRD3(41)	86.7 ± 2.6	89.2 ± 3.1	89.3 ± 3.8	90.4 ± 3.2	91.5 ± 2.8
DRD4(143)	92.3 ± 0.8	92.7 ± 1.1	93.7 ± 1.3	93.7 ± 1.4	94.1 ± 1.3
DRD5(7)	95.0 ± 4.5	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
HRH1(19)	89.2 ± 4.3	92.5 ± 2.7	86.7 ± 0.7	92.5 ± 2.7	92.5 ± 2.7
HRH2(22)	91.0 ± 3.5	93.5 ± 3.7	96.0 ± 3.6	96.0 ± 3.6	96.0 ± 3.6
HRH3(88)	97.2 ± 0.8	96.1 ± 1.3	97.7 ± 0.9	97.7 ± 0.5	97.7 ± 0.5
HRH4(5)	80.0 ± 11.0	70.0 ± 17.9	100.0 ± 0.0	80.0 ± 11.0	80.0 ± 11.0
Rhodopsin other receptors					
AA1R(56)	96.4 ± 1.5	96.4 ± 0.8	96.4 ± 1.5	97.3 ± 1.0	97.3 ± 1.0
AA2AR(73)	96.0 ± 1.7	97.3 ± 1.1	98.6 ± 0.8	98.0 ± 1.2	98.0 ± 1.2
AA2BR(83)	97.6 ± 1.0	98.2 ± 0.7	99.4 ± 0.6	99.4 ± 0.6	99.4 ± 0.6
AA3R(17)	97.5 ± 2.2	82.5 ± 1.8	94.2 ± 3.2	95.0 ± 4.5	95.0 ± 4.5
CLTR1(18)	89.2 ± 2.5	84.2 ± 4.1	89.2 ± 4.3	91.7 ± 3.1	91.7 ± 3.1
LT4R1(2)	50.0 ± 25.0	50.0 ± 25.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
LT4R2(2)	50.0 ± 25.0	50.0 ± 25.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MTR1A(91)	97.3 ± 1.1	96.8 ± 1.4	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MTR1B(90)	97.8 ± 0.9	97.8 ± 0.9	99.4 ± 0.5	99.4 ± 0.5	99.4 ± 0.5
MTR1L(75)	98.7 ± 0.7	99.3 ± 0.6	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
PAFR(1)	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4	50.0 ± 35.4
PE2R1(5)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
PE2R2(7)	100.0 ± 0.0	95.0 ± 4.5	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
PE2R3(5)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
PE2R4(5)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
R3R2(1)	50.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
TA2R(63)	100.0 ± 0.0	99.2 ± 0.7	99.2 ± 0.7	100.0 ± 0.0	100.0 ± 0.0

Family	Dirac	multitask	hierarchy	BP	PBP
Metabotropic glutamate family:					
GABR1(1)	50.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	50.0 ± 35.4	100.0 ± 0.0
GABR2(1)	50.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0	50.0 ± 35.4
MGR1(34)	98.3 ± 1.5	91.4 ± 4.7	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MGR2(6)	95.0 ± 4.5	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MGR3(5)	100.0 ± 0.0	90.0 ± 8.9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MGR5(5)	90.0 ± 8.9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MGR6(5)	100.0 ± 0.0	90.0 ± 8.9	90.0 ± 8.9	100.0 ± 0.0	90.0 ± 8.9
MGR7(6)	95.0 ± 4.5	90.0 ± 8.9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
MGR8(3)	80.0 ± 17.9	80.0 ± 17.9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Secretin family:					
VIPR1(1)	50.0 ± 35.4	100.0 ± 0.0	100.0 ± 0.0	50.0 ± 35.4	100.0 ± 0.0

Table 3.7: Prediction accuracy by GPCR for the second experiment. Mean prediction accuracy for each GPCR for the second experiment with the 2D Tanimoto ligand kernel and various target kernels. The GPCR identifiers are the GLIDA references. The numbers in bracket are the numbers ligands considered in the experiment for each GPCR. BP is the binding pocket kernel and PBP the poly binding pocket kernel.

4

Protein binding pocket similarity measure based on comparison of clouds of atoms in 3D

The chemogenomic approach presented in chapter 3 allows to predict ligands for proteins belonging to the same family or superfamily, because the similarity measures used imply that the proteins can be aligned (at least locally, in the case of the binding pocket approach). For proteins belonging to unrelated families, such sequence alignment is not relevant. However, the advantage of this approach was that it did not rely on prior knowledge of 3D structure for all the considered proteins (for the hierarchy kernel), or it required only one known 3D structure in the family (for the binding pocket kernel).

In this chapter, we will present a chemogenomic approach which is applicable to proteins belonging to totally different families, as long as their 3D structures are available. In other words, this method allows to share ligand information between proteins to imrpove performance in prediction of protein-ligand interaction, under the restriction that 3D structures need to be known.

4.1 Background

Predicting which molecules can bind to a given binding site of a protein with known 3D structure is important to decipher the protein function, and useful in drug design to identify drug precursors or predict potential side effects if a drug candidate is predicted to bind to many pro-

tein pockets. A classical assumption in structural biology is that the 3D structure of a protein is related to its molecular function, i.e., the nature of its partner molecules. Most available programs for structure visualization provide tools for 3D structure superposition and comparison, which may help to predict the nature of a protein ligand from those of other proteins with overall similar 3D structure (155).

However, proteins that do not display any overall sequence or structure similarity may present similar binding sites, and consequently also share similar ligands. Therefore, comparison of binding pockets is a more appropriate approach in order to predict if two proteins bind similar ligands (156), and many ligand prediction methods rely on local 3D comparisons at the level of the binding site; whether or not the overall 3D structures overlap.

For example, (157) compared pockets described with real spherical harmonic expansion coefficients. These coefficients are used to describe the shape of a binding pocket. (158) used a specialized geometric hashing procedure as the core of the SitesBase web server. (159) developed a method that detects multiple common sets of points. An approach proposed by (160) is based on the representation of binding pockets by triangle-discretized spheres. (161) and (162) considered graph-based representations of binding pockets and applied graph matching algorithms. Finally, (163; 164) combines the identification of a binding site on a whole protein 3D structure and its comparison to a reference binding site, using a geometric hashing procedure.

Our contribution in this chapter is twofold.

First, we propose a new similarity measure to compare binding pockets of proteins. For that purpose, we represent a binding pocket by a cloud of atoms in the 3D space, potentially baring labels such as partial charges or atom types. The method relies on the representation of local protein structures are rigid bodies, and we therefore represent a protein pocket as a cloud of points with fixed relative positions. The method performs a superposition of two pockets even if their corresponding proteins present no overall sequence or 3D structure similarity. Then, a pocket similarity is measured based on a convolution kernel between clouds of points. One important difference between this approach and most existing methods is that it does not require any pairwise matching of atoms (or superatoms), or residues, in order to compare protein binding pockets. Instead, we attempt to capture the similarity of atom densities in the 3D space. This confers smoothness properties to the proposed similarity measure.

Second, we propose to use a classification method to predict ligands for target pockets, according to their similarity scores with a set of pockets with known ligands. This approach is able to handle the difficult case where different families of pockets binding the same ligand

are present. This may be observed when the ligand is flexible and can be bound in various conformations by pockets displaying different topologies, as shown in the case of ATP in the result section.

An important question that we will debate is how to compare the quality of similarity measures. Although the area under ROC curves (AUC scores) are commonly used (156), we show that classification-based scores better compare the performances of similarity measures for ligand prediction. We underline that it is not possible to define an intrinsic quality for a similarity measure, because there is no absolute reference. Similarity measures can only be compared according to the question of interest. Here, we evaluate quality of similarity measures with respect to their ability to predict a ligand for a pocket.

We test our method on a benchmark proposed by other authors, in order to compare our new method to other published algorithms. We also test the methods on a new benchmark containing non redundant protein pockets binding ligands of similar sizes, typical of that of drug molecules, corresponding to a more realistic problem. We provided this new dataset as a publicly available benchmark.

In the following, we will first present methods used to encode, superpose and compare pockets. We will briefly review other related methods used in this study as baseline methods to which we will compare our new methods. We present the performance criterion used to evaluate methods. Then the benchmark datasets used in this study are presented before presenting the results.

4.2 Methods

4.2.1 Convolution kernel between clouds of atoms

In our model, a binding pocket is described by a set of atoms in the 3D space. Our objective is to construct a similarity measure between pockets, which may be used to identify pockets binding the same ligand.

Let $P = (x_i, l_i)_{i=1}^{N}$ denote a binding pocket consisting of N atoms, where $x_i \in \mathbb{R}^3$ is a 3D vector representing atom coordinates, and l_i is a label (discrete or real valued) that may be used to store additional information on the atoms (for example, atom type, atom partial charge, or amino acid type).

4. PROTEIN BINDING POCKET SIMILARITY MEASURE BASED ON COMPARISON OF CLOUDS OF ATOMS IN 3D

A classical approach for pocket comparison is to iteratively align two pockets and further count the number of overlapping atoms, usually within a tolerance in the range of 1Å. Different implementations of this principle can be found in such methods as the Tanimoto index (165), the SitesBase algorithm (Poisson index, (166)) , or the MultiBind algorithm (159). The alignment is made to maximize the number of overlapping atoms, which is generally a good indicator of pocket similarity.

However, atoms may have different positions but play equivalent roles in ligand binding (for example, the side chain of a basic residue may bind a phosphate group of an ATP molecule from different positions), and the role of one atom in one pocket may be played by a group of atoms in another one. These observations suggest the idea of an alternative smooth score which would not count the number of overlapping atoms, but rather use a weighted number of atoms having similar positions. We first consider the case where labels are ignored, and only atom coordinates are used to measure the similarity between pockets. Then, we explain how the information on atom labels may be introduced in the new similarity measure.

Given two pockets P_1 and P_2 the similarity measure $K(P_1, P_2)$, we used the similarity measure is defined as follows

$$K(P_1, P_2) = \sum_{x_i \in P_1} \sum_{y_j \in P_2} e^{\frac{-||x_i - y_j||^2}{2\sigma^2}}. \tag{4.1}$$

This similarity measure corresponds to a classically used positive definite gaussian kernel.

Parameter sigma can be view as a smoothing parameter allowing comparison of density of points instead of pairwise comparison between atom pairs belonging to pocket P_1 and P_2.

In addition, it may be considered as a true scalar product on atom clouds that represent binding pockets (75).

Consequently, as for all scalar product, it can define a distance :

$$D(P_1, P_2) = \sqrt{K(P_1, P_1) + K(P_2, P_2) - 2K(P_1, P_2)}\,. \tag{4.2}$$

This distance defines a distance between pockets.

The parameter σ characterizes the sensitivity of the similarity measure (formula 4.1) to points relative displacements. When σ is small, only atoms which are close to each other significantly contribute to $K(P_1, P_2)$. On the contrary, when σ is large, almost all pairs of atoms contribute to $K(P_1, P_2)$.

However, formula (4.1) is not fully appropriate in practice, because the proposed measure is not invariant upon rotations and translations of the binding pockets.

Indeed, suppose that two identical pockets are placed in two different coordinate systems, which is the case for most protein structures deposited at the PDB. The similarity measure between these two identical pockets will not be equal to one because the distance between each corresponding atoms of the two pockets is not equal to zero.

Therefore, to overcome this problem we define a similarity measure *sup-CK* as the maximum of (4.1) over all possible rotations and translations of one of the two pockets:

$$\text{sup-CK}(P_1, P_2) = \max_{R, y_t} \sum_{x_i \in P_1, y_j \in P_2} e^{\frac{-||x_i - (Ry_j + y_t)||^2}{2\sigma^2}} \tag{4.3}$$

Where R is an orthonormal rotation matrix and y_t is a translation vector. This transformation aims to bring the two pockets in the same coordinate system, and find the best possible superposition of the two pockets by rotation or translation of one pocket over the other.

The problem we encounter now is that to evaluate *sup-CK*, we now need to maximize a non-concave function over the set of rotations and translations, which may have many local maxima. *Sup-CK* is not a positive definite measure anymore, but can still be used as a similarity score. Exact maximization of this non-concave function is a hard optimization problem. An approximate solution can be estimated by running a gradient ascent algorithm, starting from many different initial points, and taking the best local maximum. Choosing initial points near the global optimal can then help find a better solution and accelerate the optimization. In the case of binding pockets, we found experimentally that, rather than starting from random initial points, a good approximation of the optimal translation vector y_t is the vector which translates the geometric center of P_2 into the geometric center of P_1:

$$y_t = \frac{1}{N_1} \sum_{x_i \in P_1} x_i - \frac{1}{N_2} \sum_{y_i \in P_2} y_i$$

Similarly, an approximation of the optimal rotation matrix R is the rotation that superposes the first principal axis of P_2 with the first principal axis of P_1, the second one with the second one, and the third one with the third one.

Once this starting point as been found, a gradient ascent method can be used to find the maximum. This requires to calculate the gradient of the function in (see formula 4.3) with respect to R and y_t.

This optimization step will not be detailed here, but it defines the best pocket superposition, according to the *sup-CK* similarity measure.

As mentioned above, it may be interesting to use additional information on binding pocket atoms, such as atom types or charges. Let us suppose that this information is represented by labels l_i (which may be discrete or real variables, or multidimensional vectors) and that it is associated to a similarity measure. For example, to measure the similarity between categorical labels like atom types, one can use the Dirac function $1_{l_i=l_j}$. In our experiments, we used atom partial charges as atom labels, with a Gaussian kernel $K_{\mathrm{L}}(l_i, l_j) = e^{-\frac{(l_i-l_j)^2}{\lambda}}$. Of course, other similarity measures may be employed.

These atom labels can be used to re-weight the contribution of two atoms x_i and y_j by $K_{\mathrm{L}}(l_i, l_j)$ in (4.3):

$$\text{sup-CK}_L(P_1, P_2) = \max_{R,y_t} \sum_{\substack{x_i \in P_1 \\ y_j \in P_2}} e^{-\frac{(l_i-l_j)^2}{\lambda}} e^{\frac{-||x_i-(Ry_j+y_t)||^2}{2\sigma^2}} \tag{4.4}$$

where parameter λ controls the sensitivity of our measure to atom labels, for example to partial charges. When λ is large, the impact of labels is negligible, which corresponds to a purely geometrical approach. When λ is close to zero, only pairs of atoms which have the same partial charge contribute to our measure. In general, the smaller λ, the greater the contribution of the atom labels to the binding pocket similarity measure. Since the function K_{L} does not depend on R and y_t in (4.4), the same optimization procedure for pockets superposition can be used to optimize (4.3) or (4.4).

Finally, it is important to notice that the *sup-CK* measure of similarity can be used to compare *any* set of atoms in 3D. As mentioned in the introduction section, the superposition method and the similarity measure may be applied to superpose and compare pockets, even when they belong to proteins displaying no sequence and no overall structure similarity. This point will be illustrated in Results on the example of two unrelated ATP binding proteins.

4.2.2 Related methods

In the following, we briefly recall the principals of a few other methods proposed to measure similarity between pockets, because we compare them to the *sup-CK* method defined in the present study.

Spherical harmonic decomposition (SHD). (157) proposed to model pockets by star-shapes built using the SURFNET program. The star-shape representation is defined by a function of

spherical coordinates $f(\theta,\phi)$, representing the distance from the pocket center to the pocket surface for a given (θ,ϕ). To measure the similarity of binding pockets P_1 and P_2, the corresponding functions f_1 and f_2 are first decomposed into spherical harmonics, and the pocket similarity is then computed as the standard Euclidean metric between vectors of decomposition coefficients. They transform the binding pocket to a standard frame of reference and compute the real spherical harmonic expansion coefficients that best approximate the shape of the pocket.

(156) presented three different variants of *SHD*, using only the shapes or the sizes of the binding pockets (keeping only the zero-th order in the spherical harmonics expansion), and their combination. As the zeroth order of the spherical harmonic coefficients reflects the general size of a shape, the division of all coefficients by the zeroth order coefficient, places the shapes on the same scale and thereby removes the influence of different sized objects.

Poisson index (sup-PI). As mentioned in the Background section 4.1 of this chapter, many binding pockets similarity measures are based on pocket alignment with further counting of overlapping atoms. This kind of approach is used in the *Poisson index* model (166). More precisely, the *Poisson index* model is based on a normalized number of overlapping atoms $PI(P_1, P_2) = \frac{L}{\#P_1+\#P_2-L}$, where L is the number of overlapping atoms, and $\#P_1$ and $\#P_2$ are the numbers of atoms in P_1 and P_2, respectively. The PI score may be computed for any pocket superposition method. While (166) used the geometric hashing algorithm to perform superposition, we used the superposition made by the *sup-CK* method.

Multibind. (159) represents pockets by pseudo-atoms labeled with physicochemical properties. Pockets are aligned using a geometric hashing technique. This algorithm was mainly designed for multiple alignment of binding sites, but it may be used for pairwise alignment of pockets, as performed in this study.

Other simple methods. We also considered two simple methods based on the comparison of simple binding pockets characteristics. These methods represent each pocket by an ellipsoid constructed on the basis of the pocket's principal axis. The first one, referred to as *Vol*, estimates the similarity between pockets P_1 and P_2 by the absolute value of the difference between the volumes of their corresponding ellipsoids: $Vol(P_1, P_2) = |Vol(P_1) - Vol(P_2)|$. The second one, called *Princ-Axis*, estimates the similarity score between pockets by $\sum_{i=1}^{3}(\lambda_i^{P_1} - \lambda_i^{P_2})^2$, where $\lambda_i^{P_1}$ and $\lambda_i^{P_2}$ are the lengths of the three principle axis of pockets P_1 and P_2, respectively.

Combination of sup-CK and Vol. Since volume information was found to be important by (156), we also tested a linear combination of the *sup-CK* and *Vol* methods, called *sup-CK-Vol*,

where the coefficient of linear combination is learned as other model parameters (σ, λ, or the distance cutoff R discussed in the Datasets section) in the double cross validation scheme. This linear combination takes advantage of the *Vol* method to separate pockets binding ligands of very different sizes like PO4 and NAD, and of the *sup-CK* algorithm to allow finer discrimination.

Sequence. To compare our method based on local 3D similarity to a simple and classical approach based on sequence comparison, we conducted a pairwise alignment of all protein sequences for the different datasets, in order to build a matrix of distance between proteins based on sequence similarity. This matrix was built with the algorithm of Needleman and Wunsch, using the default settings (167; 168).

4.2.3 Performance criteria

There are various ways to measure the similarity between binding pockets, and some of them were discussed in the previous section. To evaluate the quality of a given similarity measure, one may compare it to some "ideal" similarity measure, but the problem is that such measure does not exist. As an example, if two alternative similarity measures $SM1$ and $SM2$ compare two pockets P_1 and P_2 so that $SM1(P_1, P_2) = 0.3$ and $SM2(P_1, P_2) = 0.4$, there is no way to decide which one is the best, because we do not have any absolute reference. The choice of the optimal measure, thus, depends on the problem of interest. In the context of ligand prediction, the quality of a similarity measure can be evaluated according to its ability to cluster together pockets that bind the same ligand. This can then help to predict ligands for previously unseen pockets. To evaluate this clustering ability, we considered two different scores.

AUC score. (156) used the AUC score which is computed as follows. Let us consider a set of pockets $(P_1, \ldots, P_N)$ and a similarity measure SM. To estimate the AUC score of a given pocket P_*, we rank all other pockets according to their similarity to P_*, $SM(P_i, P_*)$ (descending order), and we plot the ROC curve, i.e., the number of pockets binding the same ligand versus the number of pockets binding a different ligand among the top n pockets, when n varies from 0 to N. The quality of SM is measured by the surface of the area under the ROC curve, which defines the AUC score. An "ideal" SM function will rank all pockets binding the same ligand as P_* on the top of the list, leading to an AUC score equal to 1.0. On the contrary, if these pockets have random positions in the ranked list, the AUC score will be equal to 0.5 (worst possible case). Finally, the overall AUC score of a method equals its mean value over all pockets.

While the AUC score represents an intuitive and classical way to evaluate the quality of similarity measures, it may fail in some situations. Consider the case of a dataset containing two types of pockets L_1 and L_2 (i.e. binding two different ligands), and a similarity measure that correctly clusters pockets according to their type. If clusters are close to each other (see blue squares in Figure 4.1), the AUC score of pockets situated near the border (pockets p_1 and p_2 in Figure 4.1) will be low. The situation becomes even worse if pockets binding ligand L_1 form several clusters, as shown in Figure 4.1, leading to low AUC scores for almost all pockets binding ligand L_1. This similarity measure will have an overall poor AUC score on this dataset, although it produces perfect separation of pocket types. This may happen when the database contains proteins that underwent convergent evolution, or that bind the same ligand under very different conformations. Therefore, a poor AUC score does not necessarily correspond to a poor pocket separation, and AUC scores may not be suited to evaluate the quality of similarity measures with respect to the question of ligand prediction.

Classification error. These remarks lead us to employ another quality score based on a classification error. To estimate the quality of the similarity measure SM, we try to predict a ligand (i.e. a class) for each pocket from that of its neighbors. The smaller the classification error (proportion of bad predictions), the better the similarity measure.

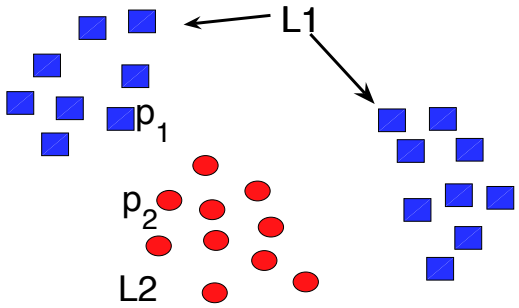


Figure 4.1: AUC score versus classification error as an evaluation of binding pocket similarity measure. Red circles represents pockets fixing ligand L1, blue squares represents pockets fixing ligand L2. The AUC score does not reflect the fact of good pocket clusterization, while the classification error does.

In this work, we used a K nearest neighbors (KNN) classifier. A KNN is a method for classifying objects based on closest training examples. An object is classified by a majority

vote of its neighbors, with the object being assigned to the most common class amongst its K nearest neighbors (K is a positive integer, typically small). If K = 1, then the object is simply assigned to the class of its nearest neighbor.

To evaluate the classification error, we applied a leave-one-out double cross validation methodology. Namely, each pocket P from the dataset is considered one by one, and all other pockets are used as the training set. Parameters of the model (K number of neighbors, σ and λ in the case of the *sup-CK* method) are estimated on the training set via cross-validation technique, and the class (i.e. the ligand) of the pocket P under consideration is predicted using the training set and the estimated parameters of the model. More precisely, in the case of a dataset containing 100 proteins, double cross validation is performed according to the following scheme: each of the 100 pockets is extracted in turn from the dataset in a leave one out procedure. Then, each of the other 99 pockets is selected in turn and its class is predicted from the 98 remaining pockets. This operation is repeated for different values of σ and λ, and the σ^* and λ^* values providing the highest number of well predicted pockets (over 99) are retained and used to predict a class for the initially extracted pocket.

Note, that all datasets contained proteins that presented less than 30% global sequence identity (167), to ensure that there were no duplicates or very close elements in the datasets. This allowed to use a leave-one-out scheme without risk of bias.

4.2.4 Data

For all protein structures, binding pockets were extracted as follows: protein atoms situated at less than R Åof one of the ligand atoms were selected, where R is a parameter of the model (as the number of neighbors k, or the σ and λ parameters), and is also learned in the double cross-validation scheme. In most cases, the optimal value of R was found to equal to 5.3 Å, a value which was retained in this study.

Note that this value of R is in the range of the distance separating atoms over which most physical interactions are considered to become negligible.

However, experiments where R is varied are also presented in the discussion section. Finally, pockets are represented by 3D clouds of atoms labeled by their partial charge, attributed according to the GROMACS (FFG43a1) force field (169). Atom partial charges were assigned according to the protein structure alone, in absence of the ligand. However, the presence of a ligand would potentially modify these calculated charges, but this could not be taken into account since the method aimed at predicting the ligand. In addition, other labels representing

chemical properties such as amino-acid type, hydrogen donor or acceptor, or hydrophobic atom could in principle be included, but was not considered in this study.

We considered three benchmark datasets. The first one, proposed by (156) and referred to as the *Kahraman dataset*, comprises 100 protein crystal structures in complex with one of ten ligands (AMP, ATP, PO4, GLC, FAD, HEM, FMN, EST, AND, NAD). The second one is an extended version of the Kahraman dataset (called *extended Kahraman Dataset* below), in which we added protein structures in complex with one of the same ten ligands, leading to a total of 972 crystal structures (see Additional file 4.4). The added proteins presented pairwise sequence identities less or equal to 30%, to avoid potential bias by inclusion of close homologues.

ligand	atoms count	molecular weight	hydrogen-bond acceptors	hydrogen-bond donors	rotatable bonds
AMP	23	345.21	9	4	4
ATP	31	503.15	13	4	8
PO4	5	95.98	3	1	0
GLC	12	180.16	6	5	1
FAD	53	785.55	15	10	13
HEM	43	616.49	4	2	8
FMN	31	456.34	8	6	7
EST	20	272.38	1	2	0
AND	21	288.42	2	1	0
NAD	44	663.43	14	9	11
Average	**28.3±15.0**	**420.7± 222.8**	**7.5±5.1**	**4.4±3.2**	**5.2±4.9**

Table 4.1: Ligands descriptors for the Kahraman dataset. AMP: adenosine monophosphate, ATP: adenosine-5'-triphosphate FAD, flavin-adenine dinucleotide, FMN: flavin mononucleotide, GLC: alpha-D-glucose, HEM: protoporphyrin containing Fe, NAD: nicotinamide-adenine-dinucleotide, PO4: phosphate ion, AND: 3-beta-hydroxy-5-androsten-17-one, EST: estradiol.

The Kahraman dataset comprises ligands of very different sizes and chemical natures, as shown in Table 4.1. However, the real challenge is to test methods on pockets that bind ligands of similar sizes. Therefore, we created a third dataset comprising 100 structures of proteins in complex with ten ligands of similar size (ten pockets per ligand), see Table 4.2.

ligand	atom count	molecular weight	hydrogen-bond acceptors	hydrogen-bond donors	rotatable bonds
PMP	16	247.17	4	4	4
SUC	23	342.3	11	8	5
LLP	24	361.33	5	6	11
LDA	16	229.4	1	0	11
BOG	20	292.37	6	4	9
PLM	18	255.42	2	0	14
SAM	27	399.45	8	7	7
U5P	21	322.17	8	3	4
GSH	20	306.32	6	6	11
1PE	14	208.25	5	1	11
Average	**19.9± 4.0**	**296.4±61.5**	**5.6±3.0**	**3.9±2.9**	**8.7±3.5**

Table 4.2: Ligands descriptors for the homogeneous dataset PMP: 4'-deoxy-4'-aminopyridoxal-5'-phosphate, SUC: sucrose, LLP: 2-lysine(3-hydroxy-2-methyl-5-phosphonooxymethyl- pyridin-4-ylmethane), LDA: lauryl dimethylamine-N-oxide, BOG: b-octylglucoside, PLM: palmitic acid, SAM: S-adenosylmethionine, U5P: uridine-5'-monophosphate, GSH: glutathione, 1PE: pentaethylene glycol.

When comparing the standard deviation of the mean values of the two tables (4.1. and 4.2), one can see that the molecules of the second set of data values are more homogeneous, mainly as regards the number of heteroatoms and the molecular weight.

This dataset will be referred to as the *Homogeneous Dataset* (HD) (see Additional file 4.4).

The results presented below on this dataset may constitute a new benchmark for future work in the same area.

4.3 Results

All methods were tested on three datasets described in the Data section. The performance of all methods is evaluated on the basis of the AUC score and of the classification error (see section 4.3.2, Performance criteria). The *sup-CK* method is compared to *sup-PI*, *SHD*, *Vol*, *Princ-Axis* and *MultiBind* algorithms (see section 4.2.2, Related methods). Among the pocket extraction methods used in the *SHD* approach, we considered the results corresponding to the Interact Cleft Model (156)), which is similar to our pocket extraction method, and allows to compare the *sup-CK* and *SHD* approaches.

Algorithms, benchmark datasets and distance matrices for the SupCK method are available at http://cbio.ensmp.fr/paris/.

4.3.1 Kahraman Dataset

Results of all methods on the Kahraman Dataset are presented in Table 4.2. According to the AUC score, all methods improve the baseline value of 0.5 corresponding to a random ranking, and simple methods like *Vol* and *Princ-Axis* give surprisingly good results. For example, there is no significant difference between the AUC score of *Vol* and the AUC score of the best performing method *sup-CK$_L$-Vol*. The same effect was observed by (156) when they used simple measures based on comparison of pockets sizes on this benchmark.

As expected, the score obtained using the sequence alignment is close to the baseline value, indicating that this approach is not suitable to the problem of predicting ligand when sequences are very different.

The AUC scores of *sup-CK-Vol* (with or without partial charges) are better than those of all other methods, except for *Vol*, according to the Wilcoxon signed-rank test, involving comparisons of differences between measurements (see Figure 4.3a). The best results are obtained by the *sup-CK-Vol* algorithm, which seems to benefit from the association of volume information

Method	AUC	CE
sup-CK	0.858±0.14	0.36
sup-CK$_L$	0.861±0.13	0.27
sup-CK-Vol	0.889±0.14	0.34
sup-CK$_L$-Vol	0.895±0.12	0.26
Vol	0.875±0.14	0.39
Princ-Axis	0.853±0.13	0.35
sup-PI	0.815±0.13	0.42
SHD	0.770	0.39
MultiBind	0.715 ±0.17	0.42
Sequence	0.55±0.08	0.8

Figure 4.2: Performance on the Kahraman benchmark Performance for each algorithm is evaluated by its mean AUC score and by its classification error (CE), averaged over all pockets. (AUC score for SHD are taken directly from (156), CE scores are estimated from data provided by authors

and of more subtle geometric details provided by the *sup-CK* algorithm. Another observation, is that information on atom partial charges does not significantly improve the AUC score of the *sup-CK* methods.

To evaluate the classification error, we tried to predict a ligand (a class) for each pocket using the k-nearest neighbors classifier (see section Performance criteria). Note that in a ten class (10 ligands) classification problem, a random classifier would have an error of 0.9, which represents baseline performance for all classifiers (the smaller the error, the better the classification).

Table 4.2 shows that methods with higher AUC scores tend to have smaller classification errors, but this correlation is not strict. For example, the *SHD* and *Vol* methods have the same classification error, although the latter displayed a better AUC score than the former. Conversely, the *sup-CK* and *sup-CK$_L$-Vol* methods had similar AUC scores, but the latter performs much better than the former in terms of classification error. This indicates that the AUC score is not appropriate to compare the quality of similarity measures with respect to the problem of ligand identification, and underlines the interest of the classification approach.

The *sup-CK* and *sup-CK-Vol* algorithms have lower classification errors than other methods, which means that they are well suited to the problem of ligand prediction. Interestingly, atom partial charges information significantly reduces classification errors of both methods,

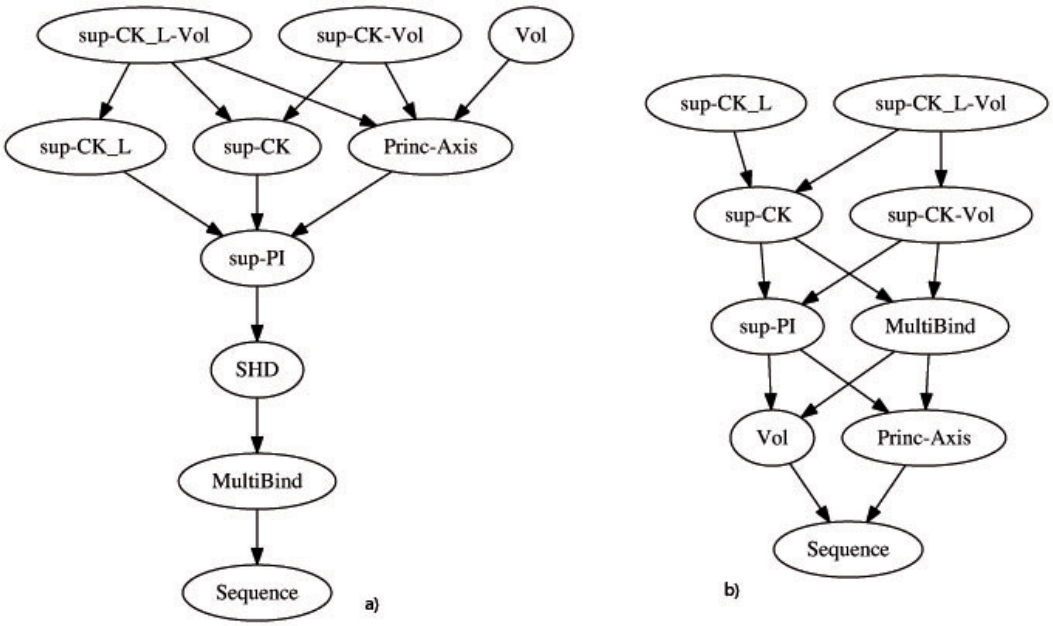


Figure 4.3: Relationship between AUC performances of the methods tested. (a) On the Kahraman dataset (b) On the Homogenous dataset. Each node corresponds to a particular method, parent nodes perform significantly better than child nodes according to the Wilcoxon signed-rank test.

which was not the case for AUC scores. The use of additional atom labels such as amino-acid type, hydrogen donor or acceptor, or hydrophobic atom may again improve the quality of ligand prediction.

No method reaches the AUC score of 1.0, corresponding to perfectly predicts the ligands. Several remarks might explain this fact. First, pockets have to be extracted from the protein structure. Whatever the employed method might be, it is difficult to extract all atoms interacting with the ligand, and only these atoms. In particular, atoms that do not interact with the ligand might have been included in the pockets, which could reduce the observed similarity between pockets that bind this ligand.

Second, ligands are flexible molecules that can adopt different conformations. Therefore, protein pockets that bind the same ligand may display various shapes. In such situations, correct prediction is still possible if the learning dataset contains pockets in which the ligand conformations correctly samples its accessible conformational space. The present dataset contains only 10 pockets per ligand, which might be too small for the most flexible ligands.

When analyzing results in Table 4.2, one must remember that the *Vol* and *Princ-Axis* methods do not require pockets superposition, while all other methods do. The superposition algorithms in these other methods are different, and the way pockets are superposed as an impact on the observed scores. However, the *sup-PI* and *sup-CK* methods only differ by their similarity measures. After superposition, *sup-PI* requires to determine the number of overlapping atoms, while *sup-CK* relies on a weighted number of atoms having close positions. This seems to confer some smoothness properties to the latter, and robustness with respect to variations observed between pockets binding the same ligand leading to better performances of the supCK algorithm.

An important point mentioned in Background is that pocket superposition with *sup-CK* does not require any sequence or structure similarities between the corresponding proteins. To illustrate this property, we analyzed in more details the results for ATP-binding proteins of this dataset. For example, the biotin carboxylase from *E. coli* (452 residues in PDB:1DV2), and the phosphoinositide 3-kinase (961 residues in PDB:1E8X) are unrelated proteins. They present no sequence similarity (they cannot be aligned), and their overall structures are totally different, as shown in Figure 4.4A. However, they bind ATP in similar conformations. When these two pockets are aligned with the *sup-CK* algorithm, their corresponding ATP molecules are found correctly superposed, as shown in Figure 4.4B, although the *sup-CK* algorithm only uses protein atoms.

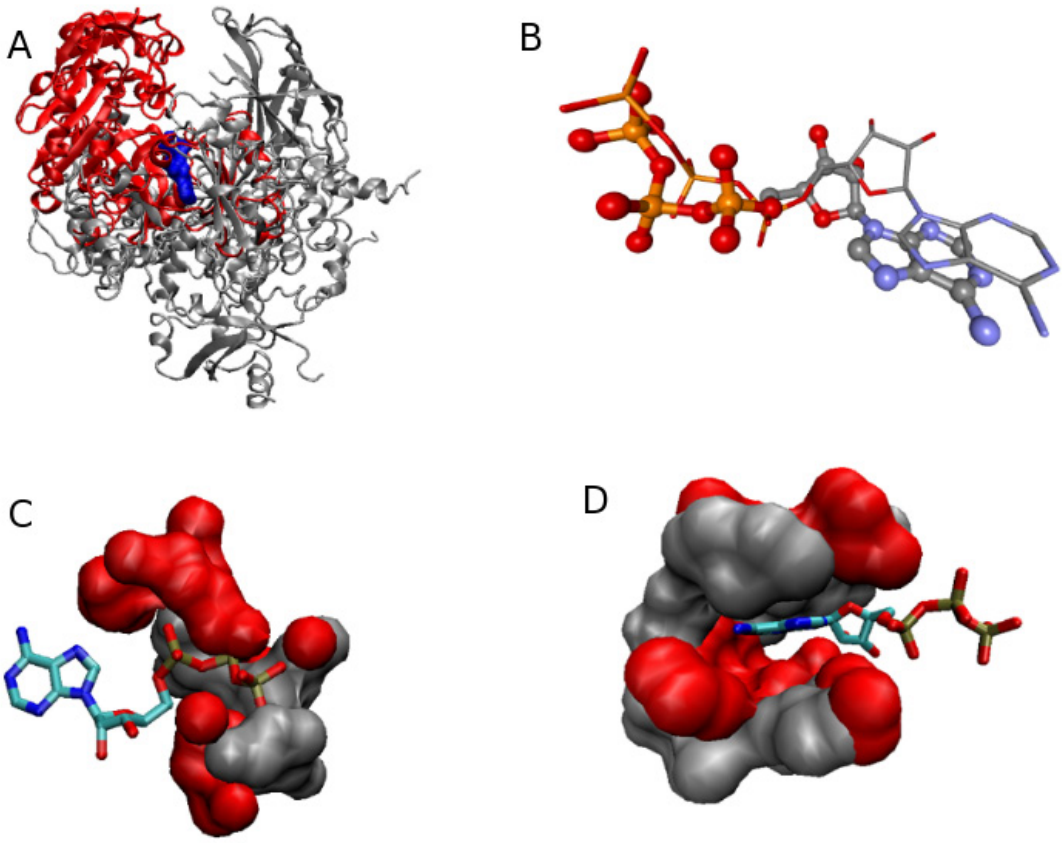


Figure 4.4: Superposition of the binding pockets of two structurally different proteins binding ATP. A) overall structures of pdb:PDB:1E8X in grey and PDB:1DV2 in red superposed according to their binding sites using Sup-CK. ATP molecules are represented in blue. B) Superposition of the ATP molecules from PDB:1DV2 and PDB:1E8X when their binding sites are superposed. C) Positively charged protein regions around ATP molecules of PDB:1E8X in grey and PDB:1DV2 in red. D) Protein hydrophobic patches around ATP molecules of PDB:1E8X in grey and PDB:1DV2 in red.

Moreover, similar residues, playing equivalent roles in ATP binding, are found in equivalent positions in the superposed structures. In particular, N951 and K807 interact with the γ phosphate of ATP in PDB:1E8X and are found close respectively to K288 and H236 that play the same role in PDB:1DV2. We also observe that, K833 interacting with the β and α phosphates of ATP in PDB:1E8X, is found close to K116 in PDB:1DV2 after pockets superposition. These residues form equivalent positively charged regions, as shown in Figure 4.4C. Similarly, the hydrophobic region interacting with the adenine ring of ATP in PDB:1E8X and involving residues W812, I831, I879, I881, V882, A885, M953, F961, and I963 is equivalent to the hydrophobic region involving residues V131, V156, I157, L204, L278, I287, I437 in the superposed PDB:1DV2 structure. These hydrophobic patches overlap after pockets superposition, as shown in Figure 4.4D. Overall, these observations indicate that the *sup-CK* algorithm proposed a relevant superposition for these two unrelated ATP-binding pockets.

Figure 4.7 shows the alignment of the two pockets, extracted from PDB:1E8X and PDB:1DV2 as clouds of atoms, and superposed by *sup-CK*. Note, that *sup-CK* did not try to superpose individual atoms, but rather superposes atom sets.

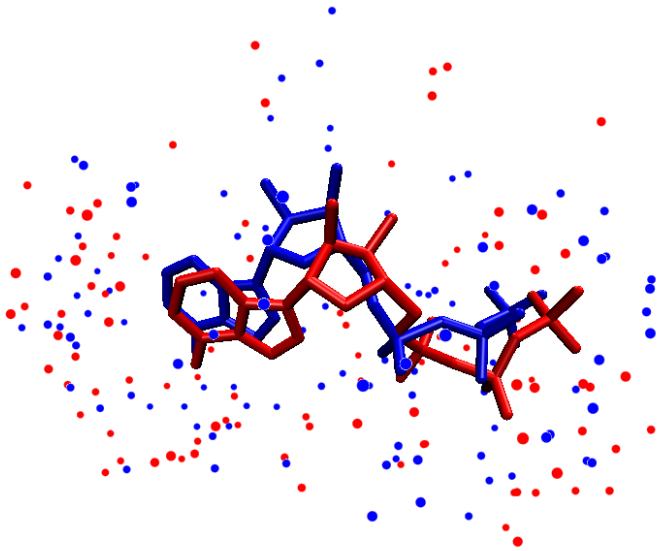


Figure 4.5: Alignment two ATP binding pockets. Alignment of two ATP pockets made by sup-CK, atoms of each pockets are represented by blue and red points, two ATP ligands are traced in licorice.

Extension of Kahraman dataset.

To evaluate the ability of the *sup-CK* method to improve its performance when trained on a larger dataset, we considered an extension of Kahraman dataset consisting of 972 of non redundant pockets that bind one of the 10 ligands of the original dataset (see Data). Therefore, the new dataset consists of 100 Kahraman pockets and 872 new pockets from the PDB.

Table 4.6 presents the classification errors observed on this dataset for different algorithms. Note that in the case of the *sup-CK* methods, the parameters were optimized on the original Kahraman dataset of 100 proteins. Column A presents the classification errors when all 972 pockets are used in the leave-one-out procedure. It shows that all methods improve when the dataset is larger. However, *sup-CK*$_L$ provides the best performance. The quality of its predictions might again improve by including more structures available at the PDB. Column B presents the results on the 100 original pockets extracted from those presented in column A. It shows that 79% of the binding pockets of the original Kahraman dataset were correctly classified by *sup-CK*$_L$, compared to 73% when they were classified using only the original dataset (a classification error of 0.27 in Table 4.2). This shows that when the learning dataset increases, the *sup-CK*$_L$ method is able to learn more and to make better predictions. Finally, column C shows the prediction errors for the 872 new pockets when the 100 original pockets are not used in the leave one out procedure. The scores in this column may be seen as a test on an external independent dataset (as mentioned above, the optimal parameters σ and λ used here were those learned only on the 100 original pockets). It shows that the performance of the *sup-CK* methods does not degrade on the 872 new pockets, and remains above those of the other methods.

It is also interesting to study the structure of the dataset according to the metric associated to the *sup-CK* method. We performed principal component analysis (170) on the pockets similarity matrix of the *sup-CK* method. Figure 4.7 represents the projection of 972 binding pockets on the first two principal components.

Overall, we observe a clustering of binding pockets according to their ligands, which illustrates the good performance of this method for ligand prediction. Looking into more details, we notice that the clusters of pockets that bind ATP, AMP or PO4 overlap. Indeed, proteins that bind ATP usually also bind AMP or PO4, although with different affinities. Furthermore, some pockets (for example pockets that bind GLC or FAD) are found far from their main cluster, or form secondary clusters, which illustrates that pockets having different geometrical characteristics may bind the same ligand. In the classification approach employed here, prediction

Method	A	B	C
sup-CK$_L$	0.19	0.21	0.18
sup-CK$_L$-Vol	0.18	0.19	0.18
Vol	0.32	0.39	0.31
Princ-Axis	0.22	0.27	0.21
sup-PI	0.24	0.33	0.23

Figure 4.6: Classification error on the extended Kahraman benchmark Classification error for all algorithms on the extended Kahraman dataset. Column A - classification error evaluated on all 972 pockets. Column B - Proportion of wrong predictions among the original 100 Kahraman pockets extracted from column A, i.e. classification error evaluated on 100 Kahraman pockets when all 972 pockets are used in the leave-one-out procedure. Column C - classification error evaluated on the 872 new pockets, when the 100 Kahraman pockets are not used in the leave-one-out procedure.

of a ligand for a given pocket uses the classes of its neighbors, which allows to better predict ligands for pockets belonging to such secondary clusters.

4.3.2 Homogeneous dataset (HD)

The Kahraman dataset contains ligands of very different sizes, which might not be typical of real problems. Therefore, we built the Homogeneous dataset because it was important to test methods on a benchmark containing pockets binding ligands of more similar sizes.

Table 4.8 shows that the performances of all algorithms are lower than on the Kahraman dataset, which illustrates that the Homogeneous dataset is a more difficult benchmark. *Vol* and *Princ-Axis* display stronger degradation of performances, with AUC scores of 0.65, and classification errors of 89% and 71%, respectively. The latter must be compared to the baseline value of 90% error for a random classifier for ten classes (ten ligands). This illustrates that size information is less discriminative on this dataset, as expected. All other methods display a stronger improvement with respect to the baseline. Interestingly, although the AUC scores of the simple *Vol* and *Princ-Axis* methods are only 5 to 10% lower than those of all other methods, their classification error is much worse, and *Vol* does not behave better than a random classifier. This again underlines the interest of the classification error score to compare the performances of similarity measures for ligand prediction.

The best AUC score is obtained by the *sup-CK$_L$-Vol* algorithm. The AUC scores of all

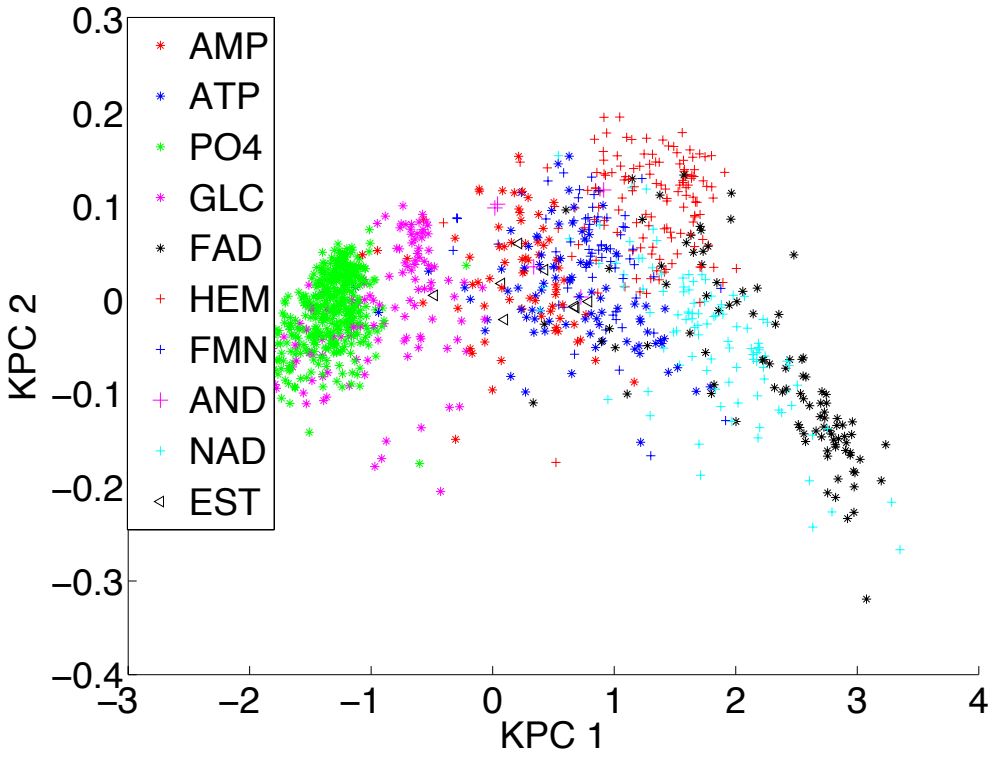


Figure 4.7: Projection of the ext-KD dataset on the first two kernel principal components defined by the similarity measure sup-CK Clustering of binding pockets according to their ligands, which illustrates the performance of this method for ligand prediction.

Method	AUC	CE
sup-CK	0.710±0.19	0.47
sup-CK$_L$	0.752±0.16	0.38
sup-CK-Vol	0.722±0.18	0.46
sup-CK$_L$-Vol	0.766±0.17	0.38
Vol	0.648±0.15	0.89
Princ-Axis	0.650±0.18	0.71
sup-PI	0.702±0.19	0.47
MultiBind	0.69± 0.14	0.48
Sequence	0.577±0.09	0.83

Figure 4.8: Performance on the HD benchmark Performance for each algorithm is evaluated by its mean AUC score and by its classification error (CE), averaged over all pockets.

other methods are significantly lower according to the Wilcoxon signed-rank test (see Figure 4.3b), except *sup-CK$_L$*. Indeed, volume information only provides a slight improvement of 1%, compared to 3% on the Kahraman dataset. On the contrary, information on partial charges leads to an improvement of 4% for the *sup-CK* and *sup-CK-Vol* algorithms, which was not observed on the Kahraman dataset. This shows that addition of physico-chemical information is critical to better compare pockets of similar sizes. The lowest classification errors are obtained by the *sup-CK$_L$* and *sup-CK$_L$-Vol* algorithms, which again shows that volume information is not critical on this benchmark. On the contrary, partial charge information leads to an improvement of 9% between *sup-CK* and *sup-CK$_L$*, and of 8% between *sup-CK-Vol* and *sup-CK$_L$-Vol*.

4.4 Discussion

Choice of optimal parameters. An important characteristic of the *sup-CK* algorithm is its ability to adapt to the variability potentially observed between pockets binding the same ligand. The *sup-CK* algorithm presents two parameters, σ and λ. Parameter σ controls the sensitivity of the similarity measure to atoms relative displacements. The larger the variability of pockets binding the same ligand, the greater the value of σ should be.

Figure 4.9a shows how the mean (over all pockets) AUC score and classification error vary with σ on the Homogeneous dataset. In both cases, the optimum is reached when σ is equal to 1. Note that we did not use the same value of σ estimated from all pockets. For each pocket,

the optimal value was estimated on the basis of the remaining 99 pockets used for training, in a double cross validation scheme, to avoid overfitting to the data. However, we observed that, in most cases (90%), $\sigma = 1$ was chosen. When information on atom partial charges is used, parameter λ (4.4) conditions the sensitivity of the method to atoms charges. Figures 4.9b and 4.9c present the mean AUC score and the classification error as functions of σ and λ. We observe that for the AUC score, the optimum is reached when σ equals 2 and λ equals 0.25, while for the classification error the optimal value of σ is equal to 4.

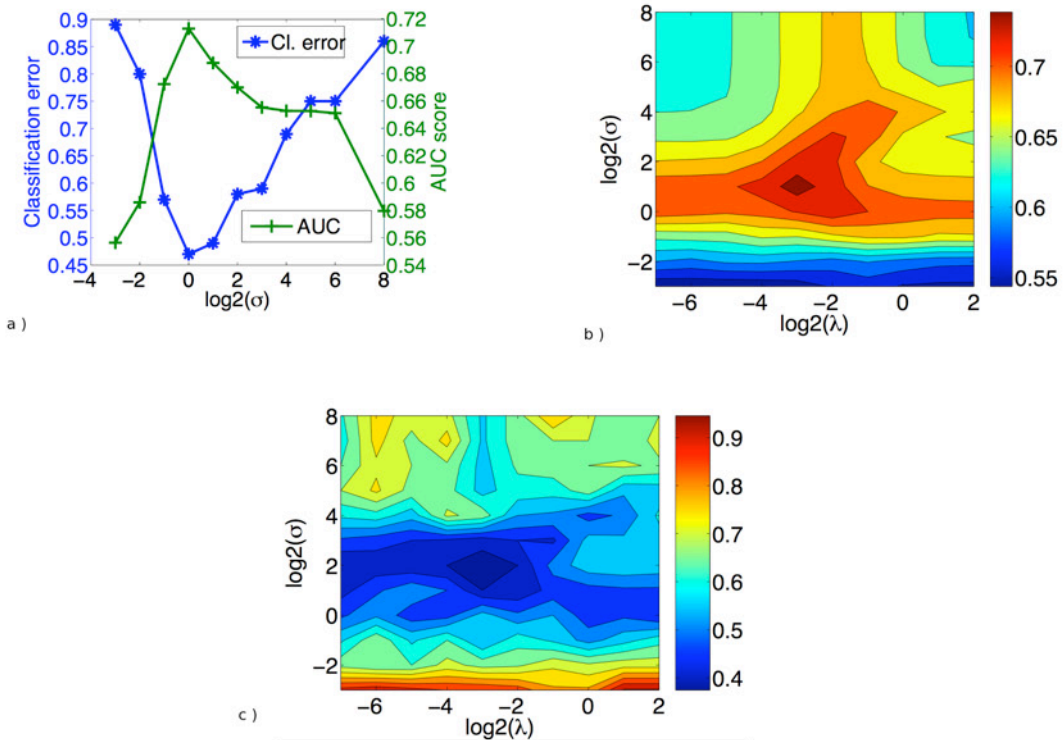


Figure 4.9: Performance on the HD dataset. (a) Mean AUC score and prediction error as functions of σ in the sup-CK method (pure geometrical version, $\lambda = \infty$), (b) mean AUC score and (c) classification error as functions of σ and λ when information on atoms partial charges is used.

While in general we suggest to learn these two parameters of the *sup-CK* algorithm on the dataset of interest, we observed that some default values provide good performance in many cases, and that they could be used in dry-runs on new datasets. For example, a good default value for σ is 1. This value was optimal for the HD dataset when we used the pure geometrical

approach, and it also gave good results on the Kahraman and extended Kahraman datasets. When partial charges are used, i.e. with the *sup-CK*$_L$ algorithm, larger default values for σ are recommended (between 2 and 4), and a good default value for λ is around 0.25.

The radius R of the extracted pocket is a parameter of the extraction pocket procedure. Figures 4.10a and 4.10b present the classification errors of *sup-CK* as a function of σ and R, respectively for the Kahraman and the HD datasets. We observe that in both cases, the optimal value of R is around 5.3 Å, which corresponds to a good default value. However, Figures 4.10a and 4.10b show that the performance of the method is still interesting for values varying between 4.5 and 8 Å. Importantly, they also show that the optimal value of σ does not depend on R. Finally, K is a parameter of the K nearest neighbors classifier (KNN classifier). Ideally, it should also be learned, but values of K between 3 and 5 usually work well.

Robustness of the method with respect to pockets definition. It is important to discuss the impact of using the R parameter, a cutoff distance used for pocket definition. This could lead to situations where an atom is excluded from the pocket in one protein, when a similar atom is included in the pocket of another protein to which it is compared. However, as briefly mentioned in the background section, the principle of the method is to compare pockets based on the optimal superposition of their clouds of atoms. The method does not define or use pairwise matching of atoms of the two pockets, as most other available methods do. Figure 4.7 illustrates this point: the method did not lead to local pairwise superposition of blue and red points, but rather proposed a global superposition of the red and blue atoms densities. Therefore, the method is expected to be robust with respect to potential inclusion or exclusion of a small number of atoms in one of the pockets. As mentioned in the above paragraph, the fact that the performance of the method remains interesting when R varies between 4.5 and 8 Åis also an indirect illustration of this idea. One could wonder if the use of atom labels such as partial charges would decrease the robustness of the method with respect to pockets definition using R. Indeed, a cutoff distance could split a strong dipole in one of the proteins, and not in the other (for example an N-H group). However, the addition of atom labels like partial charges is only one option of the method. Results using only atom positions (corresponding to a pure geometrical approach) already show good performances. Addition of partial charges labels still improves the results, despite the risk that strong dipoles might have been cut. This can probably be explained by the facts that such events are rare, and that the method searches an overall best superposition of atoms densities, despite possible local mismatches in atoms positions or

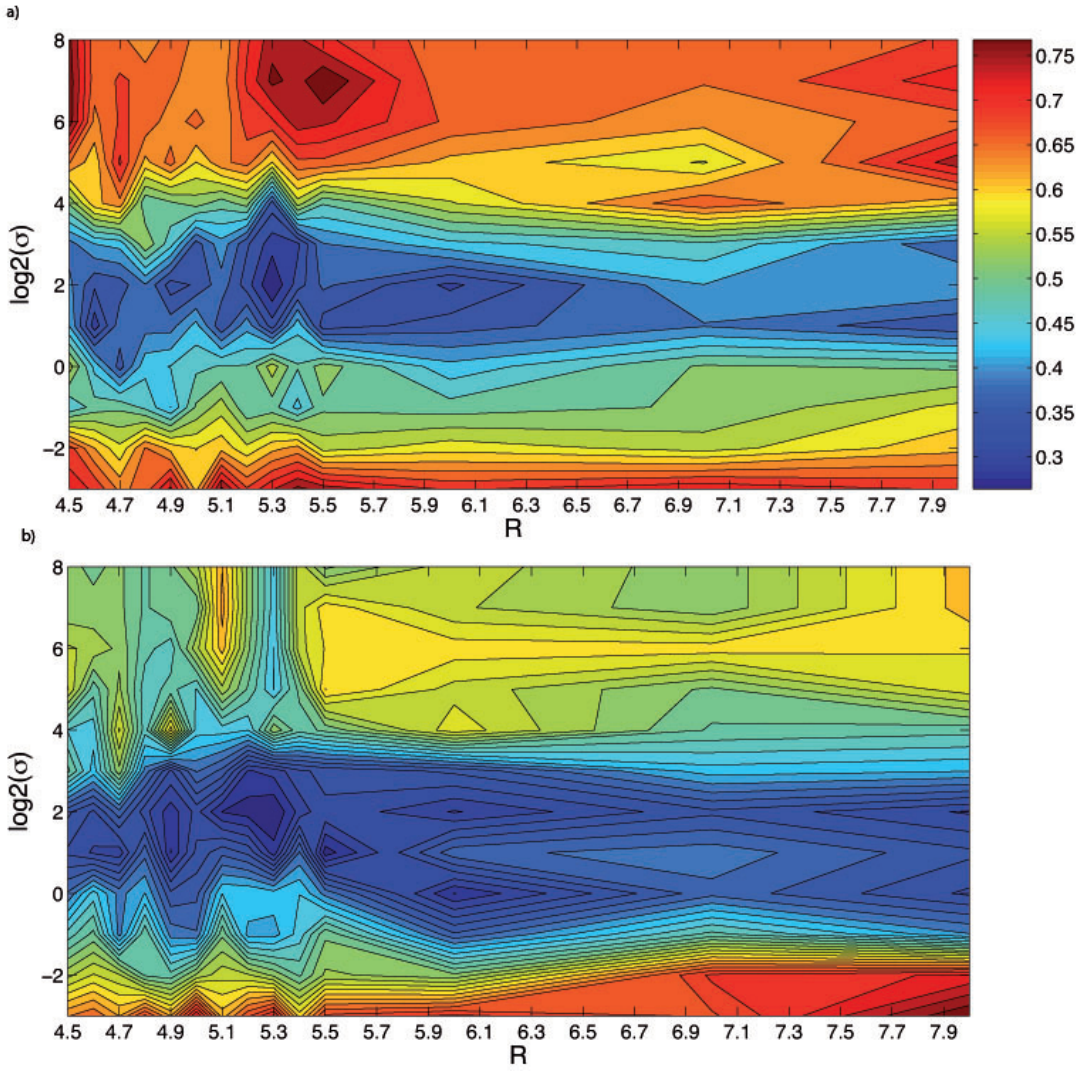


Figure 4.10: Classification error of the sup-CKL algorithm as a function of R and σ ($\lambda = 0.25$) (a) Kahraman dataset, (b) HD dataset.

labels. Nevertheless, it would be interesting to explore other cutoff criteria taking atom labels into account (including other types of labels such as hydrogen bond acceptor, donor, ...), in future developments of the method.

Pocket extraction. We did not tackle the problem of pocket detection, which relies on totally different algorithms than those discussed here, and which was out of the scope of this work. However, the similarity measured between two pockets strongly depends on pocket definition. We extracted pockets as the set of all protein atoms within about 6Åof the bound ligand. Similar approaches were used by (156) (Interacted Cleft Model), and similar pockets may also be retrieved by methods like *Q-SiteFinder* (171) without any information on ligand coordinates. Another alternative could be to employ one the the various programs that have been developed to locate depressions on protein surfaces, particularly in the case where no holo structure is available (172), or in the case of orphan proteins for which the ligand and the binding site is unknown. However, existing pocket extraction algorithms have difficulty to define the rim of a binding pocket, and tend to extract protein cavities that are larger than the binding pocket itself, as defined by the ensemble of residues involved in ligand binding. Although we observed that our method had some robustness with respect to the definition of the binding pocket, global similarity measures like those proposed here may loose some performance on automatically extracted pockets.

Protein functions. The problem of ligand prediction for proteins is related to the problem of predicting the protein molecular function. We analyzed the repartition of the ATP binding pockets generated by the *sup-CK* similarity measure on the extended Kahraman dataset. Figure 4.11 presents the projection of ATP pockets annotated as transferases or ligases, on the first two principal components of the similarity matrix associated to *sup-CK*. We observed that these two families of enzymes are essentially separated. Although these are very preliminary results, they show that *sup-CK* method may be a useful tool, in conjunction with other approaches, for the prediction of protein molecular functions.

In the Result Section, we showed the example of the PDB:1E8X and PDB:1DV2 unrelated structures, binding ATP in similar conformations, and whose pockets were correctly superposed by the *sup-CK* method. In the case of even more dissimilar pockets, binding ATP in different conformations, *sup-CK* still allows superposition of the pockets so that similar regions overlap. For example, biotin carboxylase (452 residues in PDB:1DV2) and phosphoinositide

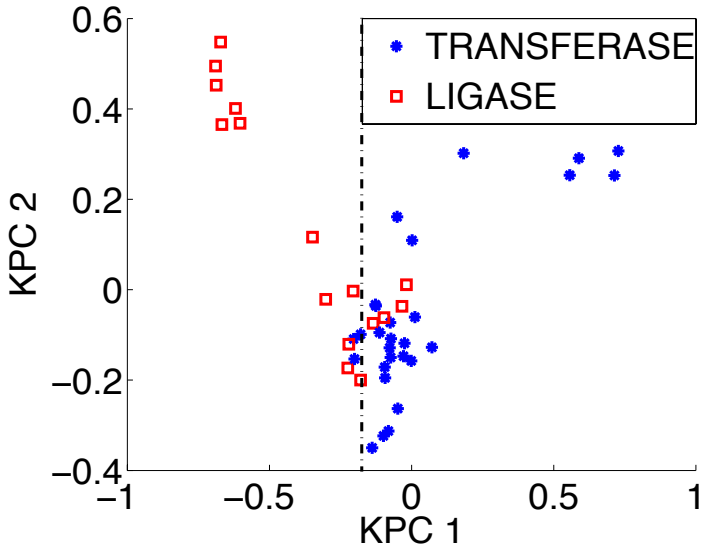


Figure 4.11: Projection of ATP binding pockets on the two first kernel principal components of sup-CK. Repartition of the ATP binding pockets generated by the sup-CK similarity measure on the extended Kahraman dataset. Red squares represent ligases, blue stars represent transferases.

3-kinase (961 residues in PDB:1E8X) of the Kahraman dataset have no sequence or structures homologies, and bind ATP in different conformations as shown in Figure 4.12A. Indeed, 1DV2 is mainly constituted of α-helices whereas 1DV2 is mainly constituted of β-strands. However, according to the *sup-CK* superposition of these two pockets, shown in Figure 4.12B, the two ATP binding sites and the two ATP molecules are found to overlap. Note that these two pockets where correctly classified by *sup-CK* (an ATP ligand was correctly predicted), on the basis of other similar pockets present in the dataset. The SupCK methods proposed a relevant pocket superposition for these highly different proteins with significant pockets deviations, since regions of these two pockets with similar physicochemical properties are found globally superposed.

Apo structures. The *sup-CK* algorithm had a good performance in ligand prediction for holo structures. It also showed robustness with respect to atom displacements. This is an important characteristic for future application of the method to real case studies where the ligand is unknown, and one must extract pockets from apo structures. Local structural rearrangements are frequent upon ligand binding, and methods displaying some smoothness with respect to atoms positions are required when working with apo structures. This would also be necessary for

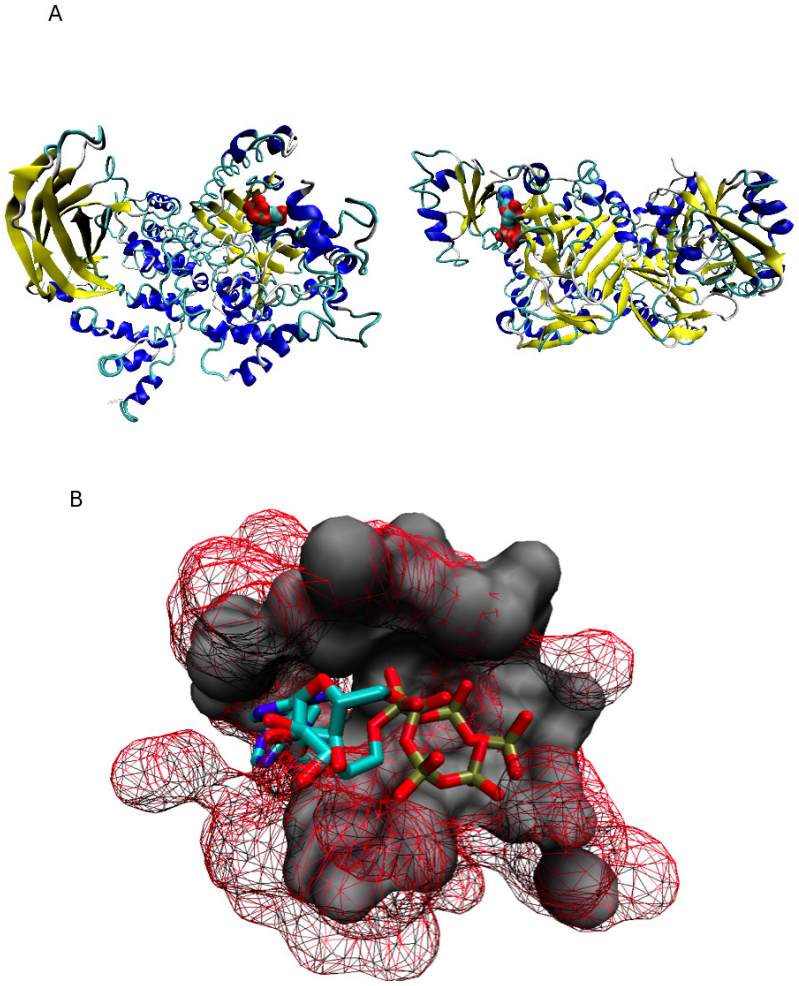


Figure 4.12: Superposition of the binding pockets of two structurally different proteins binding ATP. (A) Left: Proteins structures of phosphoinositide 3-kinase (PDB:1E8X), right: Proteins structures of biotin carboxylase (PDB:1DV2). The two ATP molecules are represented in red. (B) Superposition of the two ATP binding sites are found to overlap.

proteins with no available experimental structure but for which a homology model can be constructed, since the modeled pocket may somewhat differ from the true, but unknown, pocket. We expected that, for large flexible ligands, the performance of the *sup-CK* method might decrease, but this was not observed for the two datasets that we used (Kahraman dataset and Homogeneous dataset). However, we cannot rule out the possibility that this could be observed if the method is trained on other small training datasets.

Computational issues. The running time of the *sup-CK* method depends on the value of the stopping criterion used in the gradient ascent method, and on the number of atoms. In our experiments, the algorithm running time varied between 0.2 and 1.3 seconds (2.5 GHz CPU) per pockets pair. This running time is already quite reasonable to process large protein databanks. The method is presented on datasets of moderate sizes because our aim was to validate the methodology. However, it can be applied on ligand prediction problems, where the number of pockets (and ligands) included in the learning dataset needs to be larger. For future applications in the domain of screening using all ligands available in the Protein Data Bank, a pre-filtering on the basis of simple pocket descriptors (like volume or size) could further accelerate the *sup-CK* method. Future application of the method proposed could include identification of new ligands for protein pockets according to those known for the most similar pockets. This is of interest in the context of identification of drug precursors or of side effects prediction.

4.5 Conclusion

We have developed a new method to measure the similarity between protein binding sites. In this method, binding pockets are described as clouds of points in the 3D space, each point corresponding to an atom. These points may bare additional labels representing various characteristics such as atom partial charges, atom types, or other atomic features. The proposed method showed good performance in the classification of binding pockets according to their respective ligands. It relies on the search for the best global superposition of clouds of atoms, which confers robustness with respect to binding site definition or variations in ligand conformation. This method may be used to compare any type of binding sites in the 3D space, even in absence of overall sequence or structure similarity between their corresponding proteins.

4.6 additional files

pdb	ligand	pdb	ligand	pdb	ligand	pdb	ligand
12as	AMP	1anf	GLC	1ecy	GLC	1jgl	EST
1a0i	ATP	1ank	AMP	1ee9	NAD	1jhg	PO4
1a49	ATP	1aop	PO4	1efp	FAD	1ji0	ATP
1a6q	PO4	1aqa	HEM	1egy	HEM	1jju	HEM
1amu	AMP	1aqe	HEM	1eje	FMN	1jjv	ATP
1ayl	ATP	1aqz	PO4	1el7	FAD	1jni	HEM
1b8a	ATP	1arz	NAD	1elj	GLC	1jo9	HEM
1b8o	PO4	1as2	PO4	1ep1	FAD	1js1	PO4
1bdg	GLC	1ash	HEM	1etp	HEM	1jsc	FAD
1brw	PO4	1ass	PO4	1eu1	GLC	1jsw	GLC
1c0a	AMP	1atj	HEM	1eu3	PO4	1ju2	FAD
1cq1	GLC	1atn	ATP	1ew6	HEM	1jwb	AMP
1cqj	PO4	1atr	PO4	1ex2	PO4	1jwh	PO4
1cqx	FAD	1avq	PO4	1eyj	AMP	1jxz	PO4
1ct9	AMP	1awk	ATP	1eyv	PO4	1jzn	GLC
1d0c	HEM	1b0b	HEM	1ezv	HEM	1k0g	PO4
1d1q	PO4	1b0u	ATP	1f0i	PO4	1k27	PO4
1d7c	HEM	1b12	PO4	1f0x	FAD	1k28	PO4
1dak	PO4	1b14	NAD	1f0y	NAD	1k39	PO4
1dk0	HEM	1b1y	GLC	1f1g	PO4	1k3i	GLC
1dnl	FMN	1b2y	GLC	1f2f	PO4	1k3s	PO4
1dv2	ATP	1b37	FAD	1f3p	FAD	1k4m	NAD
1dy3	ATP	1b3r	NAD	1f4t	HEM	1k6x	NAD
1e2q	ATP	1b49	PO4	1f8r	FAD	1k7v	GLC
1e3r	AND	1b4s	PO4	1f9d	GLC	1k9s	PO4
1e8g	FAD	1b5t	FAD	1fb8	PO4	1k9y	AMP
1e8x	ATP	1b76	ATP	1fcd	FAD	1kae	NAD
1e9g	PO4	1b7v	HEM	1fft	HEM	1kbi	FMN
1ej2	NAD	1b8u	NAD	1fgj	HEM	1kfr	HEM
1ejd	PO4	1bag	GLC	1fik	PO4	1kj8	ATP
1eqg	HEM	1bbh	HEM	1fk8	NAD	1kmn	ATP
1esq	ATP	1bcf	HEM	1fla	FMN	1ko5	ATP
1euc	PO4	1bcp	ATP	1fmw	ATP	1kol	NAD
1evi	FAD	1bd3	PO4	1foh	FAD	1kp2	ATP
1ew0	HEM	1bdb	NAD	1fpp	PO4	1kp8	ATP
1ew2	PO4	1bf3	FAD	1fs7	HEM	1kpf	AMP
1f5v	FMN	1bff	PO4	1fsw	PO4	1kqf	HEM
1fbt	PO4	1bg9	GLC	1ft9	HEM	1kqn	NAD
1fds	EST	1bih	PO4	1fvi	AMP	1kr7	HEM

Continued on next page

pdb	ligand	pdb	ligand	pdb	ligand	pdb	ligand
1gn8	ATP	1bin	HEM	1fwn	PO4	1krh	FAD
1gwe	HEM	1bjw	PO4	1fxx	PO4	1ktg	AMP
1gyp	PO4	1bpd	PO4	1fyd	AMP	1kus	PO4
1h69	FAD	1bpi	PO4	1g19	PO4	1kv8	PO4
1h6l	PO4	1brl	PO4	1g21	ATP	1kv9	HEM
1hex	NAD	1brr	GLC	1g28	FMN	1kwf	GLC
1ho5	PO4	1bsj	PO4	1g31	PO4	1kxj	PO4
1hsk	FAD	1bvb	HEM	1g5t	ATP	1kyq	NAD
1ib0	NAD	1bvr	NAD	1g63	FMN	1kyv	PO4
1iqc	HEM	1bw9	NAD	1ga2	GLC	1l3p	PO4
1j99	AND	1bwk	FMN	1geg	GLC	1l8o	PO4
1ja1	FMN	1bxi	PO4	1giq	NAD	1lc0	PO4
1jp4	AMP	1bxk	NAD	1gk0	PO4	1lfk	HEM
1jq5	NAD	1bzq	PO4	1glf	PO4	1lhr	ATP
1jqi	FAD	1c0i	FAD	1go7	PO4	1lj8	NAD
1jr8	FAD	1c1l	GLC	1gos	FAD	1llu	NAD
1k1w	GLC	1c1s	PO4	1gpm	PO4	1lm3	HEM
1k87	FAD	1c52	HEM	1gpw	PO4	1lqk	PO4
1kht	AMP	1c53	HEM	1gr0	NAD	1lss	NAD
1kvk	ATP	1c6o	HEM	1gs4	PO4	1lvl	FAD
1l5w	PO4	1c8x	PO4	1gt8	FAD	1lw3	PO4
1l7m	PO4	1c8z	PO4	1gts	AMP	1lw7	NAD
1lby	PO4	1c9k	PO4	1gv4	FAD	1m1f	PO4
1lhu	EST	1cbf	PO4	1gwm	GLC	1m32	PO4
1lyv	PO4	1cbq	PO4	1gww	GLC	1m83	ATP
1mew	NAD	1cc5	HEM	1gzf	NAD	1maa	PO4
1mi3	NAD	1cch	HEM	1h2e	PO4	1mb9	AMP
1mvl	FMN	1cdd	PO4	1h2h	NAD	1mbb	FAD
1naz	HEM	1cdt	PO4	1h3e	ATP	1md9	AMP
1nf5	GLC	1cel	GLC	1h53	PO4	1mec	PO4
1np4	HEM	1cen	GLC	1h54	PO4	1mg2	HEM
1o04	NAD	1cf3	FAD	1h7b	PO4	1mh9	PO4
1o9t	ATP	1cfm	HEM	1h9j	PO4	1miw	ATP
1og3	NAD	1cgn	HEM	1hbg	HEM	1mjh	ATP
1p4c	FMN	1cja	AMP	1hbi	HEM	1mky	PO4
1p4m	FMN	1cl6	HEM	1hdi	AMP	1mmu	GLC
1po5	HEM	1cle	PO4	1he4	FMN	1mo9	FAD
1pox	FAD	1cmb	PO4	1hgy	GLC	1mpu	PO4
1pp9	HEM	1cme	NAD	1hi1	ATP	1mq4	PO4
1qax	NAD	1cpq	HEM	1hkq	PO4	1muu	NAD
1qb8	AMP	1cpt	HEM	1hlb	HEM	1muy	GLC
1qf5	PO4	1crk	PO4	1hn5	ATP	1mx3	NAD

Continued on next page

pdb	ligand	pdb	ligand	pdb	ligand	pdb	ligand
1qhu	HEM	1crx	PO4	1hn9	PO4	1mxd	GLC
1qkt	EST	1cu1	PO4	1ho4	PO4	1mz4	HEM
1qla	HEM	1cwu	NAD	1hq0	PO4	1mzh	PO4
1qpa	HEM	1cy1	PO4	1hq3	PO4	1n40	HEM
1rdq	ATP	1d0i	PO4	1hru	PO4	1n5i	ATP
1rlz	NAD	1d0s	PO4	1hsz	NAD	1n5w	FAD
1s7g	NAD	1d3g	FMN	1hto	AMP	1n97	HEM
1sox	HEM	1d4c	FAD	1hwy	NAD	1nd6	PO4
1t2d	NAD	1d9v	PO4	1i0r	FMN	1nel	PO4
1tb7	AMP	1del	AMP	1i19	FAD	1nep	PO4
1tco	PO4	1dgr	PO4	1i77	HEM	1nfb	NAD
1tid	ATP	1dgs	AMP	1i7l	ATP	1nfp	FMN
1tox	NAD	1dhr	NAD	1i82	GLC	1ng4	FAD
2a5f	NAD	1di0	PO4	1i8t	FAD	1ngk	HEM
2cpo	HEM	1dk4	PO4	1ia1	PO4	1nir	PO4
2gbp	GLC	1dks	PO4	1idr	HEM	1nni	FMN
2npx	NAD	1dld	NAD	1ie7	PO4	1nox	FMN
3grs	FAD	1dli	NAD	1ieq	GLC	1npi	PO4
8gpb	AMP	1dm1	HEM	1igs	PO4	1npl	PO4
117e	PO4	1do8	NAD	1ii0	ATP	1nrh	NAD
1914	PO4	1dor	FMN	1ii7	PO4	1nrw	PO4
19hc	HEM	1dpg	PO4	1imd	PO4	1nsf	ATP
1a2y	PO4	1ds7	FMN	1iqr	FAD	1nsj	PO4
1a40	PO4	1dse	HEM	1is2	FAD	1ntf	HEM
1a47	GLC	1dve	HEM	1itc	GLC	1nvm	NAD
1a65	GLC	1dw0	HEM	1ith	HEM	1nx6	PO4
1a7v	HEM	1dxe	PO4	1iug	PO4	1nxg	NAD
1a8p	FAD	1e04	PO4	1iw0	HEM	1ny5	PO4
1a9x	PO4	1e1q	PO4	1izo	HEM	1o2b	FAD
1a9y	NAD	1e24	ATP	1j09	ATP	1o58	PO4
1ad3	NAD	1e3j	PO4	1j0i	GLC	1o83	PO4
1aer	AMP	1e3z	GLC	1j20	AMP	1o9b	NAD
1af6	GLC	1e4g	ATP	1j70	PO4	1o9x	HEM
1ag1	PO4	1e55	GLC	1j77	HEM	1obb	NAD
1ag9	FMN	1e5d	FMN	1j7k	ATP	1obd	AMP
1aj9	HEM	1e6c	PO4	1j8r	GLC	1ofc	GLC
1aka	PO4	1e9x	HEM	1jcm	PO4	1ogo	GLC
1akd	HEM	1ea0	FMN	1jdc	GLC	1oj6	HEM
1alh	PO4	1ebf	NAD	1jds	PO4	1ojr	PO4
1an5	PO4	1eca	HEM	1jft	PO4	1omo	NAD
1ece	GLC	1ecj	AMP	1jg9	GLC	1ooy	PO4
1or4	HEM	2cz8	FAD	1vqv	PO4	2hia	PO4

Continued on next page

pdb	ligand	pdb	ligand	pdb	ligand	pdb	ligand
1ore	AMP	2czc	NAD	1w9w	GLC	2hld	PO4
1orr	NAD	2d0d	PO4	1wa6	PO4	2hmu	ATP
1oz0	PO4	2d0t	HEM	1wbh	PO4	2hnh	PO4
1ozf	PO4	2d1q	AMP	1wcf	PO4	2ho4	PO4
1p0k	PO4	2d2m	HEM	1wdk	NAD	2hoy	PO4
1p1j	PO4	2d37	FMN	1woq	GLC	2hq9	FAD
1p35	PO4	2d3q	HEM	1wvq	PO4	2hqj	PO4
1p3y	FAD	2d5m	FMN	1ww4	GLC	2hrl	GLC
1p49	PO4	2d8a	NAD	1wxx	PO4	2hry	PO4
1p9l	NAD	2dc6	PO4	1wzc	PO4	2hse	PO4
1pie	PO4	2dcl	AMP	1x01	ATP	2hsh	PO4
1pj5	FAD	2ddo	ATP	1x77	FMN	2hti	FAD
1pjc	NAD	2dfz	GLC	1x86	PO4	2huw	PO4
1pjs	NAD	2dj5	PO4	1xjo	PO4	2hxp	PO4
1pkf	HEM	2dkc	PO4	1y30	FMN	2hy1	PO4
1ps9	FAD	2dql	PO4	1y56	FAD	2hyr	GLC
1pt7	PO4	2dsd	AMP	1y89	PO4	2hzm	PO4
1ptm	PO4	2dv1	HEM	1yb0	PO4	2i02	FMN
1pvw	PO4	2dwj	GLC	1yqz	FAD	2i0z	FAD
1pwb	GLC	2dxq	PO4	1yr9	PO4	2i1o	PO4
1q08	PO4	2e2o	GLC	1yrh	FMN	2i3c	PO4
1q16	HEM	2e5f	PO4	1yrr	PO4	2i51	FMN
1q33	GLC	2e5y	ATP	1yw1	GLC	2i58	GLC
1q3f	PO4	2ead	GLC	1ywf	PO4	2i6j	PO4
1q97	ATP	2efb	HEM	1z0z	NAD	2i7h	FMN
1qf6	AMP	2egk	PO4	1z5g	PO4	2i9a	PO4
1qfc	PO4	2ehh	PO4	1z6a	PO4	2i9p	NAD
1qhb	PO4	2esr	GLC	1z6l	FAD	2iag	HEM
1qhg	ATP	2ets	PO4	1z7m	PO4	2ib5	PO4
1qhx	ATP	2evs	GLC	1z84	AMP	2ibg	PO4
1qlm	PO4	2eww	ATP	1z9n	HEM	2ifa	FMN
1qrr	NAD	2exr	FAD	1zbu	AMP	2ig3	HEM
1qwt	PO4	2ez2	PO4	1zc0	PO4	2ig6	FMN
1qz4	PO4	2f10	PO4	1zcn	PO4	2iiz	HEM
1r0x	ATP	2f17	AMP	1zm1	GLC	2ily	ATP
1r2j	FAD	2f2e	GLC	1zui	PO4	2im8	PO4
1r37	NAD	2f5v	FAD	1zwk	PO4	2imd	PO4
1r5i	PO4	2f6d	PO4	1zwx	PO4	2iml	FMN
1r72	NAD	2f6s	PO4	2a0z	GLC	2in3	PO4
1rer	PO4	2f7m	PO4	2a19	PO4	2inw	PO4
1rfm	NAD	2f7o	PO4	2a3l	PO4	2iof	PO4
1rkd	PO4	2f84	PO4	2a5y	ATP	2ipi	FAD

Continued on next page

pdb	ligand	pdb	ligand	pdb	ligand	pdb	ligand
1rkv	PO4	2f8m	PO4	2a7x	AMP	2irv	PO4
1rkx	NAD	2faq	ATP	2a96	PO4	2isi	PO4
1rli	PO4	2fb1	PO4	2acx	PO4	2isj	FMN
1rlj	FMN	2ffi	PO4	2aep	GLC	2iss	PO4
1rmg	GLC	2fg9	FAD	2aiu	HEM	2isy	PO4
1rn4	PO4	2fh6	GLC	2aml	PO4	2itf	HEM
1rp4	FAD	2fhd	PO4	2an4	PO4	2iuc	PO4
1rv3	PO4	2fjb	AMP	2aqx	ATP	2ivd	FAD
1rw9	PO4	2fkn	NAD	2ark	PO4	2ivf	HEM
1rwj	HEM	2fmy	HEM	2art	AMP	2ivt	AMP
1rxc	PO4	2fn0	PO4	2au5	PO4	2iw0	PO4
1ry2	AMP	2fn6	PO4	2avn	PO4	2ixa	NAD
1ryr	ATP	2for	PO4	2axr	FAD	2ixe	ATP
1rz1	FAD	2fr7	HEM	2b3b	GLC	2iyg	FMN
1s20	NAD	2fre	FMN	2b3d	FAD	2izz	NAD
1s5l	HEM	2fsg	ATP	2b3n	PO4	2j0p	HEM
1s5m	GLC	2fug	FMN	2b44	PO4	2j0x	PO4
1s68	AMP	2fuq	PO4	2b67	FMN	2j1d	PO4
1s96	PO4	2fw5	HEM	2b69	NAD	2j3m	ATP
1sb7	PO4	2fwr	PO4	2b9w	FAD	2j44	GLC
1sb8	NAD	2fyq	PO4	2bhy	GLC	2j6r	PO4
1ses	AMP	2g09	PO4	2bis	GLC	2j84	AMP
1sez	FAD	2g0t	PO4	2bra	FAD	2j9d	AMP
1sf3	PO4	2g1u	AMP	2bs5	GLC	2j9l	ATP
1sfs	PO4	2g25	PO4	2bu2	ATP	2j9r	PO4
1su2	ATP	2g37	FAD	2bvf	FAD	2jae	FAD
1t0i	FMN	2g5c	NAD	2bvl	GLC	2jbh	PO4
1t53	ATP	2g5g	HEM	2bwa	GLC	2jbo	PO4
1t57	FMN	2g8s	PO4	2c0k	HEM	2jbs	FMN
1t5b	FMN	2gag	FAD	2c0u	FAD	2jcb	PO4
1t6y	FMN	2gax	PO4	2c1w	PO4	2je2	PO4
1tg7	PO4	2gbl	ATP	2c30	PO4	2jen	GLC
1to3	PO4	2gd9	PO4	2c38	AMP	2jfr	PO4
1tqa	HEM	2gdv	GLC	2c4n	PO4	2jfu	PO4
1tqn	HEM	2gdz	NAD	2c54	NAD	2jgd	AMP
1tyw	GLC	2gfh	PO4	2c5s	AMP	2nad	NAD
1u2s	GLC	2gj3	FAD	2c6p	PO4	2nn8	GLC
1u9z	AMP	2gjl	FMN	2c91	PO4	2nnc	PO4
1uam	PO4	2gju	PO4	2cap	PO4	2nox	HEM
1udn	PO4	2gk6	PO4	2ccl	PO4	2npi	ATP
1uev	ATP	2gm3	AMP	2cfa	FAD	2ns2	PO4
1uf9	ATP	2gm7	PO4	2cfm	AMP	2ns9	PO4

Continued on next page

pdb	ligand	pdb	ligand	pdb	ligand	pdb	ligand
1ukz	AMP	2gmh	FAD	2cg9	ATP	2nt1	PO4
1ulc	GLC	2gmk	AMP	2ch6	GLC	2nt8	ATP
1um0	FMN	2gpj	FAD	2chp	PO4	2nvu	ATP
1uoz	GLC	2gqf	FAD	2cja	ATP	2nwb	HEM
1usc	FMN	2gru	NAD	2ck3	PO4	2nxf	PO4
1uu0	PO4	2grx	PO4	2cm6	PO4	2nyj	ATP
1uw3	PO4	2gsu	AMP	2cmw	PO4	2nzc	PO4
1uwg	PO4	2gte	PO4	2cn3	GLC	2o08	PO4
1uwv	PO4	2gtl	HEM	2cns	PO4	2o09	HEM
1v04	PO4	2gv8	FAD	2cul	FAD	2o0h	ATP
1v1b	ATP	2gvg	PO4	2cvj	FAD	2o0m	PO4
1v26	AMP	2gvy	GLC	2cx7	PO4	2o16	PO4
1v2b	GLC	2gwl	NAD	2cxn	PO4	2o4c	NAD
1v2i	PO4	2gxq	AMP	2cy3	HEM	2o4v	PO4
1v2x	PO4	2h0u	FMN	2o6p	HEM	2pbl	PO4
1v33	PO4	2h5f	PO4	2o9z	PO4	2pbz	ATP
1v9f	PO4	2h8x	FMN	2oaf	PO4	2pce	PO4
1v9y	HEM	2hae	NAD	2oaq	PO4	2ph5	NAD
1vdr	PO4	2hbl	AMP	2ob1	PO4	2pi8	PO4
1vhn	FMN	2hcr	AMP	2obn	PO4	2pia	FMN
1vj5	PO4	2hdo	PO4	2oej	PO4	2pla	NAD
1vjp	NAD	2hek	PO4	2ofx	PO4	2pmb	PO4
1vkf	PO4	2hfn	FMN	2ogx	ATP	2pnk	PO4
1vkk	PO4	2hhc	PO4	2oh5	ATP	2ppq	PO4
1vl8	PO4	2hhg	PO4	2ohh	FMN	2ppv	PO4
1vlp	PO4	2hhz	PO4	2oiv	PO4	2pq7	PO4
1vlv	PO4	2hi4	HEM	2ojw	PO4	2pqv	PO4
2oys	FMN	2qgz	PO4	2ok7	FAD	2ptf	FMN
2oyy	HEM	2qjc	PO4	2onk	PO4	2ptq	AMP
2ozt	PO4	2uuu	FAD	2oov	PO4	2pup	PO4
2p09	ATP	2uv8	FMN	2oox	AMP	2pv7	NAD
2p0e	PO4	3ck9	GLC	2osx	GLC	2q0v	PO4
2p0f	PO4	3pfk	PO4	2otd	PO4	2q3e	NAD
2p0k	PO4	3sil	PO4	2ou5	FMN	2q7g	ATP
2p6u	PO4	4kbp	PO4	2oun	AMP	2qck	PO4
2p8i	PO4	5cro	PO4	2p9e	PO4	2pb9	PO4

Table 4.3: Pdf file containing a table describing all proteins used in the Homogeneous dataset. (PDB name, EC number, ID Uniprot, protein classification, chain, Ligand)

pdb name	EC number	ID Uniprot	protein classification	protein chain	ligand
1A0G	2.6.1.2	P19938	transferase	A	PMP
1A0T	—	P22340	outer membrane protein	P	SUC
1A8I	2.4.1.1	P00489	glycogen phosphorylase	A	LLP
1AIA	2.6.1.1	P00509	transferase(aminotransferase)	A	PMP
1AIJ	—	P0C0Y9	photosynthetic reaction center	M	LDA
1AR1	1.9.3.1	P01636	complex (oxidoreductase/antibody)	B	LDA
1AUA	—	P24280	phospholipid-binding protein	A	BOG
1AX4	4.1.99.-	P28796	tryptophan biosynthesis	A	LLP
1B4W	3.1.1.4	O42187	hydrolase	A	BOG
1B56	—	Q01469	lipid-binding	A	PLM
1BJW	2.6.1.1	Q56232	aminotransferase	B	LLP
1BW0	2.6.1.5	P33447	transferase	A	LLP
1C8U	3.1.2.-	P0AGG2	hydrolase	A	LDA
1CL1	4.4.1.8	P06721	methionine biosynthesis	A	LLP
1CMC	—	P0A8U6	dna-binding regulatory protein	B	SAM
1CS1	4.2.99.	P00935	lyase	A	LLP
1D7K	4.1.1.1	P11926	lyase	A	LLP
1DBT	4.1.1.2	P25971	lyase	A	U5P
1DUG	2.5.1.1	P08515	blood clotting	B	GSH
1DXR	—	P06010	photosynthetic reaction center	M	LDA
1EEM	—	P78417	transferase	A	GSH
1EH5	3.1.2.2	P45478	hydrolase	A	PLM
1EIZ	2.1.1.-	P0C0R7	transferase	A	SAM
1F7S	—	Q39250	plant protein	A	LDA
1FG7	2.6.1.9	P06986	transferase	A	PMP
1FGX	2.4.1.3	P08037	transferase	B	U5P
1FW1	2.5.1.1	O43708	isomerase/transferase	A	GSH
1FX8	—	P0AER0	membrane protein	A	BOG
1G8I	—	P62166	metal binding protein	A	1PE
1G8O	2.4.1.1	P14769	transferase	A	U5P
1HMY	2.1.1.3	P05102	transferase(methyltransferase)	A	SAM
1I5E	2.4.2.9	P70881	transferase	A	U5P
1I78	3.4.21.	P09169	hydrolase	B	BOG
1I9G	—	O33253	transferase	A	SAM
1IUG	—	Q5SKR1	transferase	A	LLP
1IYH	5.3.99.	O60760	isomerase	A	GSH
1J04	2.6.1.4	P21549	transferase	A	LLP
1JG8	4.1.2.5	Q9X266	lyase	A	LLP
1JGI	2.4.1.4	Q9ZEU2	transferase	A	SUC
1JJ0	3.2.1.1	P00698	hydrolase	A	SUC
1JLV	2.5.1.1	Q7KIF2	transferase	A	GSH

Continued on next page

pdb name	EC number	ID Uniprot	protein classification	protein chain	ligand
1K87	1.5.99.	P09546	oxidoreductase	A	1PE
1K8Q	3.1.1.3	P80035	hydrolase	A	BOG
1KMO	—	P13036	membrane protein	A	LDA
1KTA	2.6.1.4	O15382	transferase	A	PMP
1L0G	3.5.2.6	P00811	hydrolase	A	SUC
1M66	1.1.1.8	P90551	oxidoreductase	A	PLM
1M98	—	P83689	unknown function	A	SUC
1MDO	—	Q8ZNF3	transferase	A	PMP
1MGP	—	Q9X1H9	lipid binding protein	A	PLM
1MSK	2.1.1.1	P13009	methyltransferase	A	SAM
1NT2	—	O28191	rna binding protein	A	SAM
1NW3	—	Q8TEK3	transferase	A	SAM
1O57	—	P37551	dna binding protein	C	1PE
1O6U	—	O76054	transferase	A	PLM
1OJD	1.4.3.4	P27338	oxidoreductase	A	LDA
1P91	2.1.1.5	P36999	transferase	A	SAM
1PQ2	1.14.14	P10632	oxidoreductase	B	PLM
1PT2	2.4.1.1	P05655	transferase	A	SUC
1Q0R	—	Q54528	hydrolase	A	1PE
1QZZ	—	Q54527	transferase	A	SAM
1R30	2.8.1.6	P12996	transferase	A	SAM
1R4W	2.5.1.1	P24473	transferase	A	GSH
1S7G	3.5.1.-	O30124	transcription	A	1PE
1SZ7	—	O43617	transport protein	A	PLM
1THQ	—	P37001	transferase	A	LDA
1TJ4	3.1.3.2	P74325	hydrolase	A	SUC
1UC2	—	O59245	unknown function	A	SUC
1UMX	—	P0C0Y9	photosynthetic reaction center	H	LDA
1UU1	2.6.1.9	Q9X0D0	transferase	A	PMP
1W2T	3.2.1.2	O33833	hydrolase	A	SUC
1WLJ	3.1.-.-	Q96AZ6	hydrolase	A	U5P
1XKW	—	P42512	membrane protein	A	LDA
1Y10	4.6.1.1	Q11055	lyase	B	1PE
1Y1A	—	Q99828	metal binding protein	B	GSH
1YLJ	3.5.2.6	Q9L5C8	hydrolase	A	SUC
1ZC9	4.1.1.6	P16932	lyase	A	PMP
1ZX8	—	Q9X187	unknown function	C	1PE
2B56	—	Q86MV5	transferase/rna binding protein	A	U5P
2BLN	2.1.1.2	P77398	transferase	A	U5P
2BMU	2.7.4.-	Q8U122	transferase	A	U5P
2BYN	—	Q8WSF8	receptor	B	1PE
2C37	3.1.13.	Q9UXC2	hydrolase	A	U5P

Continued on next page

pdb name	EC number	ID Uniprot	protein classification	protein chain	ligand
2C81	2.6.1.-	Q8G8Y2	transferase	A	PMP
2CJG	—	P63509	transferase	A	PMP
2CZV	3.1.26.	O59150	hydrolase	C	BOG
2E7U	5.4.3.8	Q5SJS4	isomerase	A	PMP
2FIK	—	P11609	immune system	A	PLM
2FLS	—	Q9NS18	oxidoreductase	A	GSH
2HAW	3.6.1.1	P37487	hydrolase	B	1PE
2HD0	3.4.22.	Q9ELS8	hydrolase	E	BOG
2IDB	4.1.1.-	P0AAB4	lyase	A	1PE
2IMD	2.5.1.1	Q51948	transferase	A	GSH
2IU8	2.3.1.-	O84245	transferase	B	PLM
2J4J	2.7.4.2	Q97ZE2	transferase	A	U5P
2NWL	—	O59010	transport protein	A	PLM
2P4B	—	P0AFX9	signaling protein	B	BOG
2PBJ	5.3.99.-	Q9N0A4	lyase	A	GSH
2Z73	—	P31356	membrane protein	A	BOG
3B6H	5.3.99.-	Q16647	isomerase	A	BOG

Table 4.4: Pdf file containing a table describing all proteins used in the Homogeneous dataset. (PDB name, EC number, ID Uniprot, protein classification, chain, Ligand)

5

Discussion and Perspectives

5.1 Description of the chemogenomic space

In chemogenomics, compound libraries are combined with protein information, and the ultimate goal is to identify all possible interactions between ligands and proteins of the proteome. However, the size of the protein-ligand space makes any systematic experimental characterization impossible. Indeed, the number of molecules with drug-like molecular weight (up to about 600 Da) is very large. Moreover, the human genome project has identified and characterized more than 25000 genes in the human DNA (173), leading to an even larger number of proteins due to alternative splicing and post-translational modifications. Adding a biological dimension to high-throughput screening means that, even with the impressive technological advances made, current capacities are no longer sufficient to tackle the experimental testing of tens of thousands of compounds against thousands of targets. The chemogenomic matrix is thus very sparse since experimental data, in the form of binding affinity values such as inhibition constants (Ki) and inhibitory concentrations (IC50), is available only for a very limited number of protein-ligand complexes. Chemogenomic approaches therefore tried to fill this matrix by prediction of protein-ligand interactions. However, as presented in chapter 3, these approaches have to face two main problems: how to encode and compare molecules, how to encode and compare proteins.

The protein and ligand spaces have traditionally been studied as separate entities. Since conventional drug discovery is focused on ligand optimization, the chemical space has been a well studied research topic (174).

5.1.1 Description of the chemical space

The chemical space of small molecules has been extensively studied in chemoinfomatic research. As shortly reviewed in chapter 2, large number of ligand descriptors has been used in drug discovery. Ligand descriptors are typically classified by the dimensionality of the representation of the compound (175). One-dimensional (1D) descriptors are computed from the atomic composition of the molecule. They correspond typically to global molecular properties, such as the molecular weight and hydrophobicity, the number of atoms of particular types or hydrogen bond donors and acceptors. 2D descriptors are derived from the graphical representation of a chemical structure, and include 2D binary fingerprints. Finally, three-dimensional (3D) descriptors are generated from 3D representation of the molecules, although 3D approaches suffer from the limitation that the active 3D conformer might not be known. Based on these various descriptions, numerous methods were proposed to quantify molecular similarity. Today, a large panoply of approaches is available to handle the chemical space in chemogenomic approaches.

5.1.2 Description of the biological space

The field of describing the biological space, i.e. encoding and comparing proteins, has been less studied, and constitutes the main bottleneck for development of large-scale chemogenomic approaches. We will shortly review the main strategies that have been developed to encode and compare proteins. Proteins are commonly represented according to their sequence or their 3D structure, which has lead to two main types of approaches: sequence-based and structure-based approaches. The full amino-acid sequence is the most straightforward information, and enables a relevant clustering of proteins into families such as ”GPCRs” or ”kinases”. Similarly, analysis of global 3D topologies of proteins has lead to structural classification of proteins, as illustrated by the SCOP database.

5.1.2.1 Sequence-based approaches

The sequence representation allows comparison of two proteins using sequence alignment algorithms such as BLAST (176). The alignment score can then be used to derive a distance measure between these two proteins. In particular, kernels methods developed for protein sequences (177) could in principle be used in the SVM framework for chemogenomics presented in chapter 3. However, this would require that a sequence alignment is reachable, which is

possible only for proteins belonging to the same family, as already pointed. As a consequence, chemogenomic approaches in which the biological space is encoded on the basis of protein sequences is limited to proteins belonging to the same family. However, to our knowledge, full-length sequence similarity measures have never been implemented in SVM for chemogenomics. Chemogenomic approaches based on protein sequences have compared and classified proteins based on ligand-binding sites by using sequence motifs. For example, in the case of GPCR, a study focussed on residues known from molecular recognition studies such as site-directed mutagenesis to be important for binding of the ligand (178; 179). However, comparison of such binding site sequences was used to derive GPCR classifications, but was not implemented in a chemogenomic approach to predict ligands for GPCRs. As presented in chapter 3, another approach consists in aligning all GPCRs sequences, extracting key residues supposed to map the ligand binding site and concatenating these key residues into an un-gapped sequence which can be later used to derive a distance matrix based on sequence identity (132), sequence similarity (131) or physicochemical properties (180). An exhaustive cavity-based clustering of 372 human GPCRs has been proposed using such a strategy (132). Interestingly, it reproduces the full sequence-based tree suggesting that only a few residues are really important when comparing targets across a family. This simplification enables a much simpler analysis of features (binding site regions), which are responsible for selective or permissive ligand binding by simply looking at residue conservation. Previous studies based on comparison of extracted sub-sequences of residues involved in the ligand-binding site have not been used to predict ligands, and the SVM-based approach presented in chapter 3 is the first published method to achieve this goal. However, the main limitation of the proposed method, common to all sequence-based approaches (whether they are based on the full protein sequence or on a sub-sequence), is that they can only be applied to proteins belonging to the same family.

5.1.2.2 Structure-based approaches

An alternative strategy, which could provide a means to overcome this limitation, is to encode and compare proteins based on their 3D structure, when they are available. For example, one could measure the structural similarity between two proteins based on their overall 3D structure similarity using a structural alignment algorithm such as MAMMOTH (181). Kernel methods have been proposed based on structure similarity according to MAMMOTH (182), and used for the prediction of enzyme functions or SCOP class. In principle, they could also be used in chemogenomics in order to predict ligands. However, as in the case of sequences, approaches

based on comparison of overall 3D structures can only be performed on proteins of similar fold, i.e. proteins belonging to the same structural family.

Another family of structure-based encoding of proteins consists in focussing on the ligand-binding site. One can start from a structural alignment of two proteins, and then describe them by computed molecular interaction fields derived from the cavities. These fields can then be used in vectors that encode for the proteins, allowing comparison of binding sites. Again, this approach can only be applied to proteins of similar structural families. However, its has been successfully applied to protein kinases (183), serine proteases and GPCRs (184), or matrix metallo-proteinases (185). These studies aimed at explaining pockets selectivity, and thus guiding the design of compound libraries towards the desirable selectivity pattern. However, they could be used as such for ligand prediction, using a chemogenomic approach similar to that presented in chapter 3.

Other structure-based approaches are better suited to compare proteins with different overall 3D structures, describing proteins by physicochemical properties. The molecular surface of a binding pocket can be discretized in either chemically labelled points (186) or graphs (187) and then aligned to maximize overlap with any reference. A database of protein surfaces (eF-site) has been browsed to predict the function of a hypothetical archaeon protein (MJ0226) by detection of a mononucleotide binding site (187). Surface-based comparisons are, however, relatively slow and thus incompatible with large-scale comparison of binding pockets. Faster methods have been developed (163; 188; 189; 190; 191). They all have in common to represent an active site of interest by pseudocenters (dummy atoms located close to every side chain of interest) encoding physicochemical properties (H-bonding capacity, aromaticity, hydrophobicity, charge) of their cognate residues, pseudocenters being linked together by edges and thus defining a molecular graph. Alignment is operated for example by detection of maximal common subgraphs (clique detection) (192). A nice example of binding site similarities for distant proteins has been exemplified by Weber et al. (193), who detected cross-reactivity of arylsulfonamide-based COX-2 inhibitors with human carbonic anhydrase (HCA) based on the similarity of COX-2 and HCA binding pockets. A problem with these matching techniques is that the computed similarity score (usually dependent on the number of atom/pseudocenter/triangle matches) is not always easy to interpret, notably for active sites of different dimensions, because large actives sites will have a tendency to present more matches than small ones even if the latter are more similar. Therefore, normalized distance metrics similar to those used for comparing ligands are needed. An alternative approach is to evaluate the

similarity of potential ligand binding envelopes for known X-ray structure of apo or holoproteins (194). A first draft of the human pocketome, a collection of all possible ligand binding envelopes for a set of 943 crystallized human proteins, has been proposed (An et al., 2005) and clustered by envelope similarity. Interestingly, the ligand envelope-based tree only partially matches alternative trees based on the amino-acid sequence of the target proteins or on bound-ligand similarities (194). However, note that all these structure-based methods have been used to compare proteins, cluster and classify them, or detect potential ligand cross-reactivity, but they have not been used to make large-scale ligand prediction, as chemogenomics methods intend to perform.

5.2 Extension of the proposed methods to SVM-based Chemogenomics methods

As mentioned above, we have proposed in chapter 3 a sequence-based chemogenomic method within the SVM framework. It is able to make ligand predictions on large-scale in the chemical space (i.e. predict ligands within large and chemically diverse chemical databanks) for proteins belonging to a given family. The method has been applied to the case of GPCRs, although in principle, it can be applied to other families of proteins such as kinases or proteases. In the above paragraph, we have recalled that structure-based description of proteins at the level of binding pockets can be used to handle the case of proteins belonging different families. Let us discuss how these methods could help to extend the SVM-based chemogenomic approaches for large-scale prediction in the biological space, i.e. predict ligands for large and biologically diverse protein datasets. In chapter 4, we presented a method that encodes proteins by an explicit cloud of points in 3D space, corresponding to the protein atoms belonging to the binding pocket. The method defines a similarity measure to compare any pair of proteins of known 3D structures, even if they present totally different overall 3D structures. It was not applied for large-scale comparison of proteins, because the aim was first to evaluate its performance on a limited benchmark, which however contained very different protein structures. Only ten different ligands where included for the proteins of this benchmark. In a cross validation study, we showed that the similarity measure proposed in this thesis (supCK) presented good performance to predict the ligand of a given pocket, among these ten possible ligands. Let us discuss how this method could be extended in order to be implemented in a SVM-based chemogenomic approach for larger scale predictions.

As we mentioned in chapter 4, the supCK function defines a ”good” similarity measure, but unfortunately, it is not a kernel function. This means that the similarity matrix that is built when comparing an ensemble of proteins is not definite positive, i.e. its eigenvalues are not all positive. This is a typical situation arising when designing a kernel in computational biology, as well as in other fields. For example, let us consider a finite set of proteins, and s a measure of similarity between these proteins, leading to a similarity matrix S that is not definite positive. How can we make a symmetric positive definite kernel matrix out of a pair-wise similarity matrix ? There is no single answer to this problem, but mainly three types of approaches that have been proposed to derive a kernel from such similarity score matrices.

- The first way to convert s into a valid kernel is called empirical kernel map. One replaces the similarity matrix S of this set of proteins by K, whose eigenvalues are equal to the square of those of S. They are therefore all positive, and K is now a definite positive matrix. Liao and Noble successfully applied this technique to transform an alignment score between protein sequences into a powerful kernel for remote homology detection: a protein is represented as a vector of log E-values from a pair-wise sequence comparison algorithm (195). It was also used in our group to convert structure-based similarity matrices generated by the MAMMOTH algorithm, using the log of the E-value returned by MAMMOTH. The resulting MAMMOTH-derived kernel was used to predict enzyme functions and Gene Ontology (129).

- One can add to the diagonal training matrix S a non-negative constant large enough to make it positive definite. This is equivalent to adding a constant to all eigenvalues in order to make them all positive. This approach was also used by Liao and Noble to detect protein sequence homology, with performance comparable to that of the first method.

- Another way is to perform eigenvalue decomposition of the similarity matrix, and to remove all negative eigenvalues. It was pointed out that this method preserves clusters in data, and showed promising experimental results in classifying protein sequences based on the FASTA scores (196). In this case, one uses the decomposition of the similarity matrix where n^2 is the size of S (S is a square matrix, and n corresponds to the number of proteins in the training set), and represent the column vectors of the diagonalization matrix for S. Then, one can define a kernel by keeping only positives eigenvalues.

The three types of approaches were found to display similar performance of test studies (124).

A future development of the work presented in this thesis would be to employ one these three types of approaches to convert supCK similarity matrices presented in chapter 4 into definite positive matrices. These matrices could then be used in SVM-based chemogenomic approaches in order to predict ligands on large protein datasets containing evolutionary unrelated proteins of known 3D structures.

5.3 Other structure-based kernels for proteins.

In chapter 4, we encode proteins by a cloud of points in 3D space, corresponding to the atoms defining the ligand-binding pocket. To compare two protein pockets, the similarity measure supCK requires to best superpose the two corresponding clouds of points. Although the method was found to present good performances, a drawback is that the results depend on this superposition step, and that this step is computationally costly. Therefore, an interesting research axis would be to test other structure-based similarity methods that do not require pocket superposition, and to derive new protein kernels that would be used in SVM chemogenomics. The methods described in the previous paragraph to build definite positive matrices from similarity matrices can be used for any protein pocket similarity measures that are independent from protein superposition. One such example is the similarity measure developed by Kahraman, to which the supCK method was compared in chapter 4 (156). However, most of the methods that are independent from protein superposition do not encode binding pockets explicitly by its cloud of atoms. They are therefore highly simplified representations of the pockets, which might lead to decreased performances with respect to supCK. Therefore, a methodological improvement could be to test other kernels based on explicit atom representations, but for kernels independent of any prior pocket superposition. In chapter 3, we used a 3D pharmacophore kernel developed for small molecules in order to predict ligands for GPCRs. This kernel only uses as input atom coordinates, and possibly various atom labels such as atom type or partial charge. It was developed in our group and is publicly available in the ChemCPP software. This kernel could be applied to the atoms of the cloud of atoms that define the protein pockets in a straightforward manner, without requiring any pocket superposition. This 3D pharmacophore kernel could be used as such in the SVM chemogenomic scheme similar to that presented in chapter 3. This would allow extend the presented chemogenomic SVM-based method for the

prediction of ligands, to the case of large datasets containing proteins that do not belong to the same family, as long as their 3D structures are known.

5.4 The learning database

All prediction methods rely on prior knowledge, i.e. a learning dataset. In order to predict protein-ligand interactions, the learning dataset must contain a list of known interactions which are used to start filling the chemogenomic matrix. However, chemogenomic studies have to deal with the paucity of protein-ligand interaction data, that usually cover only a small part of the protein-ligand space, particularly in the protein space. Therefore, most papers in chemogenomics report building of a specific dataset prior to the testing of methods. The question of how to build this learning dataset has not been discussed in detail in this thesis, although it is a crucial question in bioinformatics, often representing the limiting step.

For chemogenomic studies in which the biological space is encoded in a sequence-based approach, as we did in chapter 3, various databases are available, usually devoted to specific families of proteins. We used the GLIDA database that gathers known protein-ligand interactions for GPCRs. The IUPHAR database could also be interesting, since it contains known interactions for GPCRs, ion channels, and nuclear receptors, three major classes of drug targets (197).

Other databases contain interaction information for more diverse families of proteins such as ChemBank which stores raw data from screening assays (198), and DrugBank which contains information on drugs and their known targets(199). However, as mentioned earlier, sequence-based chemogenomic approach cannot handle protein diversity and are restrained make to predictions in a given family of proteins, based on a knowledge database of protein-ligand interactions within this family.

In the structure-based approaches presented in chapter 4, which encode the ligand-binding site, chemogenomic methods can learn from datasets containing very diverse proteins and/or protein structures, and the Protein Data Bank (PDB) constitutes the natural source of information. However, because membrane associated proteins are very difficult to crystallize, the main drawback of PDB is that it contains very few structures of protein receptors or channel ions. In the process of drug discovery, this database is very helpful for structure-based rational design by exploiting protein-bound conformations of known ligands, depicting their environment and the local flexibility of well-identified protein binding sites. However, described ligands are

considered from a structural point of view. This means that no difference is implied between a compound known to activate/inhibit the corresponding target and a molecule (e.g., solvent, detergent, and metal ion) devoid of pharmacological effect on that target. Carlson's group linked experimental binding data to 3-D structures from the PDB in the Binding MOAD database (200). The selection procedure combined with the bibliographic search ensured the choice of the appropriate ligand within biologically relevant complexes. This database resource covers about 10.000 ligand-protein complexes, including about 6000 different ligands. This outstanding collection of data greatly benefits the characterization of molecular recognition as well as the development of structure-based drug discovery techniques. However, a chemogenomic approach based on the encoding of the protein pocket as a cloud of points would require a well-defined collection of suitable binding sites including exact 3-D coordinates that are not available in any of the above-mentioned databases. The group of Rognan has created such a specialized database by parsing PDB files called sc-PDB (201). Selection was based on ligand properties, and a unique drugable cavity was assigned to each complex. The binding site was defined by all the protein residues with at least one atom within $6.5\mathring{A}$ of any ligand atom. The sc-PDB contains 2721 unique ligands within the 6415 complexes for which the binding site is stored in PDB format. This database could be of great interest to train SVM-based chemogenomic methods using kernels for binding sites such as the pharmacophore 3D kernel, or kernels derived from supCK similarity matrices.

5.5 Extension to proteins of unknown structures by homology modeling

Unlike sequence-based chemogenomics, structure-based chemogenomics allows to learn from protein-ligand available information across different families of proteins. However, its main limitation is the number of structures of protein-ligand complexes available at the PDB, in order to build learning datasets. Indeed, there are many more known protein-ligand interactions, than protein-ligand structures available at the PDB. Despite structural genomics efforts, it is unlikely that the three-dimensional structures of the entire human proteome will be available soon. One possible way to overcome this limitation is to use homology modeling to derive 3D models for proteins of unknown structures. These modeled structures could be used either to enrich the learning dataset, but also to extend the number of proteins for which prediction of interactions could be done. For example, the growing number of GPCR structures allows

to undertake homology modeling for GPCR proteins belonging to the family A (202). From these models, one could extract structures of protein binding pockets such as those used in chapter 4: sequence alignment of GPCRs from family A with the two GPCRs of known structures used in chapter 3 allows to identify key residues for ligand binding. The corresponding binding pocket in 3D would then be extracted from the corresponding GPCR model. Another approach would be to perform direct structural alignment of the models with the known structures using structural alignment algorithms such as STAMP in order to extract the binding pockets. The important family of kinases, involved in major human diseases like cancer, have also been modeled at the human kinome level (203). These modeled structures could be used in chemogenomic studies, together with the large number of data available about kinases inhibitors (see for example: (204)). We have shown that the sup-CK method presents smoothness properties for the comparison of cloud of atoms. This could be an interesting advantage when applied on homology models that may suffer from atom position inaccuracies, particularly for side-chain atoms.

5.6 Relation with docking

Docking programs can search for the best fit between two or more molecules by considering several parameters obtained from receptor and ligand atomic coordinates, such as geometrical complementarity, atomic VDW radius, charge, torsion angles, intermolecular hydrogen bonds, and hydrophobic contacts. Docking is therefore intended to model accurately the physical phenomenon of the binding of a molecule at the surface of a protein. As a result, docking applications return the predicted orientations (poses) of a ligand in the targets binding site. The posing process usually returns numerous possible conformations and several positions for a molecule. Scoring functions, which are able to evaluate intermolecular binding affinity or binding free energy, are employed to optimize and rank the results to obtain the best orientation after the docking procedure and selecting the best pose (205).

However, before performing the docking, it is necessary to perform a number of steps that cannot be easily automated. For example, docking conditions are usually tuned when redocking a ligand for which the structure of the protein-ligand complex is known. The docking conditions best reproducing the known complex are then used for large-scale docking against the studied receptor. Such tuning, for each receptor, makes it difficult to use when studying many proteins at once. In other words, docking approaches are well suited to large scale studies

in the ligand space, but are not easily applicable to large scale studies in the protein space. One advantage of chemogenomic approaches is that they do not require such tuning, and therefore, are applicable for large scale studies both in the protein and ligand spaces.

Another difference lies in the fact that docking is not a method that can ”learn” in a direct manner. In other words, knowing the best pose and affinity value of a ligand for a given protein might help to better dock other molecules in this protein. However, it will not help to better find the best docked pose and docking score of this ligand for another protein. On the contrary, in chemogenomic approach, any new information about protein-ligand interaction that becomes available in the learning database will potentially improve prediction of protein-ligand interactions.

However, docking could be very helpful to enlarge the learning dataset in the case of structure-based chemogenomics. Indeed, among the known protein-ligand interactions, many arise from in vitro affinity or enzymatic tests. Yet, there is no crystallographic complex for all these interactions, but in some cases, the corresponding proteins have a known structure in the PDB. In such cases, docking could be used to predict the structure of the protein-ligand complex, and extract atoms of the protein pocket (i.e. atoms in contact with the ligand). These predicted pockets would be added to the learning dataset, which would in turn improve future chemogenomics predictions. In other words, docking could help to enlarge available learning datasets.

Similarly, docking could be used in the case of homology models of proteins, further enriching the data that can be used by in chemogenomics.

The main limitation of chemogenomics with respect to docking, is that chemogenomics predict protein-ligand interactions, but do not provide the position of the ligand in the pocket: the geometry of the protein-ligand complex remains unknown. Therefore, further analysis such as ligand optimization to enhance protein-ligand affinity, as can be done in classical drug design approaches based on docking, is not easy. At this point, docking would be useful to model interactions predicted by chemogenomics.

5.7 From prediction of protein-ligand interactions to prediction of biological effects

Currently, the average time needed to develop a new drug takes between ten and twelve years, and the cost is estimated at several hundreds of millions of euros. Most drugs are small com-

pounds that interact with their targeted proteins. Drugs failure in clinical trials is mainly due to unexpected side-effects caused by interactions with proteins that are not the main target. In other words, the beneficial and unwanted effects of drugs are due to the overall spectrum of interaction of the drug against the human proteome. This underscores the need of large-scale approaches in the protein space, for predicting drug-protein interactions.

Indeed, the discovery of secondary targets in the final stages of a drug development is an important and recurring problem in the pharmaceutical industry. This problem has so far not been resolved, and the development of method that would help to predict potential off-targets at early stages of drug development would be of great interest for pharmaceutical industry. Chemogenomics may provide an early answer to this question of specificity, since this method can list potential secondary targets.

As the number of interactions data increases, and of 3D structures of complexes in the PDB increases, the quality of the predictions of chemogenomics and the number of proteins for which it is possible to make predictions will increase. This feature is encouraging because it allows to consider a consistent prediction of side effects in the medium term.

6

Conclusion

In conclusion, this thesis has shown that the use of machine learning methods is effective for predicting protein ligand (PL) interactions. This approach, as the name suggests, ”learns” from examples to recognize experimentally validated PL couples that could exist or not. To show the relevance of this method, a data set has been designed and made available to the community.

For ligands, we used descriptors corresponding to an encoding of the 2D or 3D structure. In the 2D approach, a molecule is described by a binary vector whose elements are determined by a graph that describes its chemical structure. In the 3D approach, molecules are described by the set of triplets of atoms that compose it, and the distances between these atoms. For proteins, two kernels have been designed. The hierarchical kernel evaluates the similarity between proteins based on their distance in the hierarchy of GPCRs, and the binding pocket kernel that assesses the similarity of the amino acids forming the binding sites of ligands. In the latter, the proteins of the GPCR family whose structure has been experimentally determined are structurally align in order to identify amino acids involved in ligand binding. The sequences of other GPCRs were then aligned with these two proteins and amino acids corresponding to the binding site were concatenated into a vector that allowed their comparison. Chemogenomics space is encoded by the tensor product of proteins and ligands spaces, and the distances between the pairs (protein, ligand) in this space is estimated by the product of the kernels calculated on the proteins and ligands. This method is only able to predict new interactions within a protein family whose members have a sequence and a shape globally similar enough to be compared.

With the idea of expanding the learning database and knowing that proteins with various shapes and sequences may bind to the same kind of ligands, we developed a similarity measure which is able to compare binding sites of proteins. For this, we used 3D structures of

proteins from the PDB. The binding sites, extracted from the ligands environment, are represented by clouds of atoms. The similarity between two proteins is then evaluated by the similarity between the clouds of atoms of their ligand binding pockets. This method involves a 3D alignment of the atoms forming the two pockets, by rotation and translation. The alignment is achieved by promoting the regrouping of atoms from the two pockets with similar properties in nearby regions of space. Using a data set from the literature and two others created for this purpose and made public, we have shown that the similarity measure is able to recognize pockets binding the same ligand. We also showed that the classification error is a better measure of prediction performance than the AUC which is conventionally used. This method has the advantage of comparing the binding sites of any two proteins, regardless of their similarities and their families, as long as 3D structures are available.

Finally, we have indicated possible ways of exploration to transform the similarity measure in a kernel, in order to provide a chemogenomics method which benefits from performances and characteristics of the SVM kernel methods.

Bibliography

[1] John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. How many drug targets are there? Nat Rev Drug Discov, 5(12):993–996, Dec 2006. 4

[2] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci U S A, 44(2):98–104, Feb 1958. 8

[3] Chiranjib Chakraborty, Chi-Hsin Hsu, Zhi-Hong Wen, and Chan-Shing Lin. Recent advances of fluorescent technologies for drug discovery and development. Curr Pharm Des, 15(30):3552–3570, 2009. 9

[4] Mara E Bosch, Antonio J R Snchez, Fuensanta S Rojas, and Catalina B Ojeda. Optical chemical biosensors for high-throughput screening of drugs. Comb Chem High Throughput Screen, 10(6):413–432, Jul 2007. 9

[5] Geoff Holdgate. Isothermal titration calorimetry and differential scanning calorimetry. Methods Mol Biol, 572:101–133, 2009. 9

[6] William B Peters, Verna Frasca, and Richard K Brown. Recent developments in isothermal titration calorimetry label free screening. Comb Chem High Throughput Screen, 12(8):772–790, Sep 2009. 10

[7] Walter Huber and Francis Mueller. Biomolecular interaction analysis in drug discovery using surface plasmon resonance technology. Curr Pharm Des, 12(31):3999–4021, 2006. 10

[8] Maria Miller. The early years of retroviral protease crystal structures. Biopolymers, 94(4):521–529, 2010. 11

[9] Jingshan Ren and David K Stammers. Structural basis for drug resistance mechanisms for non-nucleoside inhibitors of hiv reverse transcriptase. Virus Res, 134(1-2):157–170, Jun 2008. 11

[10] Sanjay Bhattacharya and Husam Osman. Novel targets for anti-retroviral therapy. J Infect, 59(6):377–386, Dec 2009. 11

[11] Scott C Blanchard, Barry S Cooperman, and Daniel N Wilson. Probing translation with small-molecule inhibitors. Chem Biol, 17(6):633–645, Jun 2010. 11

[12] Lee Fielding. Nmr methods for the determination of protein-ligand dissociation constants. Curr Top Med Chem, 3(1):39–53, 2003. 11

[13] J. Clarkson and I. D. Campbell. Studies of protein-ligand interactions by nmr. Biochem Soc Trans, 31(Pt 5):1006–1009, Oct 2003. 11

[14] Gaetano T Montelione and Thomas Szyperski. Advances in protein nmr provided by the nigms protein structure initiative: impact on drug discovery. Curr Opin Drug Discov Devel, 13(3):335–349, May 2010. 11

[15] Adam Golebiowski, Sean R Klopfenstein, and David E Portlock. Lead compounds discovered from libraries: part 2. Curr Opin Chem Biol, 7(3):308–325, Jun 2003. 12

[16] Michle N Schulz and Roderick E Hubbard. Recent progress in fragment-based lead discovery. Curr Opin Pharmacol, 9(5):615–621, Oct 2009. 13

[17] Gyrgy M Keseru and Gergely M Makara. Hit discovery and hit-to-lead approaches. Drug Discov Today, 11(15-16):741–748, Aug 2006. 13

[18] Ajay N Jain. Virtual screening in lead discovery and optimization. Curr Opin Drug Discov Devel, 7(4):396–403, Jul 2004. 13

[19] Brian K Shoichet, Susan L McGovern, Binqing Wei, and John J Irwin. Lead discovery using molecular docking. Curr Opin Chem Biol, 6(4):439–446, Aug 2002. 13

[20] Peter Kolb, Rafaela S Ferreira, John J Irwin, and Brian K Shoichet. Docking and chemoinformatic screens for new ligands and targets. Curr Opin Biotechnol, 20(4):429–436, Aug 2009. 13

[21] J. Xu and A. Hagler. Chemoinformatics and Drug Discovery. Molecules, 7:566–600, 2002. 14

[22] C. Manly, S. Louise-May, and J. Hammer. The impact of informatics and computational chemistry on synthesis and screening. Drug Discov. Today, 6(21):1101–1110, Nov 2001. 14

[23] A. L. Hopkins and C. R. Groom. The druggable genome. Nat. Rev. Drug Discov., 1(9):727–730, Sep 2002. 15, 68

[24] Helen M Berman, Tammy Battistuz, T. N. Bhat, Wolfgang F Bluhm, Philip E Bourne, Kyle Burkhardt, Zukang Feng, Gary L Gilliland, Lisa Iype, Shri Jain, Phoebe Fagan, Jessica Marvin, David Padilla, Veerasamy Ravichandran, Bohdan Schneider, Narmada Thanki, Helge Weissig, John D Westbrook, and Christine Zardecki. The protein data bank. Acta Crystallogr D Biol Crystallogr, 58(Pt 6 No 1):899–907, Jun 2002. 15

[25] W.Patrick Walters, Matthew T Stahl, and Mark A Murcko. Virtual screening–an overview. Drug Discovery Today, 3(4):160 – 178, 1998. 15

[26] Thomas Lengauer, Christian Lemmen, Matthias Rarey, and Marc Zimmermann. Novel technologies for virtual screening. Drug Discov Today, 9(1):27–34, Jan 2004. 15

[27] R. D. Brown and Y. C. Martin. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. J Chem Inf Comput Sci, 37:1–9, 1997. 16, 20

[28] R. Todeschini and V. Consonni. Handbook of Molecular Descriptors. Wiley-VCH, New York, 2002. 16, 43

[29] R. D. Brown and Y. C. Martin. An evaluation of structural descriptors and clustering methods for use in diversity selection. SAR QSAR Environ Res, 8(1-2):23–39, 1998. 16

[30] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug. Deliv. Rev, 46(1-3):3–26, Mar 2001. 16, 56

[31] J. Gasteiger and T. Engel, editors. Chemoinformatics : a Textbook. Wiley, New York, NY, USA, 2003. 17, 20

[32] G. Moreau and P. Broto. Autocorrelation of molecular structures: Application to SAR studies. Nouv. J. Chim., 757:764, 1980. 17

[33] M.J. McGregor and V. Pallai. Clustering of Large Databases of Compounds: Using the mdl "Keys" as Structural Descriptors. J Chem Inf Comput Sci, 37:443–448, 1997. 17

[34] L. Xue, F. L. Stahura, J. W. Godden, and J. Bajorath. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. J Chem Inf Comput Sci, 41(2):394–401, 2001. 17, 20

[35] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis. Frequent Substructure-Based Approaches for Classifying Chemical Compounds. IEEE T. Knowl. Data. En., 17(8):1036–1050, August 2005. 17

[36] C. Helma, T. Cramer, S. Kramer, and L. De Raedt. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. J. Chem. Inf. Comput. Sci., 44(4):1402–11, 2004. 17

[37] A. Inokuchi, T. Washio, and H. Motoda. Complete mining of frequent patterns from graphs: mining graph data. Mach. Learn., 50(3):321–354, 2003. 17

[38] M. Thimm, A. Goede, S. Hougardy, and R. Preissner. Comparison of 2D similarity and 3D superposition. Application to searching a conformational drug database. J Chem Inf Comput Sci, 44(5):1816–1822, 2004. 19

[39] P. Bultinck, T. Kuppens, X. Gironès, and R. Carbó-Dorca. Quantum similarity superposition algorithm (QSSA): a consistent scheme for molecular alignment and molecular similarity based on quantum chemistry. J Chem Inf Comput Sci, 43(4):1143–1150, 2003. 19

[40] C. Lemmen and T. Lengauer. Computational methods for the structural alignment of molecules. J. Comput. Aided. Mol. Des., 14(3):215–232, Mar 2000. 19

[41] T.G. Dietterich, R.H. Lathrop, and T. Lozano-Perez. Solving the Multiple Instance Problem with Axis-Parallel Rectangles. Artificial Intelligence, 89(1-2):31–71, 1997. 19

[42] N. Perry and V. J. van Geerestein. Database Searching on the basis of Three-Dimensional Similarity Using the sperm Program. J Chem Inf Comput Sci, 32:607–616, 1992. 19

[43] H. Kubinyi. Comparative Molecular Field Analysis. In J. Gasteiger, editor, Handbook of Chemoinformatics. From Data to Knowledge, Volume 4, pages 1555–1574. Wiley-VCH, Weinheim, 2003. 19

[44] C. A. Pepperrell and P. Willett. Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances. J Comput Aided Mol Des, 5(5):455–474, Oct 1991. 19

[45] M. Wagener, J. Sadowski, and J. Gasteiger. Autocorrelation of molecular surface properties for modeling. corticosteroid binding globulin and cytosolic. ah. receptor. activity by neural networks. J. Am. Chem. Soc., 117:7769–7775, 1995. 19

[46] R.E. Carhart, D.H. Smith, and R. Venkataraghavan. Atom Pairs as Molecular Features in Structure-Activity Studies: Definitions and Applications. J Chem Inf Comput Sci, 25:64–73, 1985. 19

[47] X. Chen, A. Russinko III, and S. S. Young. Recursive Partitioning Analysis of a Large Structure-Activity Data Set Using Three-Dimensional Descriptors. J Chem Inf Comput Sci, 38:1054–1062, 1998. 19

[48] S. J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, and P. Baldi. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. Bioinformatics, 21(Suppl. 1):i359–i368, Jun 2005. 19

[49] O. F. Güner. Pharmacophore Perception, Development, and Use in Drug Design, volume 2 of IUL Biotechnology Series. International University Line, 2000. 19, 45

[50] S. D. Pickett, J. S. Mason, and I. M. McLay. Diversity profiling and design using 3D pharmacophores : Pharmacophores-Derived Queries (PQD). J. Chem. Inf. Comput. Sci., 36(6):1214–1223, 1996. 19

[51] M. J. McGregor and S. M. Muskal. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. J Chem Inf Comput Sci, 39(3):569–574, 1999. 20

[52] H. Matter and T. Pötter. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. J. Chem. Inf. Comput. Sci., 39(6):1211–1225, 1999. 20

[53] E. Abrahamian, P. C. Fox, L. Naerum, I. T. Christensen, H. Thøgersen, and R. D. Clark. Efficient generation, storage, and manipulation of fully flexible pharmacophore multiplets and their use in 3-D similarity searching. J. Chem. Inf. Comput. Sci., 43(2):458–468, 2003. 20

[54] J. Saeh, P. Lyne, B. Takasaki, and D. Cosgrove. Lead hopping using SVM and 3D pharmacophore fingerprints. J Chem Inf Model, 45(4):1122–1133, Jul 2005. 20

[55] P. Mahé, L. Ralaivola, V. Stoven, and J.-P. Vert. The pharmacophore kernel for virtual screening with support vector machines. J. Chem. Inf. Model., 46(5):2003–2014, 2006. 20, 44, 48, 66

[56] A. R. Leach and V. J. Gillet. An introduction to chemoinformatics. Kluwer Academic Publishers, 2003. 20

[57] Philippe Ferrara, Holger Gohlke, Daniel J Price, Gerhard Klebe, and Charles L Brooks. Assessing scoring functions for protein-ligand interactions. J Med Chem, 47(12):3032–3047, Jun 2004. 21

[58] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification. Wiley-Interscience, 2001. 21

[59] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer, 2001. 21

[60] C. Hansch and T. Fujita. A method for the correlation of biological activity and chemical structure. J. Am. Chem. Soc, 86:1616–1626, 1964. 22

[61] A.K. Saxena and P. Prathipati. Comparison of mlr, pls and ga-mlr in qsar analysis. SAR. QSAR. Environ. Res., 14:433–445, 2003. 22

[62] D. Rogers and A. J. Hopfinger. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J Chem Inf Comput Sci, 34:854–866, 1994. 22

[63] Y. C. Martin, J. B. Holland, C. H. Jarboe, and N. Plotnikoff. Discriminant analysis of the relationship between physical properties and the inhibition of monoamine oxidase by aminotetralins and aminoindans. J. Med. Chem., 17(4):409–413, Apr 1974. 22

[64] C. Hansch, J. E. Quinlan, and G. L. Lawrence. Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. J. Org. Chem., 33:347 – 350, 1968. 22

[65] J. Zupan and J. Gasteiger. Neural Networks in Chemistry and Drug Design. Wiley-VCH, 1999. 22

[66] G. Schneider and P. Wrede. Artificial neural networks for computer-based molecular design. Prog Biophys Mol Biol, 70(3):175–222, 1998. 22

[67] D.M. Hawkins, S.S. Young, and A. Rusinko. Analysis of a large structure-activity data set using recursive partitioning. Quantitative Structure-Activity Relationships, 16:296–302, 1997. 23

[68] R. Burbidge, M. Trotter, B. Buxton, and S. Holden. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput. Chem., 26(1):4–15, December 2001. 23

[69] V. N. Vapnik. Statistical Learning Theory. Wiley, New-York, 1998. 23, 24

[70] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the 5th annual ACM workshop on Computational Learning Theory, pages 144–152, New York, NY, USA, 1992. ACM Press. 24

[71] J. Shawe-Taylor and N. Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press, New York, NY, USA, 2004. 24

[72] B. Schölkopf and A. J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, 2002. 24

[73] B . Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In M. Fayyad and R. Uthurusamy, editors, Proceedings of the First International Conference on Knowledge Discovery & Data Mining. AAAI Press, 1995. 24

[74] T. Joachims. Learning to Classify Text Using Support Vector Machines. Kluwer Academic Publishers, 2002. 24

[75] B. Schölkopf, K. Tsuda, and J.-P. Vert. Kernel Methods in Computational Biology. MIT Press, The MIT Press, Cambridge, Massachussetts, 2004. 24, 43, 48, 78

[76] J. Bockaert and J. P. Pin. Molecular tinkering of G protein-coupled receptors: an evolutionary success. EMBO J., 18(7):1723–1729, Apr 1999. 31

[77] D. A. Deshpande and R. B. Penn. Targeting G protein-coupled receptor signaling in asthma. Cell. Signal., 18(12):2105–2120, Dec 2006. 31

[78] S. J. Hill. G-protein-coupled receptors: past, present and future. Br. J. Pharmacol., 147 Suppl 1:S27–S37, Jan 2006. 31

[79] L. A. Catapano and H. K. Manji. G protein-coupled receptors in major psychiatric disorders. Biochim. Biophys. Acta, 1768(4):976–993, Apr 2007. 31

[80] B. B. Fredholm, T. Hökfelt, and G. Milligan. G-protein-coupled receptors: an update. Acta Physiol., 190(1):3–7, May 2007. 32

[81] S. H. S. Lin and O. Civelli. Orphan G protein-coupled receptors: targets for new therapeutic interventions. Ann. Med., 36(3):204–214, 2004. 32

[82] Wesley K Kroeze, Douglas J Sheffler, and Bryan L Roth. G-protein-coupled receptors at a glance. J Cell Sci, 116(Pt 24):4867–4869, Dec 2003. 32, 33

[83] Steven M Foord, Tom I Bonner, Richard R Neubig, Edward M Rosser, Jean-Phillipe Pin, Anthony P Davenport, Michael Spedding, and Anthony J Harmar. International union of pharmacology. xlvi. g protein-coupled receptor list. Pharmacol Rev, 57(2):279–288, Jun 2005. 32

[84] N. Nakayama, A. Miyajima, and K. Arai. Nucleotide sequences of ste2 and ste3, cell type-specific sterile genes from saccharomyces cerevisiae. EMBO J, 4(10):2643–2648, Oct 1985. 33

[85] K. G. Valentine, S. F. Liu, F. M. Marassi, G. Veglia, S. J. Opella, F. X. Ding, S. H. Wang, B. Arshava, J. M. Becker, and F. Naider. Structure and topology of a peptide segment of the 6th transmembrane domain of the saccharomyces cerevisae alpha-factor receptor in phospholipid bilayers. Biopolymers, 59(4):243–256, Oct 2001. 33

[86] P. S. Klein, T. J. Sun, C. L. Saxe, A. R. Kimmel, R. L. Johnson, and P. N. Devreotes. A chemoattractant receptor controls development in dictyostelium discoideum. Science, 241(4872):1467–1472, Sep 1988. 33

[87] Craig C Malbon. Frizzleds: new members of the superfamily of g-protein-coupled receptors. Front Biosci, 9:1048–1058, May 2004. 33

[88] Hsien-Yu Wang and Craig C Malbon. Wnt signaling, ca2+, and cyclic gmp: visualizing frizzled functions. Science, 300(5625):1529–1530, Jun 2003. 33

[89] Olivier Civelli, Yumiko Saito, Zhiwei Wang, Hans-Peter Nothacker, and Rainer K Reinscheid. Orphan gpcrs and their ligands. Pharmacol Ther, 110(3):525–532, Jun 2006. 34, 35

[90] J. A. Ballesteros, A. D. Jensen, G. Liapakis, S. G. Rasmussen, L. Shi, U. Gether, and J. A. Javitch. Activation of the beta 2-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. J Biol Chem, 276(31):29171–29177, Aug 2001. 34

[91] David A Shapiro, Kurt Kristiansen, David M Weiner, Wesley K Kroeze, and Bryan L Roth. Evidence for a model of agonist-induced activation of 5-hydroxytryptamine 2a serotonin receptors that involves the disruption of a strong ionic interaction between helices 3 and 6. J Biol Chem, 277(13):11441–11449, Mar 2002. 34

[92] M. J. Marinissen and J. S. Gutkind. G-protein-coupled receptors and signaling networks: emerging paradigms. Trends Pharmacol Sci, 22(7):368–376, Jul 2001. 34

[93] Susana R Neves, Prahlad T Ram, and Ravi Iyengar. G protein pathways. Science, 296(5573):1636–1639, May 2002. 34

[94] S. S. Ferguson. Evolving concepts in g protein-coupled receptor endocytosis: the role in receptor desensitization and signaling. Pharmacol Rev, 53(1):1–24, Mar 2001. 34

[95] Randy A Hall and Robert J Lefkowitz. Regulation of g protein-coupled receptor signaling by scaffold proteins. Circ Res, 91(8):672–680, Oct 2002. 34

[96] Jol Bockaert, Philippe Marin, Aline Dumuis, and Laurent Fagni. The 'magic tail' of g protein-coupled receptors: an anchorage for functional protein networks. FEBS Lett, 546(1):65–72, Jul 2003. 34

[97] A. Evers and T. Klabunde. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. J. Med. Chem., 48(4):1088–1097, Feb 2005. 35

[98] C. N. Cavasotto, A. J. W. Orry, and R. A. Abagyan. Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors. Proteins, 51(3):423–433, May 2003. 35, 67

[99] S. Shacham, Y. Marantz, S. Bar-Haim, O. Kalid, D. Warshaviak, N. Avisar, B. Inbal, A. Heifetz, M. Fichman, M. Topf, Z. Naor, S. Noiman, and O. M. Becker. PREDICT modeling and in-silico screening for G-protein coupled receptors. Proteins, 57(1):51–86, Oct 2004. 35

[100] C. Bissantz, P. Bernard, M. Hibert, and D. Rognan. Protein-based virtual screening of chemical databases. II. are homology models of G-protein coupled receptors suitable targets? Proteins, 50(1):5–25, Jan 2003. 35

[101] O. M. Becker, Y. Marantz, S. Shacham, B. Inbal, A. Heifetz, O. Kalid, S. Bar-Haim, D. Warshaviak, M. Fichman, and S. Noiman. G protein-coupled receptors: in silico drug discovery in 3D. Proc. Natl. Acad. Sci. USA, 101(31):11304–11309, Aug 2004. 35

[102] C. N. Cavasotto, A. J. W. Orry, N. J. Murgolo, M. F. Czarniecki, S. A. Kocsi, B. E. Hawes, K. A. O'Neill, H. Hine, M. S. Burton, J. H. Voigt, R. A. Abagyan, M. L. Bayne, and F. J. Monsma. Discovery of novel chemotypes to a G-protein-coupled receptor through ligand-steered homology modeling and structure-based virtual screening. J. Med. Chem., 51(3):581–588, Feb 2008. 35

[103] C. Rolland, R. Gozalbes, A. Nicolaï, M.-F. Paugam, L. Coussy, F. Barbosa, D. Horvath, and F. Revah. G-protein-coupled receptor affinity prediction based on the use of a

profiling dataset: Qsar design, synthesis, and experimental validation. J. Med. Chem., 48(21):6563–6574, Oct 2005. 35

[104] H. Kubinyi. Chemogenomics in drug discovery. Ernst Schering Res Found Workshop, 58:1–19, 2006. 36, 38

[105] S. E. Jaroch and H. Weinmann, editors. Chemical Genomics: Small Molecule Probes to Study Cellular Function. Ernst Schering Research Foundation Workshop. Springer, Berlin, 2006. 36

[106] T. Klabunde. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. Br. J. Pharmacol., 152:5–7, May 2007. 36

[107] P. R. Caron, M. D. Mullican, R. D. Mashal, K. P. Wilson, M. S. Su, and M. A. Murckoa. Chemogenomic approaches to drug discovery. Curr Opin Chem Biol, 5(4):464–470, Aug 2001. 38

[108] A. Schuffenhauer, P. Floersheim, P. Acklin, and E. Jacoby. Similarity metrics for ligands reflecting the similarity of the target proteins. J. Chem. Inf. Comput. Sci., 43(2):391–405, 2003. 39

[109] J. R. Bock and D. A. Gough. Virtual screen for ligands of orphan G protein-coupled receptors. J. Chem. Inform. Model., 45(5):1402–1414, 2005. 39, 40, 41

[110] D. Erhan, P.-J. L'heureux, S. Y. Yue, and Y. Bengio. Collaborative filtering on a family of biological targets. J. Chem. Inf. Model., 46(2):626–635, 2006. 39, 40, 41, 42

[111] L. Jacob and J.-P. Vert. Kernel methods for in silico chemogenomics. Technical Report 0709.3931v1, arXiv, 2007. 40

[112] Y. Okuno, J. Yang, K. Taneishi, H. Yabuuchi, and G. Tsujimoto. GLIDA: GPCR-ligand database for chemical genomic drug discovery. Nucleic Acids Res., 34(Database issue):D673–D677, Jan 2006. 40, 56

[113] J. R. Bock and D. A. Gough. Predicting protein-protein interactions from primary structure. Bioinformatics, 17(5):455–460, 2001. 41

[114] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized Kernels between Labeled Graphs. In T. Faucett and N. Mishra, editors, Proceedings of the Twentieth International Conference on Machine Learning, pages 321–328, New York, NY, USA, 2003. AAAI Press. 43

[115] T. Gärtner, P. Flach, and S. Wrobel. On graph kernels: hardness results and efficient alternatives. In B. Schölkopf and M. Warmuth, editors, Proceedings of the Sixteenth Annual Conference on Computational Learning Theory and the Seventh Annual Workshop on Kernel Machines, volume 2777 of Lecture Notes in Computer Science, pages 129–143, Heidelberg, 2003. Springer. 43

[116] P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, and J.-P. Vert. Graph kernels for molecular structure-activity relationship analysis with support vector machines. J. Chem. Inf. Model., 45(4):939–51, 2005. 43, 44

[117] L. Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi. Graph kernels for chemical informatics. Neural Netw., 18(8):1093–1110, Sep 2005. 44

[118] C.-A. Azencott, A. Ksikes, S. J. Swamidass, J. H. Chen, L. Ralaivola, and P. Baldi. One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties. J. Chem. Inform. Model., 47(3):965–974, 2007. 44

[119] J. Boström, J. R. Greenwood, and J. Gottfries. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. J. Mol. Graph. Model., 21(5):449–462, Mar 2003. 47

[120] T. Jaakkola, M. Diekhans, and D. Haussler. A Discriminative Framework for Detecting Remote Protein Homologies. J. Comput. Biol., 7(1,2):95–114, 2000. 48

[121] C. Leslie, E. Eskin, and W.S. Noble. The spectrum kernel: a string kernel for SVM protein classification. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Kevin Lauerdale, and Teri E. Klein, editors, Proceedings of the Pacific Symposium on Biocomputing 2002, pages 564–575, Singapore, 2002. World Scientific. 48

[122] K. Tsuda, T. Kin, and K. Asai. Marginalized Kernels for Biological Sequences. Bioinformatics, 18:S268–S275, 2002. 48

[123] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble. Mismatch string kernels for discriminative protein classification. Bioinformatics, 20(4):467–476, 2004. 48

[124] H. Saigo, J.-P. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. Bioinformatics, 20(11):1682–1689, 2004. 48, 119

[125] R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie. Profile-based string kernels for remote homology detection and motif extraction. J. Bioinform. Comput. Biol., 3(3):527–550, Jun 2005. 48

[126] M. Cuturi and J.-P. Vert. The context-tree kernel for strings. Neural Network., 18(4):1111–1123, 2005. 48

[127] P.D. Dobson and A.J. Doig. Predicting enzyme class from protein structure without alignments. J. Mol. Biol., 345(1):187–199, Jan 2005. 48

[128] K.M. Borgwardt, C.S. Ong, S. Schönauer, S.V.N. Vishwanathan, A.J. Smola, and H.-P. Kriegel. Protein function prediction via graph kernels. Bioinformatics, 21(Suppl. 1):i47–i56, Jun 2005. 48

[129] J. Qiu, J. Hue, A. Ben-Hur, J.-P. Vert, and W. S. Noble. A structural alignment kernel for protein structures. Bioinformatics, 23(9):1090–1098, May 2007. 48, 118

[130] T. Evgeniou, C. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. J. Mach. Learn. Res., 6:615–637, 2005. 49

[131] N. A. Kratochwil, P. Malherbe, L. Lindemann, M. Ebeling, M. C. Hoener, A. Mühlemann, R. H. P. Porter, M. Stahl, and P. R. Gerber. An automated system for the analysis of G protein-coupled receptor transmembrane binding pockets: alignment, receptor-based pharmacophores, and their application. J. Chem. Inf. Model., 45(5):1324–1336, 2005. 51, 66, 115

[132] Jean-Sebastien Surgand, Jordi Rodrigo, Esther Kellenberger, and Didier Rognan. A chemogenomic analysis of the transmembrane binding cavity of human g-protein-coupled receptors. Proteins, 62(2):509–538, Feb 2006. 51, 115

[133] T. Okada, M. Sugihara, A.-N. Bondar, M. Elstner, P. Entel, and V. Buss. The retinal conformation and its environment in rhodopsin in light of a new 2.2 a crystal structure. J. Mol. Biol., 342(2):571–583, Sep 2004. 52

[134] Vadim Cherezov, Daniel M Rosenbaum, Michael A Hanson, Sren G F Rasmussen, Foon Sun Thian, Tong Sun Kobilka, Hee-Jung Choi, Peter Kuhn, William I Weis, Brian K Kobilka, and Raymond C Stevens. High-resolution crystal structure of an engineered human beta2-adrenergic g protein-coupled receptor. Science, 318(5854):1258–1265, Nov 2007. 52

[135] R. B. Russell and G. J. Barton. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins, 14(2):309–323, Oct 1992. 52

[136] R. J. Lefkowitz, J.-P. Sun, and A. K. Shukla. A crystal clear view of the beta2-adrenergic receptor. Nat. Biotechnol., 26(2):189–191, Feb 2008. 52

[137] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. J. Mol. Graph., 14(1):33–8, 27–8, Feb 1996. 52, 53

[138] V. A. Avlani, K. J. Gregory, C. J. Morton, M. W. Parker, P. M. Sexton, and A. Christopoulos. Critical role for the second extracellular loop in the binding of both orthosteric and allosteric g protein-coupled receptor ligands. J. Biol. Chem., 282(35):25677–25686, Aug 2007. 54

[139] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res., 31(13):3497–3500, Jul 2003. 54

[140] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA, 89(22):10915–10919, Nov 1992. 54

[141] T. Mirzadegan, G. Benkö, S. Filipek, and K. Palczewski. Sequence analyses of G-protein-coupled receptors: similarities to rhodopsin. Biochemistry, 42(10):2759–2767, Mar 2003. 54

[142] J. Caldwell, I. Gardner, and N. Swales. An introduction to drug disposition: the basic principles of absorption, distribution, metabolism, and excretion. Toxicol. Pathol., 23(2):102–114, 1995. 56

[143] W. J. Egan, K. M. Merz, and J. J. Baldwin. Prediction of drug absorption using multivariate statistics. J. Med. Chem., 43(21):3867–3877, Oct 2000. 56

[144] D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward, and K. D. Kopple. Molecular properties that influence the oral bioavailability of drug candidates. J. Med. Chem., 45(12):2615–2623, Jun 2002. 56

[145] Y. C. Martin. A bioavailability score. J. Med. Chem., 48(9):3164–3170, May 2005. 56

[146] C. Chih-Wei, H. Chih-Chung and L. Chih-Jen. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University., 2003. 58

[147] K. Kristiansen, S. G. Dahl, and O. Edvardsen. A database of mutants and effects of site-directed mutagenesis experiments on G protein-coupled receptors. Proteins, 26(1):81–94, Sep 1996. 66, 67

[148] B. K. Kobilka. G protein coupled receptor structure and activation. Biochim. Biophys. Acta, 1768(4):794–807, Apr 2007. 67

[149] X. Yao, C. Parnot, X. Deupi, V. R. P. Ratnala, G. Swaminath, D. Farrens, and B. Kobilka. Coupling ligand structure to specific conformational switches in the beta2-adrenoceptor. Nat. Chem. Biol., 2(8):417–422, Aug 2006. 67

[150] J.-Z. Chen, J. Wang, and X.-Q. Xie. Gpcr structure-based virtual screening approach for cb2 antagonist search. J. Chem. Inf. Model., 47(4):1626–1637, 2007. 67

[151] X. Deupi, N. Dölker, M. L. Lòpez-Rodrìguez, M. Campillo, J. A. Ballesteros, and L. Pardo. Structural models of class a G protein-coupled receptors as a tool for drug design: insights on transmembrane bundle plasticity. Curr. Top. Med. Chem., 7(10):991–998, 2007. 67

[152] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, Adv. Neural. Inform. Process Syst. 19, pages 41–48, Cambridge, MA, 2007. MIT Press.

[153] Edwin Bonilla, Kian Ming Chai, and Chris Williams. Multi-task gaussian process prediction. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, Advances in Neural Information Processing Systems 20. MIT Press, Cambridge, MA, 2008.

[154] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: operator estimation with spectral regularization. J. Mach. Learn. Res., 10:803–826, 2009.

[155] Lei Xie, Li Xie, and Philip E Bourne. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. Bioinformatics, 25(12):i305–i312, Jun 2009. 76

[156] A. Kahraman, R. J. Morris, R. A. Laskowski, and J. M. Thornton. Shape variation in protein binding pockets and their ligands. J. Mol. Biol., 368(1):283–301, Apr 2007. 76, 77, 81, 82, 85, 87, 88, 100, 119

[157] R. J. Morris, R.J. Najmanovich, A. Kahraman, and J.M. Thornton. Real spherical harmonic expansion coefficients as 3d shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics, 21(10):2347–2355, May 2005. 76, 80

[158] N.D. Gold and R.M. Jackson. Sitesbase: a database for structure-based protein-ligand binding site comparisons. Nucleic Acids Res., 34:D231–D234, Jan 2006. 76

[159] A. Shulman-Peleg, M. Shatsky, R. Nussinov, and H. J. J. Wolfson. Multibind and mappis: webservers for multiple alignment of protein 3d-binding sites and their interactions. Nucleic Acids Res., 36:260–264, May 2008. 76, 78, 81

[160] C. Schalon, J-S. Surgand, E. Kellenberger, and D. Rognan. A simple and fuzzy method to align and compare druggable ligand-binding sites. Proteins, 71(4):1755–1778, Jun 2008. 76

[161] N. Weskamp, E. Hullermeier, D. Kuhn, and G. Klebe. Multiple graph alignment for the structural analysis of protein active sites. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 4(2):310–320, 2007. 76

[162] R. Najmanovich, N. Kurbatova, and J. Thornton. Detection of 3d atomic similarities and their use in the discrimination of small molecule protein-binding sites. Bioinformatics, 24(16):i105–i111, Aug 2008. 76

[163] Alexandra Shulman-Peleg, Ruth Nussinov, and Haim J Wolfson. Recognition of functional sites in protein structures. J Mol Biol, 339(3):607–633, Jun 2004. 76, 116

[164] Alexandra Shulman-Peleg, Ruth Nussinov, and Haim J Wolfson. Siteengines: recognition and comparison of binding sites and protein-protein interfaces. Nucleic Acids Res, 33(Web Server issue):W337–W341, Jul 2005. 76

[165] P. Willett, V. Winterman, and D. Bawden. Implementation of nearest-neighbor searching in an online chemical structure search system. J. Chem. Inform. Comput. Sci., 26(1):36–41, 1986. 78

[166] J.R. Davies, R.M. Jackson, K.V. Mardia, and C.C. Taylor. The poisson index: a new probabilistic model for protein ligand binding site similarity. Bioinformatics, 23(22):3001–3008, Nov 2007. 78, 81

[167] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol., 48:443–453, 1970. 82, 84

[168] P. Rice, I. Longden, and A. Bleasby. Emboss: the european molecular biology open software suite. Trends Genet., 16(6):276–277, Jun 2000. 82

[169] W. R. P. Scott, I. G. Tironi, A. E. Mark, S. R. Billeter, J. F., A. E. Torda, T. Huber, and P. Kruger. The gromos biomolecular simulation program package. J. Phys. Chem. A, 103:3596–3607, 1999. 84

[170] B. Schölkopf, A.J. Smola, and K.-R. Müller. Kernel principal component analysis. In B. Schölkopf, C. Burges, and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, pages 327–352. MIT Press, 1999. 93

[171] A. T. R. Laurie and R. M. Jackson. Q-sitefinder: an energy-based method for the prediction of protein–ligand binding sites. Bioinformatics, 21(9):1908–1916, 2005. 100

[172] F. Glaser, R. J. Morris, R. J. Najmanovich, R. A. Laskowski, and J. M. Thornton. A method for localizing ligand binding pockets in protein structures. Proteins, 62(2):479–488, February 2006. 100

[173] Tina A Eyre, Fabrice Ducluzeau, Tam P Sneddon, Sue Povey, Elspeth A Bruford, and Michael J Lush. The hugo gene nomenclature database, 2006 updates. Nucleic Acids Res, 34(Database issue):D319–D321, Jan 2006. 113

[174] Christopher M Dobson. Chemical space and biology. Nature, 432(7019):824–828, Dec 2004. 113

[175] L. Terfloth. Chemoinformatics: A Textbook, chapter Calculation of Structure Descriptors. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG., 2004. 114

[176] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, 25:3389–3402, 1997. 114

[177] C. Leslie, E. Eskin, J. Weston, and W.S. Noble. Mismatch String Kernels for SVM Protein Classification. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, Advances in Neural Information Processing Systems 15. MIT Press, 2003. 114

[178] T. Klabunde. Chemogenomics approaches to ligand design. In Ligand Design for G Protein-coupled Receptors, chapter 7, pages 115–135. Wiley-VCH, Great Britain, 2006. 115

[179] D. Rognan, R. Mannhold, H. Kubinyi, and G. Folkers. Ligand Design for G Protein-coupled Receptors, Volume 30. Wiley, 2006. 115

[180] T. M. Frimurer, T. Ulven, C. E. Elling, L.-O. Gerlach, E. Kostenis, and T. Högberg. A physicogenetic method to assign ligand-binding relationships between 7tm receptors. Bioorg. Med. Chem. Lett., 15(16):3707–3712, Aug 2005. 115

[181] Angel R Ortiz, Charlie E M Strauss, and Osvaldo Olmea. Mammoth (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci, 11(11):2606–2621, Nov 2002. 115

[182] M. Hue. Semi-supervised learning for protein structure prediction. Master's thesis, Ecole des Mines de Paris, 2004. 115

[183] Thorsten Naumann and Hans Matter. Structural classification of protein kinases using 3d molecular interaction field analysis of their ligand binding sites: target family landscapes. J Med Chem, 45(12):2366–2378, Jun 2002. 116

[184] Christian Hoppe, Christoph Steinbeck, and Gerd Wohlfahrt. Classification and comparison of ligand-binding sites derived from grid-mapped knowledge-based potentials. J Mol Graph Model, 24(5):328–340, Mar 2006. 116

[185] Bernard Pirard and Hans Matter. Matrix metalloproteinase target family landscape: a chemometrical approach to ligand selectivity based on protein binding site analysis. J Med Chem, 49(1):51–69, Jan 2006. 116

[186] M. Rosen, S. L. Lin, H. Wolfson, and R. Nussinov. Molecular shape comparisons in searches for active sites and functional similarity. Protein Eng, 11(4):263–277, Apr 1998. 116

[187] Kengo Kinoshita and Haruki Nakamura. Identification of protein biochemical functions by similarity search using the molecular surface database ef-site. Protein Sci, 12(8):1589–1595, Aug 2003. 116

[188] Stefan Schmitt, Daniel Kuhn, and Gerhard Klebe. A new method to detect related function among proteins independent of sequence and fold homology. J. Mol. Biol., 323(2):387–406, Oct 2002. 116

[189] Martin Jambon, Anne Imberty, Gilbert Delage, and Christophe Geourjon. A new bioinformatic approach to detect common 3d sites in protein structures. Proteins, 52(2):137–145, Aug 2003. 116

[190] Robert Powers, Jennifer C Copeland, Katherine Germer, Kelly A Mercier, Viswanathan Ramanathan, and Peter Revesz. Comparison of protein active site structures for functional annotation of proteins and drug design. Proteins, 65(1):124–135, Oct 2006. 116

[191] N.D. Gold and R.M. Jackson. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J. Mol. Biol., 355(5):1112–1124, Feb 2006. 116

[192] E. J. Gardiner, P. J. Artymiuk, and P. Willett. Clique-detection algorithms for matching three-dimensional molecular structures. J Mol Graph Model, 15(4):245–253, Aug 1997. 116

[193] Alexander Weber, Angela Casini, Andreas Heine, Daniel Kuhn, Claudiu T Supuran, Andrea Scozzafava, and Gerhard Klebe. Unexpected nanomolar inhibition of carbonic anhydrase by cox-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. J Med Chem, 47(3):550–557, Jan 2004. 116

[194] Jianghong An, Maxim Totrov, and Ruben Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics, 4(6):752–761, Jun 2005. 117

[195] L. Liao and W.S. Noble. Combining Pairwise Sequence Similarity and Support Vector Machines for Detecting Remote Protein Evolutionary and Structural Relationships. J. Comput. Biol., 10(6):857–868, 2003. 118

[196] V. Roth, J. Laub, J. Buhmann, and K. Mller. Going metric: Denoising pairwise data. In Advances in Neural Information Processing Systems, 2003. 118

[197] Joanna L Sharman, Chidochangu P Mpamhanga, Michael Spedding, Pierre Germain, Bart Staels, Catherine Dacquet, Vincent Laudet, Anthony J Harmar, and N. C-I. U. P. H. A. R. Iuphar-db: new receptors and tools for easy searching and visualization of pharmacological data. Nucleic Acids Res, 39(Database issue):D534–D538, Jan 2011. 120

[198] Kathleen Petri Seiler, Gregory A George, Mary Pat Happ, Nicole E Bodycombe, Hyman A Carrinski, Stephanie Norton, Steve Brudz, John P Sullivan, Jeremy Muhlich, Martin Serrano, Paul Ferraiolo, Nicola J Tolliday, Stuart L Schreiber, and Paul A Clemons. Chembank: a small-molecule screening and cheminformatics resource database. Nucleic Acids Res, 36(Database issue):D351–D359, Jan 2008. 120

[199] David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res, 36(Database issue):D901–D906, Jan 2008. 120

[200] Mark L Benson, Richard D Smith, Nickolay A Khazanov, Brandon Dimcheff, John Beaver, Peter Dresslar, Jason Nerothin, and Heather A Carlson. Binding moad, a high-quality protein-ligand database. Nucleic Acids Res, 36(Database issue):D674–D678, Jan 2008. 121

[201] Esther Kellenberger, Pascal Muller, Claire Schalon, Guillaume Bret, Nicolas Foata, and Didier Rognan. sc-pdb: an annotated database of druggable binding sites from the protein data bank. J Chem Inf Model, 46(2):717–727, 2006. 121

[202] Catherine L Worth, Gunnar Kleinau, and Gerd Krause. Comparative sequence and structural analyses of g-protein-coupled receptor crystal structures and implications for molecular models. PLoS One, 4(9):e7011, 2009. 122

[203] Michal Brylinski and Jeffrey Skolnick. Comprehensive structural and functional characterization of the human kinome by protein structure modeling and ligand virtual screening. J Chem Inf Model, 50(10):1839–1854, Oct 2010. 122

[204] Xi Chen, Yuhmei Lin, Ming Liu, and Michael K Gilson. The binding database: data management and interface design. Bioinformatics, 18(1):130–139, Jan 2002. 122

[205] Ren Thomsen and Mikael H Christensen. Moldock: a new technique for high-accuracy molecular docking. J Med Chem, 49(11):3315–3321, Jun 2006. 122

Développement d'approches de chémogénomique pour la prédiction des interactions protéine - ligand

Résumé : Cette thèse porte sur le développement de méthodes bioinformatiques permettant la prédiction d'interactions protéine - ligand. L'approche employée est d'utiliser le partage entre protéines des informations connues, à la fois sur les protéines et sur les ligands, afin d'améliorer la prédiction de ces interactions. Les méthodes proposées appartiennent aux méthodes dites de chémogénomique. La première contribution de cette thèse est le développement d'une méthode d'apprentissage statistique pour la prédiction des interactions protéines - ligands par famille. Elle est illustrée dans le cas des GPCRs. Cette méthode comprend la proposition de noyaux pour les protéines qui permettent de prendre en compte la similarité globale des GPCRs par l'utilisation de la hiérarchie issue de l'alignement des séquences de cette famille, et la similarité locale au niveau des sites de fixation des ligands de ces GPCRs grâce à l'utilisation des structures 3D connues des membres de cette famille. Pour cela un jeu de données a été créé afin d'évaluer la capacité de cette méthode à prédire correctement les interactions connues. La deuxième contribution est le développement d'une mesure de similarité entre deux sites de fixation de ligands provenant de deux protéines différentes représentés par des nuages d'atomes en 3D. Cette mesure implique la superposition des poches par rotation et la translation, avec pour but la recherche du meilleur alignement possible en maximisant le regroupement d'atomes ayant des propriétés similaires dans des régions proches de l'espace. Les performances de cette méthode ont été mesurées à l'aide d'un premier jeu de données provenant de la littérature et de deux autres qui ont été créés à cet effet.

L'ensemble des résultats de cette thèse montre que les approches de chémogénomique présentent de meilleures performances de prédiction que les approches classique par protéine.

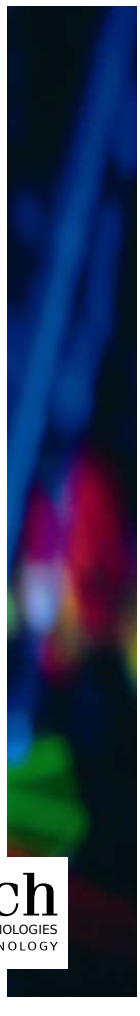
Mots clés : Chémogénomique, bioinformatique, criblage virtuel, apprentissage statistique, SVM, noyaux, mesure de similarité, structure 3D, interactions protéines ligands.

Development of chemogenomic approaches for prediction of protein-ligand interactions

Abstract: This thesis focuses on the development of bioinformatics methods for the prediction of protein-ligand interactions. The approach used throughout this thesis is to share the known information, both on proteins and on ligands to improve the performance of predictions. The first contribution is the development of a statistical learning method for the prediction of protein - ligands interactions within a family, and is illustrated in then case of GPCRs. This method involves the proposal of new kernels for proteins which take into account the overall similarity of GPCRs based on a sequenced-based hierarchy, and the local similarity of the ligand binding sites of GPCRs based on known 3D structures of known members of this family. A dataset was created to assess the ability of this method to correctly predict the known interactions. The second contribution is the development of a similarity measure between two ligands binding sites from two different (and potentially unrelated) proteins represented by clouds of atoms in 3D. This measure requires pockets alignment using rotations and translations, with the aim of finding the best possible alignment by maximizing the gathering of atoms with similar properties in the nearby regions of space. The performance of this method were measured using a first dataset described in the literature and two others that were created for this purpose.

Overall, the results show that chemogenomic approaches display better prediction performances than classical approaches.

Keywords: Chemogenomics, bioinformatics, virtual screening, machine learning, SVM, kernel, similarity measure, 3D structure, protein ligand interactions.



MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY