Biochemoinformatics: integrative computational tools at the interface between chemistry and biology

Ricard Garcia Serna

DOCTORAL THESIS UPF / 2010

Thesis Director:

Dr. Jordi Mestres

(CEXS Department)

The research in this thesis has been carried out at the Chemogenomics Laboratory (CGL) within the Unitat de Recerca en Informàtica Biomèdica (GRIB) at the Parc de Recerca Biomèdica de Barcelona (PRBB).

The research carried out in this thesis has been supported by Chemotargets S.L.

A mi madre y mi hermano.

Acknowledgements

Primer de tot volia donar-li les gràcies al meu director de tesi el Dr. Jordi Mestres, no només per tot l'esforç que ha posat en aquest treball sino també per haver-me permès fer recerca sense moure'm de la meva ciutat, cosa que per mi no té preu. Ha costat però al final ens n'hem ensortit!

Després volia agraïr els ànims i la companyia que m'ha donat tota la gent que ha anat passant aquests anys per aquest entorn laboral nostre del PRBB, tan entranyable i inspirador. Hem passat un munt de bones estones junts i almenys jo en tinc grans records... bé n'hi ha algun de concret, sobretot de la època de Bill Clinton, que preferiria no tenir. Tornant al tema que ens interessa doncs això, que moltes gràcies a tots: Albert, Alfons, Angel, Anna, Carina, Cristian, Ferran, Ingo, Jorge, Judith, Laura, Loris, Maricarmen, Marta, Miguel, Montse, Nicolas, Nikita (convida'm al teu poble quan puguis, va xD), Oscar, Pau, Xavi... i tots els que m'estigui oblidant per burro. Espero haver estat a la vostra alçada. Sé que aquesta última època no m'heu vist gaire el pèl, però és que el senyor del paràgraf anterior em feia treballar molt. Ara vindré a dinar tot sovint i vull que em tingueu al tanto de totes les xafarderies! També m'agradaria fer menció especial a Praveena [*thank you very much for being so nice with me*]; a David V. Hasselhoff, ja veuras que quan no estic estressat puc resultar simpàtic i tot; i a Baldo Oliva, que tot i que crec que no hem creuat més de tres frases en aquest sis anys, si estic aquí és gràcies a que tu em vas posar en contacte amb en Jordi i això no ho oblidaré en la vida.

També volia donar-li les gràcies a tota la gent de fora del laboratori, començant per la penya de superherois formada per Albert, Joan, Jordi (si Cire, ets tu), Marc, Mireia, Míriam, Noelia, Roger i Sergi. Hem passat tot de coses junts però com diu el nostre lema, no hem de recordar els moments mítics sino crear-ne de nous, “estamos?”. No puc tampoc oblidar a la gent del barri: l'Enric i el Jordi (quasi 20 anys que ens coneixem ja, nens!) ni al Dr. Luis Grandío (yo

también soy doctor tio!). Un especial recuerdo también para Ivan, futuro nobel de física.

I què puc dir del matriarcat Vinyallonga? Maria Rosa, Iuca, Tieta, Leta, Marc, Ter i JosepLluis vosaltres sí que m'heu donat un munt sense esperar res a canvi. Una abraçada ben forta per tots i prepareu-vos que a partir d'ara m'apunto a totes les comilones! Tot i que ja sabeu que menjo poquet... I no oblidem la secció gatuna: Fu, Urpa, Gati, Naga, Pudi, Coet, Grapa i Neula. El proper dia que us vegi us estrujaré a tots!

Quería también aprovechar para dedicar unas palabras a mi Padre, a mi Hermana, a Jose y a JosePardoGarcía. Os quiero mucho. Somos poquitos pero hacemos una buena piña, y la tenemos que hacer aún mejor. En mi nueva vida me pongo el propósito de dedicaros mucho más tiempo que hasta ahora, estáis dispuestos a aceptar el reto? I papá, que sepas que cada dia que pasa me parezco más y más a ti, ves como aún vamos a sacar algo de provecho de este piltrafilla! También un abrazo desde aquí a la tieta Palmi, que me has querido siempre un montón y yo no he sabido corresponderte adecuadamente; por suerte aún estamos a tiempo de arreglarlo. Molts petons també pel Ricard i la Meri, hem de quedar per veure totes les reformes!

No puedo acabar sin tener un momento de recuerdo para la gente que ya no está. Aunque en ese aspecto sí que tengo una mentalidad científica y a conciencia no puedo creer que haya algo más allá de la realidad física más aparente, lo que sí os puedo decir es que mientras yo esté por aquí vosotros también estaréis y que cultivo vuestra memoria para seguir aprendiendo de vosotros lo que en su momento fui tonto y no capté. ¿Cuántas cosas cambiaría si pudiera volver atrás? Pero no puedo, lo hecho hecho está y sólo me consuelo con el hecho de que en adelante voy a revertir todo lo que me disteis en la gente que quiero. Gracias.

I pel final he deixat la persona més important de totes. Hola Sara! Aquella única persona per la qual si em vinguessin a buscar uns extraterrestres per anar a explorar l'univers no hi aniria. Què més puc dir? Bé, tu ja saps que

Leónidas d'Esparta tenia un vocabulari més aviat reduït així que ho haurem de deixar aquí [:oD]. ABO forever!

"*Agarrados del viento viviremos,*
no me importa a dónde vamos."

Preface

Recent technological improvements and the corresponding explosion in data generation and collection are expanding the knowledge base of traditional isolated disciplines and transforming life sciences into a continuum integrated domain. In particular, the application of integrative approaches in chemical biology and drug discovery is having a major impact on the identification of chemical probes for biological systems and the design and optimization of safer, more efficient drugs.

With this vision, the main objective of the present thesis was the development of new methods and tools that contribute to the advancement of integrative approaches to life sciences. The document has been divided in six parts. The first part provides an overview of current trends in integrative biomedical sciences and describes the different areas that have been the focus of our research, namely, the chemical space defined by small molecules, the biological space defined by proteins of therapeutic relevance, and the phenotypical space defined by drug side effects. The next two parts introduce the primary objectives pursued and discuss the main results obtained, respectively. The final three parts compile the six publications that have resulted from this thesis, the main conclusions derived, and the general list of relevant references, respectively.

List of publications

- Cases M, García-Serna R, Hettne K, Weeber M, van der Lei J, Boyer S, Mestres J: **Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family**. *Curr. Top. Med. Chem.* 2005, **5**: 763-772

- García-Serna R, Opatowski L, Mestres J: **FCP: functional coverage of the proteome by structures**. *Bioinformatics* 2006, **22**: 1792-1793

- Garcia-Serna R, Ursu O, Oprea TI, Mestres J: **iPHACE: integrative navigation in pharmacological space**. *Bioinformatics* 2010, **26**: 985-986

- Vidal D, Garcia-Serna R, Mestres J: **Ligand-based Approaches to In Silico Pharmacology**. *Methods Mol. Biol.* 2011, **672** (in press)

- Garcia-Serna R, Mestres J: **Anticipating drug side effects by comparative pharmacology**. *Expert Opin. Drug Metab. Toxicol.* 2010 (submitted)

- Garcia-Serna R, Mestres J: **Chemical probes for biological systems**. *Drug Discov. Today* 2010 (submitted)

Part I – Introduction

Chapter I.1 – Integrative approach to biomedical sciences

I.1.1 Relevant entities in biomedical sciences

The ultimate goal of biomedical sciences could be defined as the holistic understanding of the mechanisms behind all our body processes at all levels in order to be able to reverse damaged or malfunctioning systems to a normalized state. Huge economic resources are dedicated to this vast and heterogeneous field worldwide every year both by public administrations and by some of the more powerful transnational companies from the pharmaceutical industry.

Diseases have significant social or demographic impact on human populations and thus, different interests put it under the focus of biomedical research in academic institutions and private corporations to start looking for therapeutic remedies at different levels. Individuals suffering from the anomalous set of characteristics associated with the disease are the first study cases in the path to understand the causes of the problem with the objective of experimentally apply the scientific method in search of a treatment and cure.

But at this starting point, an endless list of questions emerges. Which are the observable effects of the disease beyond individual specificities and environmental heterogeneities? Are there any organic malfunctions behind those effects? In which specific tissues reside the main alterations? Which altered metabolic processes cause those abnormal biological behaviors? What is the new set of characteristics that affected cells show and how many elements of the cellular machinery are implied in the non standard processes? Are there any foreign chemicals related with the development of the disease or is it possible to design drugs for its treatment? None of these questions is easy to answer because each of them involves different disciplines and experimental procedures as they cover different biomedical entities from metabolites to synthetic small molecules, from ADN and ARN to multiple proteins and pathways, from cells to tissues and organs, from individual phenotypes to statistics in populations, and so on.

When we consider each facet of the problem at a time, specific issues can be studied separately and successfully addressed, but the complete solution is far beyond the reaches of any single involved discipline and integrative approaches are required [1]. Human diseases have genetic, pharmacological and environmental parameters that interrelate in an intricate network that presents an altered state and needs to be reconverted to the basal situation. Beyond the study of the single perturbations in the system, the whole network needs to be considered as the fundamental unit of the analysis [2], if an ultimate solution is to be found.

The final years of the last century witnessed the emergence and evolution of many disciplines around this issue. Systems biology is the more general term that includes all approaches based on a holistic point of view, in contrast to the traditional reductionism, and has been applied in different fields in the "omics" area relaying on mathematical and computational tools [3]. Another non specific widely employed term is bioinformatics, referring to the conceptualization of relevant entities in biomedical sciences and the information technologies applied to their analysis [4]. In between the cohesive approach of systems biology and the focused points of view of each single basic domain, many disciplines are found at different integration levels. For instance chemogenomics is one of them, integrating the fields of informatics, chemistry, and pharmacology to address drug discovery at a protein family level.

Recent advancements in the various experimental disciplines, alongside with the development of modern integrative knowledge-based computational approaches may one day make possible the ultimate paradigm of biomedical sciences: the establishment of a computational model of a complete living organism at all levels of organization so that the effects of any genetic, environmental or chemical perturbation can be predicted and the most optimal therapeutic solution proposed. Advanced mathematics have already allowed the development of simple models for certain metabolic and signal transduction processes [5] but the progress in this area is still limited by currently available computing capacities [6].

I.1.2 Emergent synergy of data integration

Data derived from the measurement of the relevant features for abovementioned entities is the sap of the knowledge tree of biochemoinformatics. From an applied point of view, this information is also considered the basis of drug discovery [7], a process that requires expertise in multiple interrelated disciplines.

Although the size of accessible data increases at an exponentially growing rate, the computational capacities available to researchers are, in most cases, sufficient to deal with such amount of data. In fact, while the theoretic quantity of information encoded in nature has cosmic proportions, the amount of real information that researchers in this field manage is much lower than in other scientific domains. Then, far from being overwhelmed by the quantity of information, the actual difficulty comes from the diversity of scientific domains covered and the number of attributes that have to be considered at the time. Accordingly, from the raw unprocessed data obtained in numerous unrelated experimental assays, researchers are in need to derive high level integrated information systems that can be interrogated in search for complex answers to cross-domain questions. Obviously the efforts dedicated to this process will need to be multiplied as the diversity and richness of primary data increases.

In this respect, several steps will have to be taken in order to achieve this ambitious goal. First of all, data has to be normalized by specific weighting and scaling protocols to become comparable. Then the connectivity networks between the distinct sources of data need to be defined in order to establish links between the same elements across domains. Finally, this derived data needs to be presented and visualized in such a way that high level analyses can be easily performed, with specific representations being more appropriate depending on the nature of the source [8]. The web application iPhace is a good example of a tool designed and implemented as a specialized visualization tool at the interface of chemistry and biology (see chapter **IV.3** for further details).

One of the main obstacles in this process is the identification of the same entities and processes across the different domains involved, and sometimes

even within the same research area, to avoid connectivity problems. Some researchers have proposed that this should be the main focus of biomedical sciences as of today [9]. To solve this problem we need to go beyond format issues and take into account the deep meanings or semantics that lay behind the different syntaxes used. This can be accomplished by the design and application of directed controlled vocabularies, aggregations of precisely defined unique terms relevant to a certain field that are unambiguously identified and can be systematically organized into highly curated ontologies which have already proven to be of utmost importance in this area [10] given that higher levels of logic relational organization open the path for a deeper analysis. From the first initiatives in this field, like GO (Gene Ontology) [11], these have grown in number and diversity in the recent years at the same time that interest from researches has focused on them because of their promising applications [12]. The web application FCP is a good example of a tool designed to provide a functional classification of all experimentally determined structures currently deposited in the Protein Data Bank (see chapter IV.2 for further details).

While in the last years the scientific community has devoted huge efforts to this issue, disappointment has raised when even a priory successful approaches have not been able to be reflected in, for instance, significant increases of the success rates in the drug discovery process [13]. However, these success expectations were optimistic because when different systems of information are related, a new more complex meta-system is revealed, hence requiring more complex approaches for its analysis. The development of these new methodologies can cover, on their own, the scope of non-negligible projects and hence the fruits of this work are likely to be revealed after some time.

Network-based approaches have extensively proven their utility at this high-level analysis [14] and, for instance, the use of “intelligent” symbolic systems [15] able to make new suggestions by the computational analysis of well designed knowledge bases has been proposed. Therefore, beyond traditional data extraction and browsing of databases, data needs to be organized in flexible but specific architectures from which meta-data and

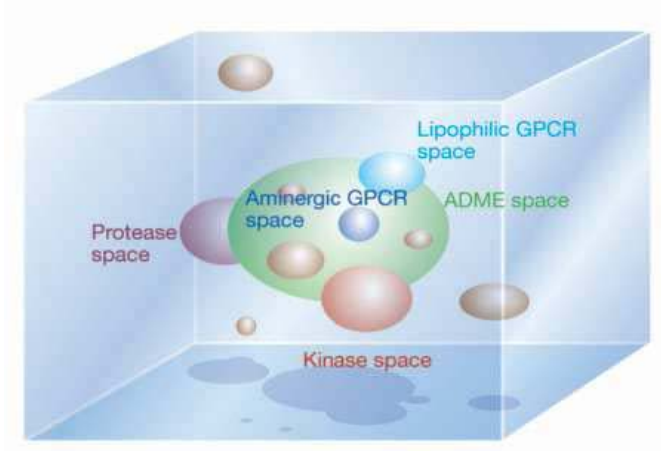
unseen correlations may be inferred.

In summary, we could state that the complexity of a living organism cannot be expected to be easily understood with partial, limited, and isolated views, but integrative approaches need to be developed instead. When data coming from different domains that refers to the same semantic entity are combined, the true signals may be synergistically reinforced while the background noise will tend to be inherently corrected. Furthermore, new characteristics might be expected to become apparent at the systems level, whereas they could not be perceived from the isolated components [16]. Many examples of the utility of this wider approach can be found in recent literature across all domains of biomedical sciences to different extents [17-21]. Along this view, the main purpose of this thesis is to contribute to the path of integrating domains and by doing so provide new methods and tools that can be used to advance in the fields of chemical biology and drug discovery.

Chapter I.2 - The chemical space

I.2.1 Definition and scope

An almost infinite number of molecules are expected to populate the vast chemical space with the number 10^100 posited as a limit [22]. However, in biomedical sciences, the biologically relevant fraction of this space is considered to contain mainly small molecules of biological origin and designed synthetic chemicals that are able to interact with some biological elements in living systems. Different attempts have been made to draw the borders of the chemical space of pharmacological interest, and some authors have assessed a number around 10^60 molecules considering only those with less than thirty atoms different from hydrogen [23]. Additional filters, such as the Lipinski's rule-of-five for oral bioavailability [24], reduce this set of biologically relevant molecules by several orders of magnitude. Although those are the theoretical numbers, the reality is that only several thousands of molecules have been actually marketed as drugs and that the number of small molecules present in our bodies with any function is also within that order of magnitude [25]. This emphasizes the fact that the characterization of this space is an issue of utmost importance in this field as in its vast majority is entirely unexplored.



Cartoon representation of the relationship between the continuum of chemical space (light blue) and the discrete areas of chemical space that are occupied by compounds with specific affinity for biological molecules. Extracted from [26].

The chemical space must be then structured and organized in order to be efficiently explored because a random approach would be not very effective due to its size. In the discovery process, for instance, molecules with certain physicochemical properties need to be located so efforts need to be focused in the appropriate portion of the space. Characterization of the molecules is the first step in this process, and it is achieved by employing several types of descriptors based on molecular structural and physiochemical features.

One-dimensional descriptors capture specific properties of a molecule in a single value and are fast and easily computable. Molecular weight, number of hydrogen bond donors and acceptors or octanol-water partition coefficient and aqueous solubility belong to this group and some of them are usually combined to delimit the space of interest like in the above mentioned Lipinsky's rule-of-five. In a level above, two-dimensional descriptors are based on the topology of a molecule given by the atoms it contains and their connectivity. Substructure methodologies are used to compare molecules looking for common fragments [27] while fingerprint-based methods codify the molecular features into strings called fingerprints which are designed to be easy to generate and compare [28]. Finally, three-dimensional descriptors capture additional molecular information on the relative position of atoms or features in three-dimensional space. An optimal balance between computational cost and prediction performance finally determines the ultimate choice of descriptors for the particular property to be modeled [29].

Beyond these numerical descriptors, other structural approaches have been successfully applied to the mapping of the chemical space by looking for common elements across predefined sets of compounds. Recursive simplification of chemical structures by applying simple rules, like removing all side chains or turning all bonds to single bonds, will reveal different levels of structural complexity, up to the atomic framework level [30]. The organization of these emerging structures in a simple-to-complex hierarchy will allow researchers to locate recurrent atomic organizations fulfilling specific interests, serving as basis for the design of new compounds [31]. Furthermore, the application of structure-based clustering methodologies has been successfully

applied to the mapping of large chemical databases [32]. However, as emphasized above, an optimal combination of different types of methodologies offers often a better option that any of the individual methodologies applied in isolation [33].

Two main sources have been historically used to populate the fraction of the chemical space of potential biological interest. First studied were molecules naturally present in living systems, which are part of the system itself so have intrinsic biological significance, and are widely used as basic components for further structural optimization in the drug discovery process [34]. Despite a reduction of about 30% in the number of drugs based on natural products in clinical or preclinical stages in this decade [35], they are still considered cornerstones in the drug discovery process and many more are likely to be revealed when new universes, like marine life or traditional Chinese medicinal products will be studied [36] through the application of the improvements in related technical proceedings [26]. The second group is composed by those structures designed *de novo* by the pharmaceutical industry with therapeutic objectives, to try to emulate the actions of natural ligands over specific proteins. These designed novel chemical entities may be modifications of natural products or synthetic molecules. If we consider drugs marketed in the last thirty years almost a half of them come from each group [29].

I.2.2 Sources of information

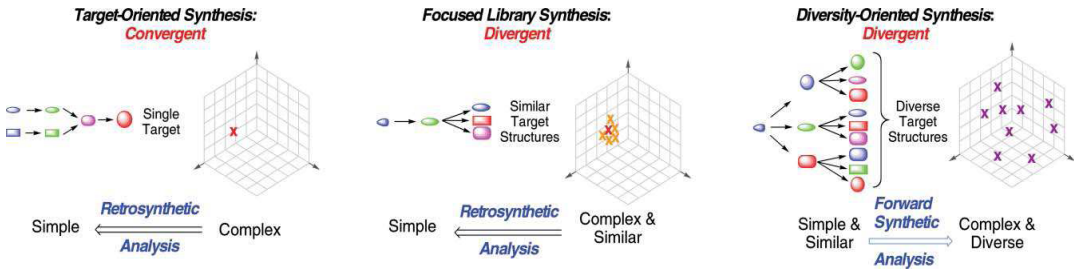
There are different kinds of databases storing information related with the chemical space defined above, but all of them specify the structures of the chemicals they contain which may have been gathered with different purposes. Beyond this, most of these sources also contain additional information about the compounds, and some of them also have valuable data related to the interactions these molecules are known to have over specific biological systems. We will now give a brief overview on the two main types of repositories of chemical and pharmacological data relevant to our purposes, in this order, plain chemical libraries and annotated chemical libraries.

I.2.2.1 Chemical libraries

Plain chemical libraries are collections of chemicals with associated information related to their intrinsic properties or marketing data. They are generated based on organic and medicinal chemistry and designed with specific purposes so the molecules they contain will have certain characteristics that fulfill their expected requirements. The main uses of these collections are to improve efficiency of drug discovery, to computationally describe drug-like molecular properties, to identify promising target candidates (biomolecules, proteins, enzymes, receptors) or promising ligand candidates (small organic synthetic or biosynthetic compounds) among others [37].

Targeted chemical libraries are built around a specific concept and have molecules with delimited characteristics that are related with certain properties, for instance, having affinity for a protein target or a family of targets like kinases or G protein-coupled receptors [38] or passing the brain-blood barrier [39]. There are different methodologies used in the design of such libraries that explore the borders of the chemical space of interest, generating diverse sets of molecules inside the stated property ranges based on biochemical and proteomical knowledge [40]. Moreover, *in silico* profiling of compounds has also been posited as a promising way of assessing the potential scope of a targeted library [41].

On the other hand, diverse chemical libraries are used to scan big portions of the chemical space in search of compounds with the required characteristics. In combinatorial libraries, for instance, compounds include modifications of a common scaffold with the addition of selected functional groups in different positions in order to maximize diversity while covering several properties of interest [42]. As the chemical space is so huge and the costs of experiments are high, having a well represented chemical library is also a key issue in the drug discovery process and different methodologies are used to design them, for instance, based on natural products modification [43]. Nowadays many strategies for counting the diversity of a chemical library are being developed and scaffold analysis is emerging as one of the most intuitive and successful [44], alongside with other graph-based methodologies [45].



Schema of the different strategies used in chemical library design. Extracted from [46].

Larger libraries emerge from the union of smaller more specific ones so most of these collections do not fully cover the chemical space and are biased for historical reasons. Although there are dozens of commercial suppliers of such libraries, an analysis showed that the overlap between them is small, meaning that none of them is able to cover a big fraction of the chemical space [47]. Independent from commercial purposes, several huge molecular repositories exist from the academic side. Chemical Entities of Biological Interest (ChEBI) [48], developed by the European Bioinformatics Institute in Cambridge, is a freely available dictionary of molecular entities focused on 'small' chemical compounds [49]. Another of these repositories is PubChem [50], a component of the NIH's Molecular Libraries Roadmap Initiative, hosting two large databases named PubChemSubstance and PubChemCompound with 62 million and 26 million records each as of today. Furthemore, PubChem presents a wide number of features and has recently started another database named PubChemBioAssay to store activity screens of chemical substances from PubChemSubstance [51].

In the ambit of drug discovery, compounds in these libraries may be screened to select those fitting the expected requirements to go into further optimization steps, so they use to avoid structures or functional groups that are known to be reactive or to have undesired ADMET (absorption, distribution, metabolism, excretion and toxicity) properties that could be a handicap in later stages of the process. Because there are many different computational

methodologies available to do this screening *in silico* with a sufficient success rate, this has become a convincing way of lowering costs on the overall drug discovery process [52].

I.2.2.2 Annotated chemical libraries

These sources play a key role in the integration of the chemical space with the other spaces involved in biomedical sciences and drug design [53] given that beyond the detailed structures and other intrinsic data, they contain information on how small molecules interact with macromolecules. The most usual contents are the response data obtained in experimental assays about the effect of a determined compound over an organism, tissue, cell line or specific protein, as well as information about the ADMET properties of the compounds or other functional data. We will call annotation to any piece of information that relates the compound with medical data or specific entities from any other space relevant to biomedical sciences. This information is usually published in peer reviewed journals and then collected into large databases but in some cases it comes directly from series of assays, the results of which are available to the public. In this thesis, we focused on the chemical libraries relating compounds to the protein targets they bind because they play a key role in current drug discovery, allowing researches to detect the relevant features enabling specific interactions. Once these are recognized, new compounds with similar characteristics may be explored *in silico* to predict their theoretical binding affinities, as we further analyze in chapter **IV.4**. Then, the power and weakness of these chemical libraries is that they usually contain sparse data, as it comes from different original sources, giving place to diverse collections of compounds covering dissimilar targets with highly incomplete activity data. Due to this diversity, researchers can build models based on the known interaction data which may cover many targets with unrelated active ligands, giving them a positive prediction success average rate. However, the lack of completeness does not lead to exhaustive models, generating errors caused by the presence of abundant unknown response data.

Among the annotated libraries relating molecules with different protein targets we can find examples with a specific scope like IUPHARdb [54] which, containing more than 2000 ligands, is the official database of the IUPHAR Committee on Receptor Nomenclature and Drug Classification and is devoted to the study of compounds interacting with G protein-coupled receptors and ion channels. Also with a specific approach, in this thesis we collaborated on the generation of a nuclear receptor directed annotated library, as will be further elaborated in chapter **IV.1**. On the other hand we have huge chemical libraries like ChEMbldb, an initiative of the European Bioinformatics Institute (EMBL-EBI), which is part of the European Molecular Biology Laboratory (EMBL) [55]. With more than half a million compounds, this is the biggest publicly available resource of quantitative and bioactivity data on the interaction between proteins, cells, organisms and small molecules as it collects more than 2.5 million annotations [56].

There are other models of annotated libraries like the represented by PDSP. This public database is focused on novel psychoactive compounds for which pharmacological and functional activity data is provided through screening on cloned human CNS proteins. Its most important characteristic is that new data, mostly binding affinities coming from biochemical assays, is generated and only a small amount are taken from literature [57].

The biggest annotated compound libraries are however proprietary and some of them, like Wombat [58] or MDDR [59] contain huge amounts of data collected from literature. Wombat is, for instance, a proprietary database developed at Sunset Molecular Discovery that as of January 2009 contains almost three hundred thousand chemicals interacting with nearly two thousand proteins. All that information has been extracted from more than fifteen thousand papers published in the last thirty years in peer reviewed journals of the field. Besides this, the new data generated via high-throughput inside the pharmaceutical industry is also proprietary, and the major resources of response and functional data for small compounds are encapsulated in private data libraries of leader companies.

I.2.3 Data integration

To have a complete overview of the data contained in several chemical libraries it is mandatory to deal with the standardization of the different formats that those libraries will present and also to identify compounds in all libraries to know if they are present in the rest of them.

First of all, the compound structures can be specified in different formats, namely, smiles, sdf, mol2, rdf, to name a few. All these formats provide information about the atoms present in a specific molecule and the bonds that link them, but not all of them give the coordinates of these atoms. Furthermore, some chemical libraries will specify the 3D structure of the compound, whereas others will just rely on the 2D topology, an aspect worth considering depending on the descriptors one needs to compute. Of course, something we need to have in mind is that the more information we encode in the way we describe the structures, the bigger storage capacity and more time consuming processes we will have to deal with, so an optimal balance between detail and management is required. One could use more specific descriptions like the sdf coding for annotated chemical libraries, which are smaller and used in processes like modeling which could require much more detail, while other much compact formats like smiles could be used for commercial chemical libraries containing millions of compounds to be screened every once in a while.

A decade ago, few methodologies were available to uniquely identify chemical structures in large databases. Today, different approaches have been taken to solve this problem. The IUPAC, International Union of Pure and Applied Chemistry, was working from 2000 to 2005 on the IUPAC Chemical Identifier Project to establish a unique label, the IUPAC Chemical Identifier, which would be a non-proprietary identifier for chemical substances. To get the InChI of a compound, the process starts with the normalization of its structure, removing redundant information, and then it follows with the canonicalization into a unique form, such that any representations of this compound would collide into a single unique graph representation. This canonical representation is serialized into a textual form called InChI containing six different layers of information related with the structure, the charges, the stereo chemistry and

other chemical features of the compound. When this InChI was found to be too long to be efficiently searched and stored, the InChI keys were developed. For their calculation, InChI string is hashed into a 25 characters length alphanumeric code were 14 of this characters result from the connectivity information of the InChI, followed consecutively by a hyphen and 8 more characters resulting from the remaining layers of the InChI. After this, a single character indicating the version of InChI used and a single checksum character are found. The chance for two different compounds to have the same InChI key is estimated in 1.3 for every 10^9 compounds, meaning a single collision into 75 databases of 10^9 compounds each [60].

At the same time, the chemical structure code (CSC) was developed in our lab. Based in structural hierarchy, this code consists on a unique six-level code for each molecule where each level encodes for a sub-structural characteristic, going from the most generic one to the final unique identifier of the molecule. The first, second and third levels are integers specifying the number of rings in the largest ring system present in the molecule, the number of bonds in the longest path and the number of branching points in the longest path, respectively. The fourth, fifth and sixth levels are unique eight-character hash codes for the molecular framework, scaffold, and the complete molecular structure, respectively.

When any of these codes is calculated for all the molecules in several chemical libraries, information on how many common compounds they have is revealed when previously it was hidden behind the different compound identifiers and structural representations. Once we are able to collapse all repetitions of a compound into a single chemical graph entity, the work continues with the standardization of the information that each source may contain related to the physiochemical and pharmacological characteristics of that compound. In this step, the pieces of information coming from different sources need to be converted into the same units when possible, and when these data are being collected, one has to be aware that errors occur and has to apply a methodology that removes wrong specific data once it is observed to be non reliable or represent an outlier point in a set of data.

Several software solutions are available to translate compound libraries from one format to another and to write additional information when the format allows that to be done. Some of these programs like OpenBabel [61] also allow the user to calculate compound identifiers and simple chemical descriptors. In this respect, of mention is also the open policy to the academic community adopted by ChemAxon [62], which provides a good list of useful applets for developing tools in chemoinformatics and life science research.

Chapter I.3 - The biological space

I.3.1 Definition and scope

For the sake of simplicity, we define the biological space as being composed of all proteins present in any living or extinct organism on earth [63], the union of all specific proteomes. The study of proteins is of utmost importance for biomedical sciences as they are the machinery of living systems. Some of them perform mechanical and structural functions, like in the muscles and the cytoskeleton, while others are specialized in catalyzing reactions that by natural means would occur at a much lower rate and are called enzymes. A third group of proteins would be involved in regulating cell signaling and cell cycle processes and would include transporting proteins and channels, along with different kinds of receptors. All these proteins are then characterized by their function and those with similar functions will usually tend to be phylogenetically related.

Amongst all this wide biological space we focused our work on the druggable portion as this is the one that will allow us to link it with the chemical space of small molecules. These proteins, not related with structural or mechanical functions are regulated by endogenous ligands that bind them in a specialized pocket that is generated when protein is folded to a 3D structure. The modulation of the behavior of a certain protein may have a great significance on the cell function and this can be achieved by designing chemicals with the features required to fit in the protein active site. To be able to do that, we need to gather as many information as possible about the proteins inside our scope, including sequence, functional, phylogenetic and physicochemical data as well as data about their known interactions with ligands or the large scale effects that arise from the modification of their behavior.

There are several families of proteins that are included into this biological space. The biggest and most well known belongs to the enzymes group, which has been classically the most studied, while other families such as

the nuclear receptors, the cytochromes, the G-protein coupled receptors and the different kinds of channels and transporters have attracted researchers' interest for specific reasons. While the complete human proteome is currently estimated to have more than twenty thousand [64] different proteins, we will focus only in the small portion that has been successfully characterized. More than 5000 enzymes, 5000 channels and transporters, about 1000 G-protein coupled receptors and a hundred nuclear receptors have been detailed at different levels and organized in several classifications.

In terms of their interaction with the chemical space, proteins need to be characterized depending on the chemical features they expose to the compounds they bind but this is difficult to achieve due to the complexity of this issue. However, as it is expected that proteins with similar functions or with common phylogenetic origins will probably have similar features in their activity pockets, some classification systems based on different data, either functional and/or structural, have been successful to a certain extent.

I.3.2 Sources of information

Proteins may be the most studied entities in biomedical sciences, as they are the elemental machineries involved in all biological processes. In this section we will have a quick glance at the different types of information available about them and their storing repositories. Although we have divided this data in functional, structural and phylogenetic, these three elements are intrinsically related and one may be used to infer or study the others.

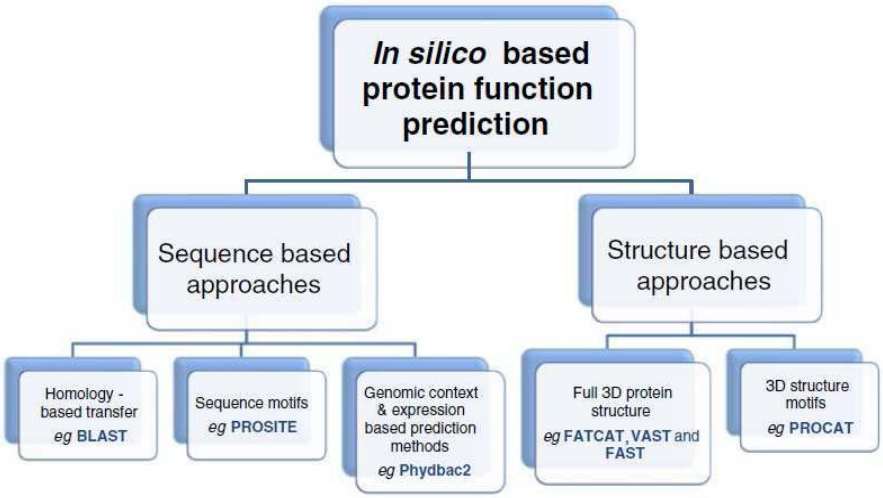
I.3.2.1 Functional data

The classical way to characterize a protein is based on the function it develops in nature although the definition of what is considered as such is not clear because protein capabilities are related to actual molecular functions, cell processes and cellular location [65]. Behaving as functional and structural units, domains are conserved portions of proteins that have specific function

and fold independently from the rest of the protein. They can be combined as basic pieces to generate new proteins, hence being considered the units for protein evolution with around 150 domain families present in all kingdoms of live [66].

Several initiatives also consider these domains as the units for functional annotation and, as most eukaryotic proteins have multiple domains, they can be related with different functions at the same time [67]. However, this approach is limited by the consideration that the researcher has to decide when a certain domain belongs to a specific type, and that is easy for very similar domains but becomes more and more difficult the more different they are, for instance in the case of proteins with a distant common ancestor. Depending on the methodology used to cluster the proteins on basis of the domains they contain, different classifications account from one thousand to several thousand different families [68]. CATH, for instance, is a well known curated classification of protein domain structures containing 14.500 domains divided in 1.150 sequence families grouped in 226 superfamilies for a total of 124 defined different foldings [69]. With a sequence based approach to the same issue, pFam, an initiative of the Sanger Institute, specifies as of today almost 12.000 families grouped in clans by sequence similarity or structural similarity [70].

The prediction of the function of a protein is a key issue in the genomic era and is mainly based in two methodologies emerging from the idea that sequence, structure and function are intrinsically related. First one is sequence comparison of its encoding gene with those of other proteins of known function. An approach that, while simple as an idea, has resulted in a proliferation of genetic sequence alignment methods because of its complexity [71]. The second group comprises structure based approaches that compare the final 3D protein foldings instead of the genetic sequences and have also given place to a multitude of methodologies [72]. Moreover, the combination of several techniques on a certain prediction is likely to increase the success rate.



Schematic overview of in silico-based protein function prediction methods. Extracted from [73]

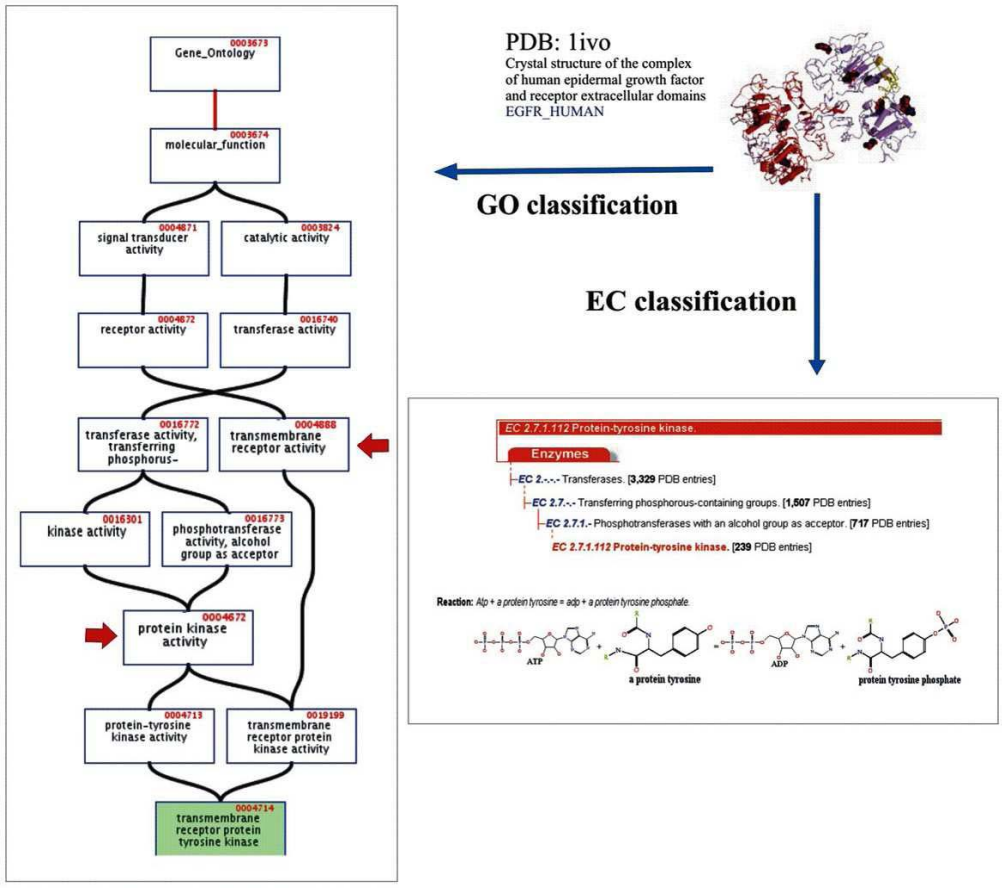
Once the protein has been annotated to a set of functions, it has to be organized along with other members of its functional class. This was addressed by different initiatives that worked on specific families of proteins, giving place to well defined but not homogeneous classification systems. The Enzyme Nomenclature Committee developed the Enzyme Commission number (EC number) [74], a standardizing and hierarchical classification scheme for enzymes based on the reactions they catalyze. Although EC numbers are associated with a recommended name for the respective enzyme, EC numbers do not specify enzymes but catalyzed reactions. This means that different enzymes that catalyze the same reaction will receive the same EC number. Up to now, this classification accounts for 4150 different entries which represent a slightly bigger number of real enzymes.

Another well known classification system is the one stated by the IUPHAR, today called International Union of Basic and Clinical Pharmacology. Centred on channels and transporters, divides almost 400 of them in three main classes depending on its function and morphology and then hierarchically classifies proteins of this type in sublevels under each of these classes. These proteins cover up to a third of the targets of marketed drugs as they are highly

relevant because of their key role in cell to cell signal transduction at nervous system, acting over a wide range of signalling pathways. The elaboration of a classification system for such a diverse family of proteins has taken more than 15 years of official reports in the journal Pharmacological Reviews [75].

As both of this classifications give proteins a hierarchical code, we can collapse the groups of proteins at different points so that their properties at each level can be characterized hence allowing us to detect common features. However, some of the classifications are very specific in the definition of the groups while others are much less detailed, depending on the number of proteins they are dealing with, and this is something to have into account when integrating data from different sources.

Recently, a new approach to the functional classification of proteins based on GO (Gene Ontology) is gaining adepts, being the main difference that instead of a hierarchical organization tree, different semantic terms related with biological processes, molecular functions and cellular components are linked with the protein in a networked fashion. With this extensive annotation process requiring expert curation, researches can increase the detail of specification of the characteristics for each entity they define and may infer new functional annotations for genes and proteins [76] although sometimes this complexity will lead to inconsistencies [77].



Hierarchical and network-type functional annotation of proteins. Extracted from [78]

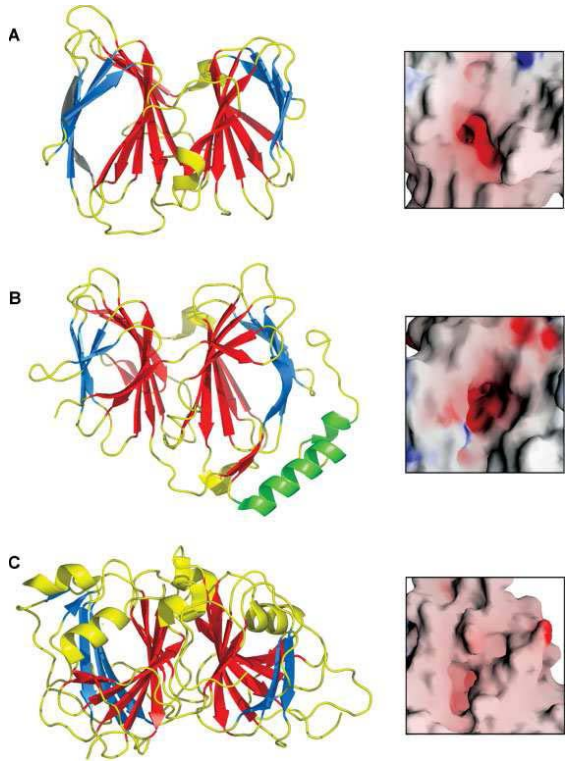
All these methodologies have generated several functional classification schemes which are summarized in web repositories like PIR (Protein information resource) [79] or KEGG (Kyoto Encyclopedia of Genes and Genomes) [80].

I.3.2.2 Structural data

The properties of a protein and the actions it is able to perform depend on its 3D structure resulting from the processing and folding of its specific amino acids sequence. In principle it is possible to say that knowing which gene

codes for a protein we should be able to know which postrancriptional processes it would suffer and how the final structure would look like. However, the complexity of the mechanisms that lead to the final three-dimensional shape of the protein make that objective very difficult to achieve through the different modeling methodologies we have nowadays, like DNA based comparative modeling or amino acid based fold recognition [78] also called homology modeling. Other template independent methodologies for the prediction of protein structures have been reported to achieve lower accuracy levels but are promising techniques for a near future [81]. However, the need to relay on experimentally determined structures is unavoidable in most cases but, as we cannot have them for all existing proteins, we need to solve as many and diverse as possible in order to be able to model the remaining ones [82].

Actually, it has been observed that although the number of genetic sequences may be almost infinite, the number of different 3D foldings is estimated under 10.000, having that a few of them comprise the vast majority of sequences while most of them have few representatives [83]. For instance, an analysis of 250 proteins not belonging to any known group revealed that two thirds of them had 3D structures fitting already known foldings [84]. Furthermore, having access to these three-dimensional structures of proteins has given place to the development of different computational methodologies capable of modeling with reasonable accuracy the binding of ligands into protein cavities at a low cost [85].



3D folding of three structurally similar proteins and detail of their active pockets. Extracted from [86]

In the last years, there has been an exponential increment in the number of structures detailed due to the advances in different techniques for protein synthesis at large scale as well as in X-ray crystallography, NMR spectroscopy and lately electron microscopy. X-ray crystallography is the first developed and most used method, and all steps have achieved a high level of automatism except from protein expression and crystallization which seem to be in need for the implementation of new tools and protocols [87]. These improvements generated a lot of information and raised the need of properly storing and organizing all this new data in a standardized and easily accessible manner which was the main objective for the development of the PDB, the protein data bank, an initiative of the Brookhaven National Laboratories.

Although the PDB started as a closed project for experts in structure

research in the early 70s [88], the exponential increase on the number of structures in the 80s and changes in the view of the issue of data privacy inside the community, gave this initiative the shape it has today [89]. With more than sixty thousand structures as of January 2010 and more than seven thousand new ones deposited each year, including DNA, RNA, proteins and mixed complexes, this is the reference repository for protein structures. Beyond this information, which is encoded in PDB files that have become a standard for the representation of this kind of data, this resource contains also chemical, biological and bibliographical related information.

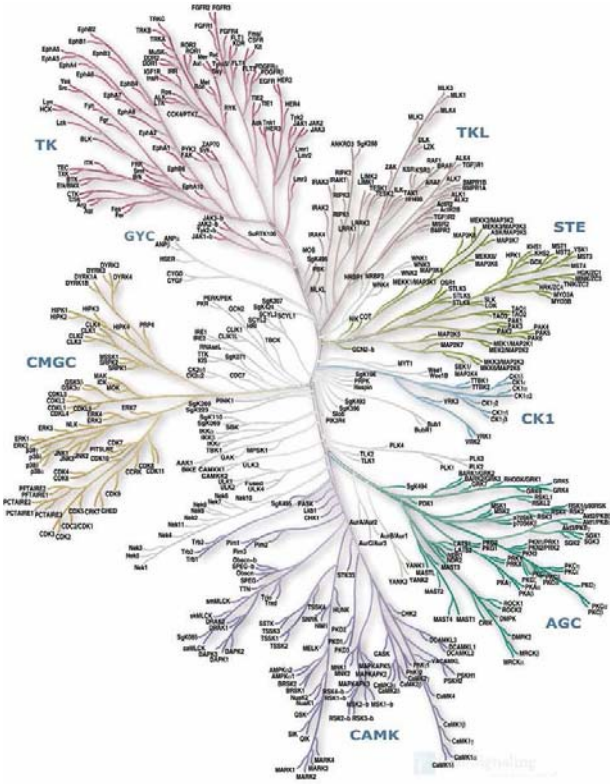
Assuming that this data is of high need and utility for biomedical sciences one would expect to have a normalized representation of the different protein families in repositories like PDB although this is not even by far close to reality. The population of structures is strongly biased towards targets with classical known therapeutic interests, mostly enzymes including kinases and proteases, in detriment of others that have recently emerged as relevant families like the aforementioned nuclear receptors, transporters, channels or G-protein coupled receptors. The web application FCP presented in this thesis is a tool dedicated to the analysis of the current distribution of solved structures among different target families, including graphic and numeric data on the coverage and bias of the different groups of proteins at all levels. The systematic approach of FCP is based in the organization of the proteins in hierarchical nomenclature systems in order to analyze the number and distribution of the structures at each level. This will be discussed in detail in chapter **IV.2**.

As this web tool shows, it is clear that this bias is trying to be corrected as new methodologies try to cover empty areas of the biological space [90]. However, the bias of the structural coverage across different groups of proteins is remaining constant while the number of new groups characterized grows almost exponentially, meaning that even when structures are being solved for previously empty groups, more are being accumulated for those that are already well known because they are easier to crystallize [91].

I.3.2.3 Phylogenetic data

In the evolutive history of every species, protein families are populated with the addition of new proteins as a result of two different processes. The first one, that gives place to paralogous proteins, consists in a duplication of the initial gene inside the genome of the species resulting in a copy that keeps its prior function while the other may diverge by accumulating random mutations. In the other hand, orthology occurs when a species differentiates in two and the proteins of the common ancestor evolve independently to derived proteins. Beyond the use of this knowledge in the mapping of the genome for the discovery of new unknown genes [92], the phylogenetic relationships between proteins are the basis for sequence homology comparative modeling and hence for structural and functional similarity annotation, given that the more related two proteins are, the more similar their sequences, three-dimensional foldings and active sites will tend to be [93]. Furthermore, structure based phylogenies have proven to be very useful because structures are more conserved than genetic sequences [94] so there is an extra margin for success and, in combination with the comparison of related sets of proteins from different proteomes, this may enable the functional annotation of uncharacterized groups [95] as well as the development of several evolutionary analysis methodologies [96,97,93].

A successful initiative in this field is the characterization of the human kinome achieved by Manning et al. in 2002 [98]. Kinases are specialized enzymes that modify the behavior of other enzymes by means of phosphorilation and are considered potential targets of research in biomedical sciences with the idea to indirectly regulate other therapeutically relevant proteins involved in critical cellular processes. About 500 known human kinases were clustered into 9 main related groups characterized by the different domains they present.



Representation of the human kinome. Extracted from [98]

Other researches focus their work on the nuclear hormone receptors, the main class of transcriptional regulators in animals with high clinical significance due to their involvement in different physiological processes at the DNA transcription level. A natural functional classification of this group of proteins would specify four groups depending on their DNA-binding properties and dimerisation preferences [99]. Instead of that approach, in 1.999 the Nuclear Receptors Nomenclature Committee established a compilation of the 48 known human nuclear receptors [100] classifying them hierarchically by sequence homology into seven main subfamilies.

Bridging functional and phylogenetic information, TCDB is an initiative of the University of California San Diego aiming to establish a classification scheme for membrane transport proteins known as the Transport Classification

(TC) system. Defining a five levels classification hierarchy, this system is analogous to the previously mentioned for enzymes, being the main difference that the definition of the groups combines information related with the specific function of the transporters and the phylogenetic relations they have [101]. Currently this database contains 3000 protein sequences divided into 550 different groups where orthologues and paralogues collapse [102].

As we can see, the characterization of the phylogenetic trees behind the therapeutically relevant protein families may bring light to some key aspects of the protein function and its interaction with the chemical space and, because of this, all of these initiatives focusing in different portions of the proteome need to be integrated to have a complete and comprehensive overview of the biological space.

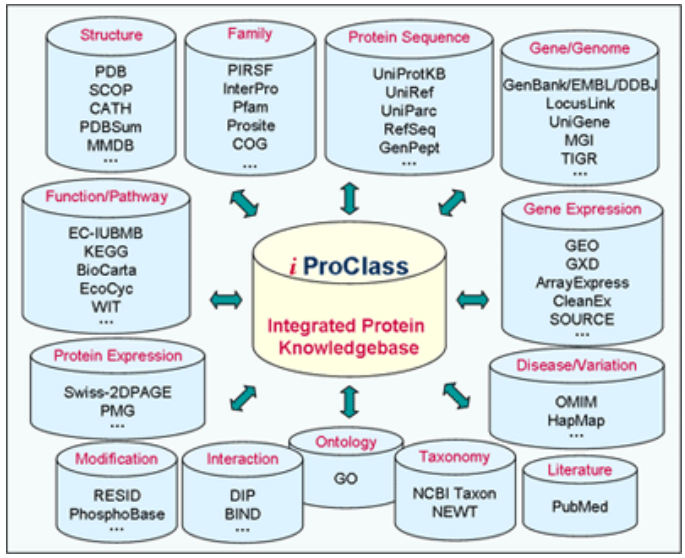
I.3.3 Data integration

The key issue at this point is linking the information coming from different sources that refers to the same entity in the biological space in order to have a complete view of its properties and features. This is not an easy job as different synonyms or denomination formats are used in literature and by the different classification systems to refer to a certain protein or group of proteins. Moreover, before the existence of any even vague guidelines for protein nomenclature, scientists referred to the proteins assigning unrelated names, contributing then to arriving at the present situation where not only a protein can have up to thirty different names but also a single name can refer to different proteins inside a specie or in different species [103]. On top of that, scientists face the problem of how to identify isoforms, mutations or other kinds of variations of the same initial protein and how to identify related proteins across species.

In an ideal situation when a new protein is described, the name given by the author should be addressed to a nomenclature expert to be reviewed in order to assure that it is syntactically correct and unambiguous. Nevertheless, since the development and optimization of high throughput methodologies for

functional annotation of proteins, the manual revision of the names has become an unaffordable task and as a result of that there has been a proliferation on nomenclature guiding initiatives that could lead researchers to give in first place an already correct name for the proteins they are describing. This used to be a hard task but recently new tools have been released to assess scientists in the application of the existing rules and guidelines to generate standardized names [104].

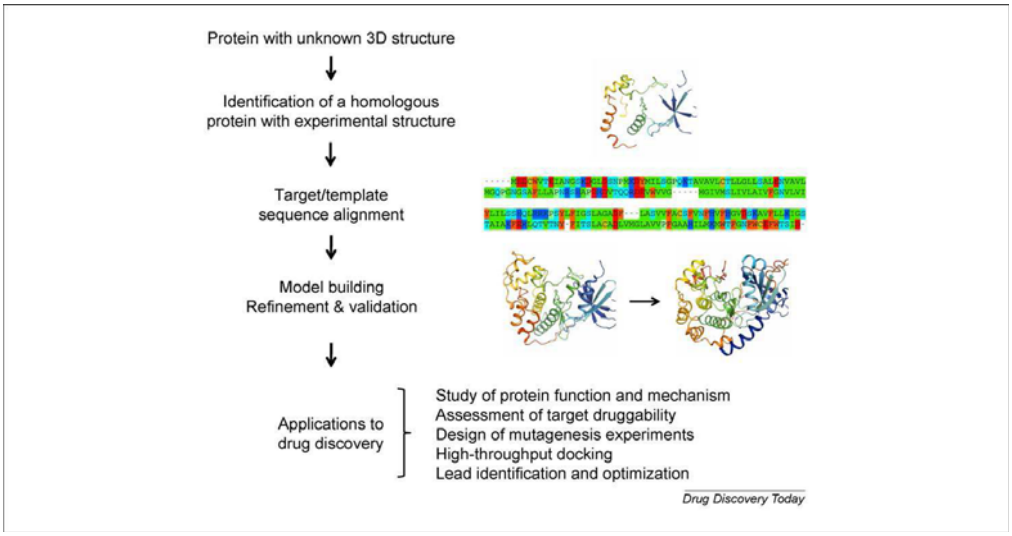
An outstanding initiative in this field is The Universal Protein Resource (UniProt) project, an effort from groups of the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR) which main mission is developing and maintaining a protein knowledge database with integrated and standardized information including the unique identification of individual proteins. Stable and carefully curated with in silico methods first and then by experts in a second phase, the data is freely available on the web through a very complete site with a simple user friendly interface. Researchers can browse the proteome of different species including three eukaryotic complete proteomes, like the human, and retrieve useful information related with each protein such as sequence, clusters of similar proteins around it or a list of synonyms [105].



iProClass integrative architechture. Extracted from [106]

Once we are able to identify a single protein with a specific name and we have available a list of synonyms that are known to refer to the same entity, we can start mining the different databases and information resources to annotate all related information. iProClass, linking UniProt entries with valuable information from more than 90 biological databases of diverse content, is another initiative of the Protein Information Resource group and an outstanding example of this approach [106].

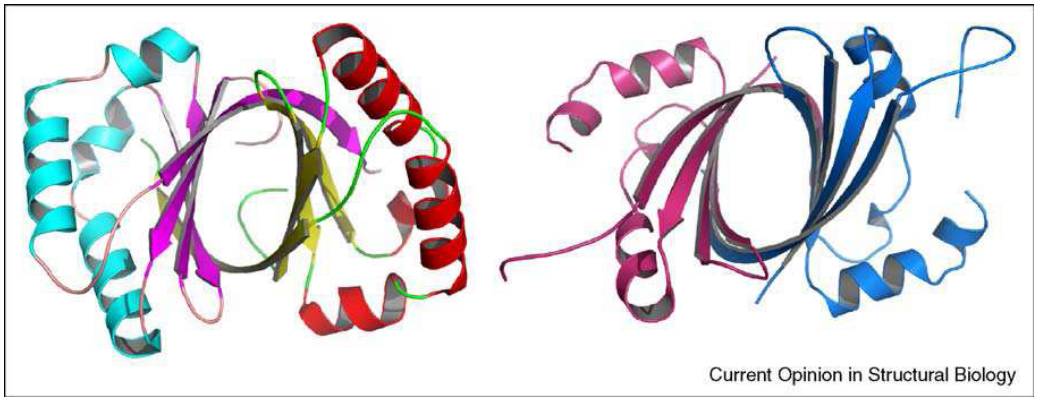
With important implications in the drug discovery process, the combination of structural information with functional data, allows researchers to define the motifs of the activity sites and their three dimensional disposition for proteins of interest, hence enabling the modeling of their possible interactions with ligands [107].



Outline of the homology modeling process and its applications in drug discovery. Extracted from [108]

Furthermore, phylogenetic data from sequence homology analysis is used to make predictions on the function of proteins coded by new identified genes [109] and also to compare the amino acids sequences of a given protein with others of known three dimensional structure in order to predict the actual folding of the uncharacterized protein [110, 111, 112]. However, despite all

these efforts, there are yet many problems to solve in this field for instance when proteins with high sequence similarity have been found having different foldings and functions at the same time that promiscuous structures have been detected to be useful for different functions [113] and genes with no sequence similarity have been found to produce proteins with the same 3D foldings [114].



Similar 3D structures of two novel monooxygenases with no sequence similarity. Extracted from [114]

All this studies can be taken to a higher level by applying the different classification systems to study the common properties and characteristics of proteins at different levels inside protein families. It is clear why an integrative and synergetic approach is necessary to have standardized annotation and nomenclatural organizations for a systematic and complete view of the biological space and its interactions with other spaces.

Chapter I.4 - The phenotypical space

I.4.1 Definition and scope

The word phenotype refers to the observable physicochemical or biological characteristics of an organism. For the sake of simplicity, we well define the phenotypic space as being composed of the different variations among individuals produced by the alteration of the normal phenotype. In the scope of our work we studied the set of variations that are originated by the alteration of the normal functions or processes of the molecular machineries by different means, focusing in the addition of compounds that are known to interfere with the behavior of specific target proteins.

When the alteration appears due to the abnormal function of one or more proteins and is detected as a disease, drugs are administered for its treatment with the idea of reverting the wrong processes to the normal state. However, one of the problems we may encounter is that this compounds or their derived metabolites usually bind more than one target protein, the computed average is expected to be around six [115,116], giving place to the desired effect but often to other secondary alterations which are called adverse drug reactions or side effects. The differences with the classic “toxic effects” of chemicals or with allergic responses to drugs are that here there is no dependence on the concentration of the compound and that the effects are pharmacologically mediated by proteins not related with the immune system. Furthermore, the main characteristic of these side effects is that they do not appear on every patient but they are related to certain risk factors like age, gender, multi-drug administration, concurring diseases, ethnic and genetic differences and other pharmacokinetic factors [117]. Despite the fact that no drug is free to cause harm, there are some groups of drugs which have been classically related with adverse drug reactions like antibiotics, diuretics or opiates [118].

In a broad sense, side effects produced by drugs are complex phenomenological observations that can be attributed to a diverse set of causes. The above mentioned polypharmacology is one of them but researchers also

consider drug-drug interactions, protein aggregation and other pharmakokynetic problems [119]. The effects of this drug toxicity can be grouped in three main categories including cell death or tissue injury, functional alteration and cancer [120]. As large scale phenotypic responses mediated by drugs we can consider from hepatotoxicity, maybe the most studied in toxicogenomics because of the critical detoxifying functions of the liver [121], to ventricular arrhythmia or aplastic anemia [122]. Adverse reactions related with already marketed drugs increased 2.7 fold from 1998 to 2005 as stated by FDA medical reports [123] and are the main causes for market drug withdrawal [124] hence being of utmost economical relevance from the point of view of pharmaceutical industry.

Being the drug discovery process so long, complex and expensive, the need to asses as early as possible the actual dangers behind drug candidates has become of primary interest and many efforts from different approaches are exploring this issue at the moment. This can be considered at first level by looking for repetitive structures across the chemicals or their metabolites that are known to have associated toxicity [125] or the possibility of covalent binding [126, 127]. Beyond that, other valid approaches not ligand centered are the study of toxicity propagation through the effects over metabolic pathways [128], and recently, also transcriptomics and metabolomics based studies. Indeed, a recently emerging field called toxicoproteomics or computational toxicology aims at the elucidation of toxicity mechanisms and other related issues by the implementation of quantitative proteomics technologies [129], genomic based approaches [130], organ specific toxicity studies [131] and different computational methodologies [132, 133] which try to predict drug related toxicity and adverse reactions. As an example of this, a methodology aimed to analyze the intricate mechanisms behind protein mediated adverse drug reactions was developed within the scope of this thesis; see further details in chapter **IV.5**.

But until the moment when predicting algorithms and models are ready with full capabilities, the search for adverse drug reactions relays on data generated experimentally either in *in-vivo* or *in-vitro* assays. Furthermore, these assays may be performed in other species than human and when data is

extrapolated to the real administration circumstances, scientists have to deal with the fact that those *in-vitro* and/or non-human models cannot be completely relied, as many cases of misleading results have been discovered when the drug was tested with humans. For instance, the success rate of prediction protocols based in animal testing in reproductive toxicity experiments has been posited to be of about 60% with a 40% rate of false positives, so the search for new pharmacologic testing methodologies adapted to the current state of the art is, for some experts, of high need for different reasons [134].

I.4.2 Sources of information

As we have just stated, several toxicity testing methodologies including animal tests, cellular assays, microarray studies and other experimental procedures, are the main original sources of preclinical toxicological data at different scales. On the other hand, medical reports contain very interesting information, the side effects detected across the population when the drug is already marketed and administered. However, being this information so varied and disperse, there is a need to rely on other sources that store and structure this raw data in higher semantic levels, more suitable to our purposes. Again, the implementation of new computational methodologies and the integration of different disciplines have been required and more than 50 repositories have appeared containing some kind of elaborated toxicological data [135,136]. One of these is the Therapeutic Target Database (TTD), developed at the University of Singapore with the objective of gathering information from literature into an interrelated database with clinical, metabolic and drug data for almost two thousand target proteins [137]. This information is, in addition, referenced and linked to other trusted databases as an example of an integrative approach.

Scientific literature, is however one of the main resources of drug related toxicity, for instance with the study of the relation of certain proteins with specific diseases. This is a key piece of the pharmaceutical knowledge as identifying the element which abnormal function is causing a disease or side effect is the first step towards the correction of that deviation to the desired

state. In the case of drug adverse reactions, this kind of data is nevertheless too incomplete to account for the totality of the lately observed alterations due to the simple reason that many side effects are not expected or even detected in preclinical stages. Despite those limitations, a relevant initiative in this field is HSDB, the Hazardous Substances Data Bank from the National Library of Medicine [138], which retrieves information from literature as well as technical documents and books to organize it in an integrated database. In the end, the information is reviewed by experts and opened to the public in a web site where a compound can be searched to find related clinical and toxicological information. From the point of view of the study on the impact of genetic variations in humans based on the differences in drug response phenotypic data, PharmGKB links drugs with diseases either by reviewing literature or by the study of the metabolic pathway interferences the drug is known to cause [139

The other main source of information is found in drug leaflets and is freely available in different web repositories as well as in the manufacturers' websites. This data, in opposition to the previous, is an overrepresentation of the possible adverse drug reactions that may be related with the compound. This happens because companies tend to overestimate the problems that could arise from a given drug in order to increase security for both the patient and the company. The details on this leaflets also depend on the country were the drug has been marketed and variations will exist due to legal and social issues.

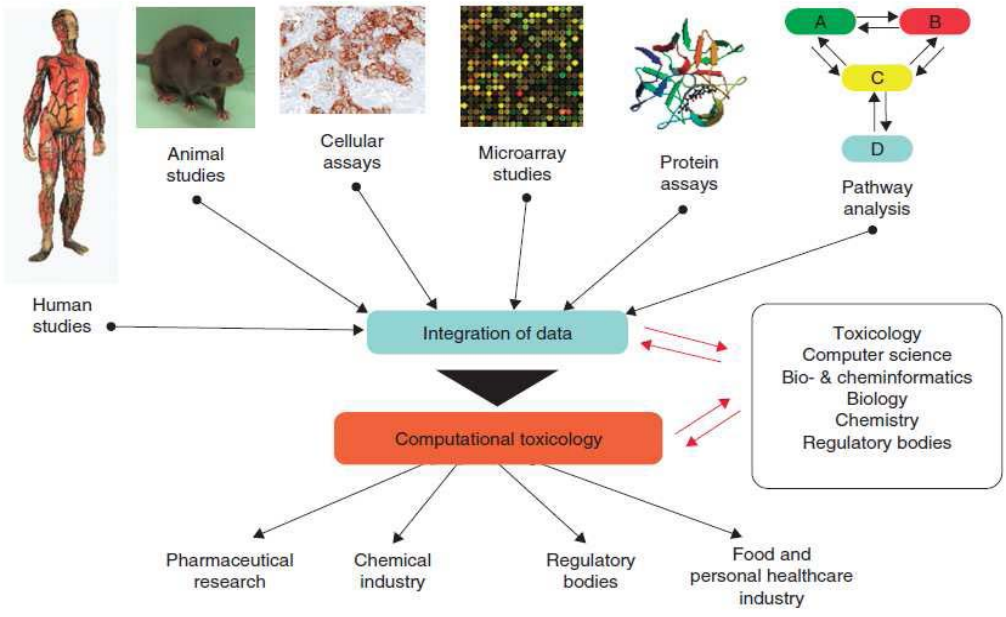
Some of the more relevant repositories of package inserts, besides the manufacturers' websites, are Dailymed [140] and Medlineplus [141] both from the National Library of Medicine. The first is focused in storing and organizing them in a searchable web database and the second one, with a wider approach, includes also other health topics as well as medical and clinical information. Other web repositories like Medscape [142] or Drugs.com [143] give reviewed advice to the open public on the use of therapeutic drugs but also contain information on their prescription and the possible secondary effects they may induce according to the details of the leaflets.

Working with this material, Campillos et al. collected data relative to the side effects of almost 900 drugs and used this information to predict and

confirm new interactions of those compounds with targets that had not been stated before [144]. Manual revision was needed to work with this kind of data and statistical analysis were performed to differentiate which side effects were significantly annotated to a certain drug from those that were appearing either because they were very common effects from many disorders or because they were artifacts generated by the overrepresentation in the package inserts. The final compilation of side effect profiles is freely available on the SIDER web site [145].

I.4.3 Data integration

Useful information coming from the different aforementioned sources needs to be gathered to make a deep analysis of this field but that is not the scope of this thesis which is focused in linking large scale effects to compounds and proteins.



Toxicological data flow and related disciplines and users. Extracted from [132]

From the proteins point of view, merging of literature relevant information with other structural, functional and pharmacological data is a hard task but when accomplished is a perfect example of the synergy expected to arise from an integrating systematic approach. As an example of this, Cases & Mestres collected a list of more than two hundred cardiovascular related targets from literature, and applying a chemogenomic approach, they were able to link them with a set of almost 45.000 compounds. Further exploration of the pharmacological profile of these ligands lead to the identification of another four hundred target proteins that extended the previously described biological space either as main cardiovascular targets or drug related secondary off-targets [146].

On the other hand, when researchers try to integrate drug centered information, two main difficulties arise. The first obvious task is related with linking the compounds in marketed drugs and their characteristics across the different formulations and brand names, and the second problem resides on the identification of the adverse drug events that refer to the same phenotypic variation but are named with a diverse set of medical terms.

Several resources on this field have tried to tackle these issues along with the integration of different data into centralized database architectures. Drugbank is the most cited repository for drug related information [147] joining chemical, pharmacological and pharmaceutical data for about 4.500 drugs including target information such as protein sequence, structure and involving pathways. Inside the drug cards displaying this data, a list of synonyms and brand names is given for each drug. A yet more extensive list is available in PharmDB, another initiative that by mining the recent literature relates drugs, targets and diseases based on an integrated pathway network [148].

Once researchers know which phenotypic responses are assigned to each compound, their synonyms and nomenclatural variations with the same meaning have to be resolved into a unique entity. Given that the wide medical vocabulary is subject to interpretation, it is complex to link related terms because some of them include others or are partial synonyms. Because of its completeness and accuracy, the Unified Medical Language System (UMLS)

developed by the National Library of Medicine, is the most used resource of biomedical and health related terminology. Thousands of terms including drug names, diseases and medical tools and techniques are extracted from medical literature and then collected into a huge metathesaurus where they are classified and their relations with other terms are pointed [149]. Moreover, the semantic and lexical relations among these terms are studied and described to assess researchers in this field.

From another point of view, medical terms can be hierarchically organized so that they can be studied at different levels of specificity. This is the objective of The Medical Subject Headings (MeSH) initiative, also developed within the NLM, that establishes a set of general categories like “Anatomy” or “Mental disorders” and then organizes related terms into sub levels up to those that are very specific [150]. Although much more informative, this approach lacks from the completeness and exhaustivity of the previously cited UMLS system, as not all medical terms are stored but only those more representative.

As we can see here again, an integrative approach to these subjects, such as mapping the vast amount of terms of the UMLS metathesaurus to the hierarchical MeSH system, may give place to a powerful characterization of the phenotypic space based on the relations between the composing entities. Furthermore, if this approach is taken at the moment of linking these entities with those in other spaces, like proteins or drugs, hidden information will emerge for the better understanding of the mechanistic undergoing processes in this field.

Part II – Objectives

The list of concrete objectives pursued in this thesis are:

1. The establishment of a protocol to extract, normalize and integrate multiple types of chemical, biological, and phenotypical data available from various public resources.
2. The exploitation of these data for the development of ligand-based target models for *in silico* pharmacology.
3. The construction of an annotated and structured repository of adverse drug reactions and its use to anticipate drug side effects.
4. The development of new integrative biochemoinformatic tools for the analysis and visualization of pharmacological data.

The first objective was accomplished by contributing to the creation of an annotated chemical library directed to nuclear receptors (see **Chapter IV.1**) and the implementation of an automated protocol for the integration of this internal resource with other publicly available annotated chemical libraries that was ultimately used to discuss the adequacy of the criteria for qualifying small molecules as chemical probes for biological systems (see **Chapter IV.6**). The construction of a general integrated repository of ligand-target interaction data tackled the sequent objective of developing novel ligand-based approaches to *in silico* target profiling (see **Chapter IV.4**). The third objective was addressed with the specification of a series of standardized strategies to solve the different issues involved in the generation of a drug side effects database connected with the previous one and its posterior analysis (see **Chapter IV.5**). Finally, the fourth objective was accomplished with the development of two web based tools, one to assess the functional coverage of the proteome by structures and the other one to navigate on ligand-target interaction data from highly curated public resources (see **Chapters IV.2** and **IV.3**).

Part III – Results and discussion

This thesis is centered on the study of some of the more relevant elements, data sources and computational methodologies currently used in biomedical sciences, focusing on those more closely related with the computational approaches to chemical biology and drug discovery. The word integration is, for sure, the most repeated term throughout this work, as it encloses the essence of the current trends in these fields and is the basis of future knowledge-based technologies to be developed in the future. Then, although different entities are involved in this study, all of them are interrelated and newly developed integrative tools have given rise to an important cross-fertilization across different fields.

Small molecules are the basic element in the drug discovery process, being designed and studied by the pharmaceutical industry with the intention to evolve into compounds that produce large scale beneficial effects on treated patients. The design and analysis procedures are evolving from their experimental basis with the incorporation of computational assessment tools at all levels of the process. Moreover, the revolution of high-throughput "omics" gave the community a considerable amount of data of all kinds, including larger molecular libraries with richer and more diverse information. The response data on the activity of compounds over druggable protein targets, for instance, holds the key to open the path for more accurate ligand-target interaction modeling technologies with a wide range of applications. However, it is difficult to study because most of this information is spread across articles published in peer reviewed journals in the last decades. As an example of how to overcome this issue, compounds having known binding affinities over nuclear receptors were collected and integrated into an annotated chemical library. The targets were then sorted into a functional classification and information on the privileged chemical structures binding specific protein groups was revealed. This library was later added to a knowledge-based architecture that collects information from public repositories of drug-target response data and organizes present proteins into hierarchical classifications. The resulting integrated database was then used as basic data framework for virtual profiling methodologies, allowing us to validate the combination of three two-dimensional descriptors defined

inside the laboratory as a useful virtual profiling methodology. Among these, SHED and FPD belong to the family of atom-pair molecular descriptors and are derived from distributions of atom-centered feature pairs extracted from the topology of molecules, whereas PHRAG is a pharmacophoric-based descriptor derived from a direct fragmentation of the molecular graph into overlapping segments of a fixed length.

Beyond this, the study of the target profiles of compounds in the database, including more than 3000 drugs, enabled the detection of putative chemical probes along with the proposal of a polypharmacology-based redefinition of this term. While the classic idea of a chemical probe comprehends compounds with high affinity and selectivity for a specific target protein, the proposal was to consider compounds accomplishing those rules for two or more targets as multiple probes. These new probes may be useful in the study of protein targets in the context of biological systems, robust biological networks were several nodes need to be attacked in order to produce an overall alteration of the system. Finally, a visualization architecture called iPhace was designed and implemented as a web site incorporating state of the art analysis tools to enable a high level overview of data present in annotated chemical libraries. The system organizes both ligands and proteins in hierarchical classifications and is able to answer complex questions with the analysis of the properties at the different levels of these hierarchies and the study of the activity profiles of compounds.

On the other hand, these aforementioned druggable proteins account for some of the most studied entities in biomedical sciences over the last fifty years because of their crucial involvement in all cellular processes. Their study is also of utmost importance from the point of view of drug discovery because drugs need to be active on proteins with specific functions and to achieve that they need to have a chemical structure capable of interacting with the target protein pocket. The detection of putative proteins from genetic sequences and the proteome wide specification of structure and function hold the focus of combined efforts of laboratories all around the world. In most cases, genetic sequences with concurring ancestors will codify for structurally related

proteins. Furthermore, proteins with similar three-dimensional foldings and active sites will tend to be involved in related activities. Hence, the study of protein phylogeny, the efforts for solving new uncharacterized structures and the implementation of functional annotation schemes, are expected to synergistically improve the results in all three fields with the use of novel technologies developed either for sequence analysis, structure solving and modeling or functional classification. In relation with this, the web tool FCP was designed and implemented for the analysis of the distribution of solved protein structures across the more relevant protein families, enabling a comprehensive overview of the current trends and historical evolution of this field. In the last version, information can be browsed through several functional classifications including schemes for kinases, nuclear receptors or G protein-coupled receptors among others. Beyond this, the user can also browse the database from the point of view of the cocrystallized ligands present in solved protein structures.

Despite the impact of all these integrative approaches, the alteration of such a complex system as the human body cannot be expected to be carried with complete success relying in current knowledge. As a result of this, when a drug is administered to a patient, several undesired side effects, not related with chemical toxicity or drug related allergies, may appear due to a combination of factors including age, gender, ethnic group and other individual specificities like genetic variations. These adverse drug events are actually the main reason for market drug withdrawal with a high economic impact on the pharmaceutical industry. Hence, the development of new methodologies to better assess drug safety and possible adverse reactions in the early stages of the drug discovery process is a relevant issue in biomedical sciences. As a basis of the development of such methodologies, a database relating marketed drugs with their reported side effects was collected and analyzed. Through the linkage of that database with the annotated chemical library previously generated, we could study the relations between the target profiles of those drugs and the side effects they produce. A network based approach showed how compounds annotated to similar adverse reactions were found to have affinity over related sets of targets

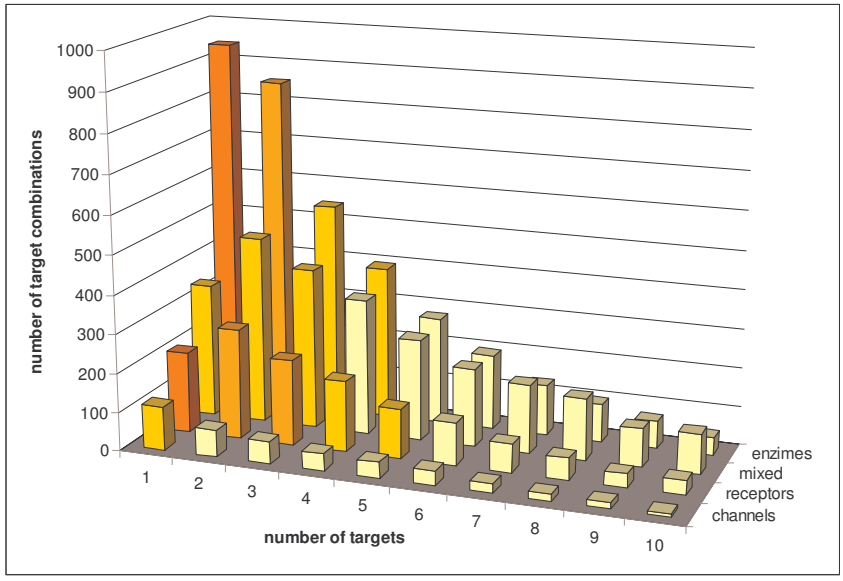
suggesting that comparative polypharmacology in preclinical stages on the drug discovery process could be a strong tool to overcome this issue.

So, how can all this knowledge improve the success rates in drug discovery? In silico target profiling, with the aim of predicting potential binding affinities of given molecules across the entire target space, is the key point were all the aforementioned fields may collide to produce synergetic advances. Applied at the initial steps of drug development, these computational methodologies enable researchers to browse the chemical space in search of compounds able to bind the desired protein profile. Most virtual profiling strategies are usually based on models created from experimental data so, in order to improve efficiency, we have to study available information to analyze the limitations it could present. As soon as we do this, we will see that any analysis we perform will be biased by the overall lack of data completeness.

Given that any compound may be highly active over one or several specific proteins and may also present residual affinities for other secondary targets, it is crucial to have highly populated ligand binding profiles to understand the mechanisms behind successful interactions. However, in our integrated database, the situation is that 52% of the compounds only have information related with 1 target, 41% with 2 to 5 targets, 6% with 6 to 15 targets and only 1.5% of them have known interaction values for more than 15 targets. Moreover, this information is highly biased towards active interactions, as inactivity data is not usually reported in literature. These problems generate a loss of accuracy in the study on how the features on those compounds enable them to bind some targets while not others, even within closely related groups. Finally, incomplete data also narrows the applicability domain of ligand-based virtual profiling methodologies that can only be used on compounds falling inside pre-established similarity ranges to the reference compounds. This suggests that there is a need to cover the portion of the chemical space of interest for each target protein with as many representative ligands as possible.

From the point of view of the proteins the situation is very similar. Even though the average number of ligands annotated to each target in our database is around 143, proteins which have been classically considered of interest in drug

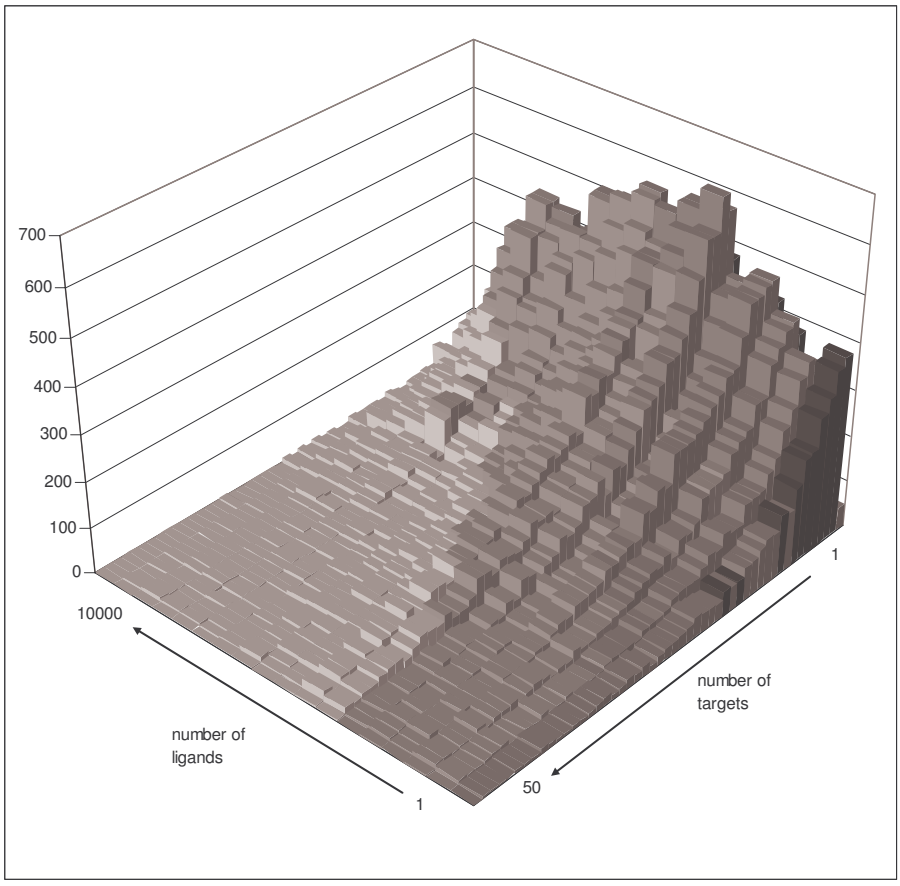
discovery have available data for thousands of compounds while 59% of the other targets have known interaction values for 5 ligands or less. Again, this means that for most of the targets it will be very difficult to have a good representation of the chemical space populated by compounds with high binding affinities, so the acquisition of an accurate understanding of what characteristics these ligands should present will be beyond our reaches. Additionally, an analysis of the different sets of targets with known binding affinities over coincident ligands showed that very few combinations of proteins have been explored and that this number decreases exponentially when the number of involved targets grows. This absence of cross-pharmacological data, relating proteins inside groups or across families by their binding affinities with the same sets of ligands, is also a major obstacle for the study of secondary interactions. In this situation, it is difficult to successfully predict the possible off-targets a drug may bind beyond its main target, causing a loss of efficiency in the drug design process. To overcome this situation we need to compare ligands independently annotated to the targets of interest relaying again in ligand-based similarity methods with the aforementioned limitations they carry.



Number of different target combinations with known affinities for a concurrent set of ligands depending on the number of targets. Darker bar colour means higher number of ligands with information for combinations in that category.

As we can see, some protein families account for more studied target combinations than others, although that also depends on the number of elements inside each group. Anyway, the number of targets sets from different families with data for the same ligands is small if we consider that this category accounts for the greatest number of possible combinations, meaning that ligands tested for proteins in a given family are not usually tested for representatives of the others.

These analyses suggest that current efforts should be focused on the determination of binding affinities of compounds with medium sized known activity profiles over mixed sets of targets containing representatives of different families. With this, we could provide complete binding profiles covering the same targets for a good set of compounds, achieving one of the possible solutions to the completeness issue: the elaboration of complete data matrices where all interaction values between the chosen sets of ligands and targets is known. These complete matrices could then be used to study the binding profiles of the specific compounds, in cross-pharmacology analyses and as benchmarks for virtual profiling methodologies. Within this thesis project, we analyzed the composition of all matrices we could build in basis of cross-pharmacology information extracted from public data, confirming that biggest matrix sizes account for a lower number of representatives and, in average, have lower completeness levels. A precise analysis of the detailed composition of these matrices suggests that a modest experimental effort in this direction would suppose a giant step towards the correction of this problem.



Distribution of matrices depending on the number of ligands and targets they contain. Darker color means higher completeness for matrices in that range.

On the other hand, in order to extract maximum knowledge from the available data, there is a need to go beyond ligand-similarity based approaches. As we said before, side effect based methodologies, not considering the chemical structure of the compounds but only the adverse reactions they might produce after being administered, are currently emerging as one of the possible complements. For a successful integration of this approach in a virtual profiling workflow we need to understand which are the relevant adverse reactions related with each drug and how they should be combined in the profiles of two drugs to support the transferring of one or more targets present in the target profile of one of these drugs to the other. Although these methods have only been successfully tested for compounds binding protein families with high

known cross-pharmacology, like g-protein coupled receptors, the robustness of this approach suggests that it may have a much wider application scope. However, the fact that this information is only available for several hundreds of well known drugs implies that we need to gather all available knowledge in order to expand as much as possible the applicability domain of this approach.

Beyond the drug discovery field, we have seen that the improvement of the current knowledge in all these disciplines is expected to come through the gathering of information coming from different fields into cohesive knowledge based schemes. New mechanistic explications for a diversity of unanswered questions are expected to emerge from evolving integrative approaches but beyond the cross relation and organization of all these data, a new landscape of methodologies and visualization tools has to be developed. This last step, going beyond the creation of high level databases, will enable researchers to extract useful knowledge with the detection of the relevant characteristics of the individual elements that become apparent only at the systems level.

Part IV – Publications

Chapter IV.1

Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family

Cases M, García-Serna R, Hettne K, Weeber M, van der Lei J, Boyer S, Mestres J

Curr. Top. Med. Chem. 2005, **5**: 763-772

Cases M, García-Serna R, Hettne K, Weeber M, van der Lei J, Boyer S, et al. Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family. Curr Top Med Chem. 2005; 5(8): 763-72.

Chapter IV.2

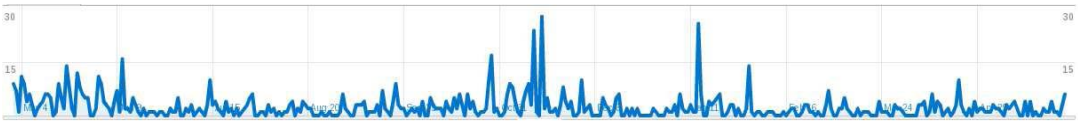
FCP: functional coverage of the proteome by structures.

García-Serna R, Opatowski L, Mestres J

Bioinformatics 2006, **22**: 1792-1793

García-Serna R, Opatowski L, Mestres J. FCP: functional coverage of the proteome by structures. Bioinformatics. 2006; 22(14): 1792-3.

Summary of activity in FCP website between 01-May-09 and 01-June-10



Number of visits per day in this period

More than 350 unique users have visited the site, which has been accessed almost one thousand times with over six thousand pages viewed in total. The day with more visits was November the 16th with 28.

Number of visits per country in this period

Visitors came from 146 cities in 38 different countries. Among the 10 most repeated countries we find Spain, EEUU, Ireland, France, India, Germany, Italy, UK, Portugal and Denmark. If we look at statistics by cities we will find Barcelona, Dublin, San Francisco, Lyon, Madrid and Bangalore with more than 10 visits each. Finally, these are the companies or research groups responsible for most of the visits:

Entity	Visits
Institut Municipal d'Investigació Mèdica	323
Universitat Pompeu Fabra	121
Trinity College Dublin	43
Ecole Normale Superieure de Lyon	31
Universit degli Studi g. d'Annunzio	16
Parc de Recerca Biomedica de Barcelona	14
University of California san Francisco	14
Danish network for research and education	11
Freie Universitaet Berlin	8
Organon laboratories ltd	5
Universite d'Angers	5

Chapter IV.3

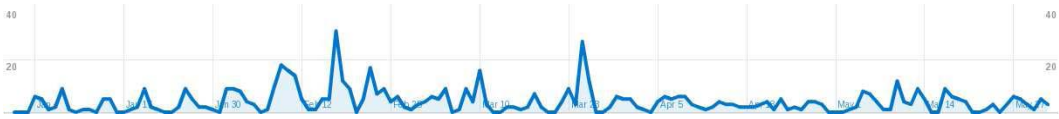
iPHACE: integrative navigation in pharmacological space.

Garcia-Serna R, Ursu O, Oprea TI, Mestres J

Bioinformatics 2010; **26**: 985-986

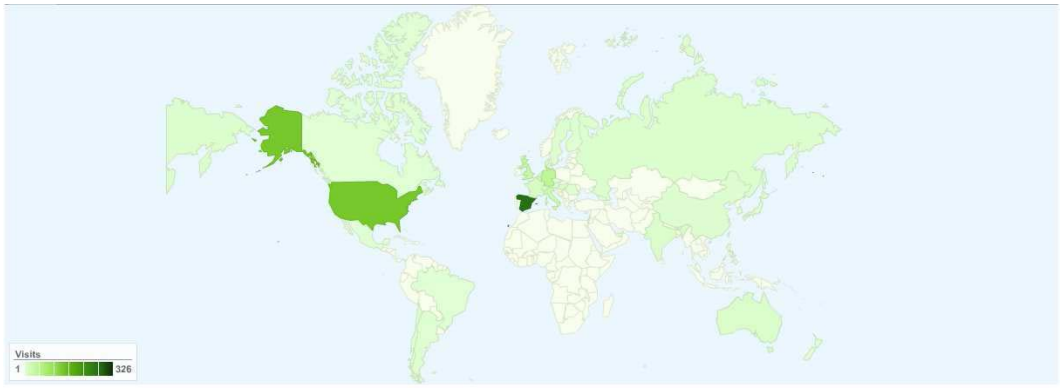
García-Serna R, Ursu O, Oprea TI, Mestres J. iPHACE: integrative navigation in pharmacological space. Bioinformatics. 2010; 26(7): 985-6.

Summary of activity in iPhace website between 01-January-10 and 01-June-10



Number of visits per day in this period

Almost four hundred unique users have visited the site, which has been accessed more than 750 times with over two thousand pages viewed in total. The day with more visits was February the 17th with 43.



Number of visits per country in this period

Visitors came from 146 cities in 33 different countries. Among the 10 most repeated countries we find Spain, EEUU, Germany, UK, Italy, Netherlands, Belgium, Denmark, Israel and France. If we look at statistics by cities we will find Barcelona, Albuquerque, Madrid, Edimburgh, Oss, Dortmund, Perugia and Copenhagen with more than 10 visits each. Finally, these are the companies or research groups responsible for most of the visits:

Entity	Visits
Universitat Pompeu Fabra	120
Institut Municipal d'Investigació Mèdica	118
University of New Mexico Healthsciences Center	44
Edinburgh University	18
Danish Network for Research and Education	16
Organon Biosciences	15
Abbott Laboratories	12
The Hebrew University of Jerusalem	11
Universit degli Studi di Perugia	10
Universitat de Barcelona	10
Hoffmann Laroche inc.	9
Pfizer inc.	7

Chapter IV.4

Ligand-based Approaches to In Silico Pharmacology.

Vidal D, Garcia-Serna R, Mestres J

Methods Mol. Biol. 2011, **672** (in press)

Vidal D, García-Serna R, Mestres J. Ligand-based approaches to in silico pharmacology. Dins de: Bajorath J (ed.). Chemoinformatics and Computational Chemical Biology. New York: Springer: Humana Press, 2011. (Methods in molecular biology; 672). p. 489-502.

Chapter IV.5

Anticipating drug side effects by comparative pharmacology

Garcia-Serna R, Mestres J

Expert Opin. Drug Metab. Toxicol. 2010 (submitted)

Anticipating drug side effects by comparative pharmacology

Ricard Garcia-Serna & Jordi Mestres*

Chemogenomics Laboratory, Research Program on Biomedical Informatics (GRIB), Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain

Drugs having similar side-effect profiles were successfully shown recently to have also affinity for a common target. Going one step further, this review highlights the fact that drugs connected by side-effect similarity have also similar affinity profiles over multiple targets. Given the current low levels of completeness of drug-target interaction data, one may expect that the number of common targets for which drug neighbours in a side-effect network have some level of affinity is likely to increase in the near future. This observation suggests that preclinical comparative pharmacology, based on *in vitro* data but supported strongly by modern *in silico* methods for drug metabolite prediction and affinity profiling, represents an attractive strategy to anticipate clinical drug side effects.

Keywords: adverse drug reactions, affinity profiles, chemogenomics, systems chemical biology, drug metabolism, robustness, perturbations

1. Introduction

It is widely recognised that any molecule capable of producing beneficial therapeutic effects has also a potential risk to produce side effects due to unforeseen pharmacological properties [1]. Unwanted side effects lead often to harmful or unpleasant reactions to the patient and, depending on its severity, they may ultimately result in the withdrawal of the drug from the market [2]. This notwithstanding, not all side effects have negative implications and some might actually reveal interesting alternative opportunities for the therapeutic repositioning of drugs [3]. Therefore, anticipating the likely side-effect profile of drugs is an aspect of key importance in current drug discovery, development, and marketing.

Unfortunately, indicative signals of many side effects are only detected once the drug is in the market [4]. The relatively limited number of participating patients and short duration of clinical trials make difficult to detect adverse drug reactions (ADRs) in patients that occur only rarely, have a long latency, or manifest exclusively in specific populations [5]. Any ADR inadvertedly missed prior to approval of a drug may potentially pose serious health threats once released into the general population [6]. In this respect, the increasing availability of electronic health records, and their integration with multiple biomedical databases, offers new opportunities for the early detection of ADRs in the postmarketing phase [7].

In recent years, the discovery of direct links between the interaction with a particular target and the development of a certain ADR has motivated the systematic testing of compounds at the early phases of drug discovery using *in vitro* biochemical and cellular assays [8-10]. For example, blockade of the hERG potassium channel may result into QT interval prolongation [11] or agonism for the serotonin 5-HT_{2B} receptor may translate into the possible occurrence of cardiac valvulopathy [12]. Many associations between single targets and possible ADRs have been established, meaning that assay panels for preclinical safety pharmacology covering between 50 and 100 targets are quite common nowadays. However, the current limited knowledge on both disease-relevant targets [13] and drug-target interactions [14-16] makes it likely that many more are still to be determined, implying that even more extensive *in vitro* safety assays would perhaps need to be implemented.

Consistently performing *in vitro* safety profiling of hundreds of compounds over multiple targets can be logistically complicated and extremely expensive and thus there is increasing interest in developing *in silico* methods for the identification of new targets

that could explain some of the side effects known for old drugs [17-20]. Indeed, it was recently shown that drugs producing similar profiles of side effects tend to share some targets [21] and viceversa, that drugs with a similar affinity profile over multiple targets tend to share some side effects [22,23]. Therefore, having the ability to predict *in silico* the complete affinity profile of small molecules and comparing it against the *in vitro* affinity profiles of known drugs may prove useful to foresee the likely side-effect profile of drug candidates [24]. Accordingly, the aim of this contribution is to perform a step-by-step analysis of the links between the affinity profile of a drug and its associated side-effect profile, form an opinion on the actual applicability of this type of approaches to anticipate drug side effects, and highlight some of the current limitations and future trends in this direction.

2. Relating drugs by side-effect similarity

The process of relating drugs by side-effect similarity [21,22] requires some careful, and in some aspects extremely tedious, prework in three essential areas. First and foremost, it is key to adopt a side-effect terminology that ideally can be further mapped to alternative existing terminologies. Subsequently, this side-effect terminology will be used to identify and parse drug side-effect terms from multiple drug resources. Finally, comparison of all drug side-effect profiles extracted will be done by means of a purposely defined similarity metric. The particular implementations of all these aspects will influence the relationship of drugs from the perspective of side effects.

The ability to recognise and process all side effects caused by drugs reported in publicly available web resources is very much dependent on the use of a standardised side-effect terminology and the completeness of its associated thesaurus [25]. In this respect, Bender et al. [23] used for example the ADR terminology present in the World Drug Index [26]. In contrast, both Campillos et al. [21] and Fliri et al. [22] used the concepts of the Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART), which are part of the Unified Medical Language System (UMLS) metathesaurus [27]. Alternatively, we propose to use the Medical Subject Heading (MeSH) terminology, where the number of terms is much lower than in UMLS but the categorisation is more specific [28]. However, mapping the two terminologies is not straightforward as all MeSH terms are included in the UMLS ontology but the reverse is not true. Accordingly, the strategy adopted was to map each specific MeSH term to its exact term in UMLS extended to all its related UMLS synonyms. For example, the

MeSH term “chronic periodontitis” will be mapped to the corresponding exact UMLS term but also to all its synonyms, such as “chronic pericementitis”, “Fauchard’s disease”, and “Riggs’ disease”, among others. In the end, one should have a side-effect dictionary composed of a list of unique MeSH side-effect terms linked to a comprehensive thesaurus covering all possible UMLS synonyms for each term.

Data for drug side effects are available from various public sources, namely, Dailymed [29], Drug Information Online [30], MedlinePlus [31], Medscape [32], and ToxNet [33], to name just a few. Unfortunately, drug terminology is not standardised among them and different resources may refer to the same drug entity using a variety of country-specific commercial names and identifiers that in some cases may also include forms of administration. For example, the drug “aripiprazole” can be found in different sources as “aripiprazole orally disintegrating tablets”, “aripiprazole solution”, “abilify”, or “abilify oral”, among others. Accordingly, a prerequisite for mining properly all this information is the adoption of a drug dictionary and associated drug name thesaurus that allows for collapsing all drug terms found in the various sources to a unique drug entity. Then, for each drug entry covered by each source, the text related to ADRs is searched for matches to the side-effect terminology described above.

Finally, different approaches can ultimately be used to translate the drug side-effect data extracted from all sources into a mathematical representation of the phenotypic profile of a drug. Usually, side-effect data are first converted into a binary fingerprint that assigns a value of 1 for any term in the side-effect dictionary that is associated with the drug and a value of 0 otherwise. Then, Fliri et al. [22] applied for example a hierarchical clustering method to group drugs having similar side-effect profiles and measure the similarity between groups using half-square Euclidean distances. Further refinements can be introduced by taking into consideration that side effects vary greatly in abundance and that not all side effects are independent of each other. For example, nausea, emesis, asthenia, and diarrhoea are among the most common side effects and many drugs causing nausea are also linked to vomiting. In order to correct for these observations, Campillos et al. [21] weighted side effects accordingly and then calculated the similarity between the side-effect profiles of two drugs by summing the product of the weights over all shared side-effect terms. In our implementation, we assign to each drug side-effect a confidence score depending on the reporting frequency among the different sources used. For example, a drug side effect reported in all five sources listed above [29-33] will be assigned a confidence score of 1.0, whereas another drug side

effect being found only in one of those sources will be given at this stage a confidence score of 0.2. Then, side-effect similarity between a drug pair can be evaluated as the probability that consulting any source a certain MeSH label is found for both drugs.

The result of all this process is a list of scores that reflect the side-effect profile similarity between all possible drug pairs. At this stage, one can apply network theory to visualise and analyse the complexity of drug connections emerging from side-effect similarities in a simple and compact manner [21]. As an illustrative example, Figure 1 shows the drug network obtained with our own implementation, in which the number of links to each node in the network was limited to the three most similar nodes. In total, the network contains 2,733 unique drug entities, covering 1,924 MeSH labels representing 4,589 UMLS terms. The inset in Figure 1 highlights a particular network path composed of 14 drugs, 8 of which were also the focus of attention in the previous work by Campillos et al. [21]. It is worth stressing the therapeutic diversity found along this path of drugs connected by common side effects. Under the Anatomic Therapeutic Chemical (ATC) classification [34], one can find five antidepressants (N06A: venlafaxine, mirtazapine, nefazodone, fluvoxamine, and fluoxetine), two antipsychotics (N05A: ziprasidone and risperidone), two antidementia (N06D: donepezil and tacrine), two antihistamines for systemic use (R06A: cetirizine and levocetirizine), one antiepilectic (N03A: gabapentin), one hypnotic and sedative (N05C: zaleplon), and one dopaminergic agent (N04B: pergolide). Also, among the 13 connections, only four drug pairs share a common therapeutic indication up to the fourth level of the ATC code. This set of drugs will be taken as the framework of reference on which the following discussion about the links between side-effect and target profiles will be centred.

3. Similar side-effect profiles reveal common targets

Connecting drugs by side-effect similarity was recently proven to be an attractive strategy to predict novel targets for drugs beyond mere chemical and sequence similarities [21]. However, in order to be able to analyse the extent by which similar side-effect profiles may indicate common targets for drugs, one needs to extract and parse previously all affinity data available at present for drugs from multiple sources, such as ChEMBLdb [35], PDSP [36], BindingDB [37], and IUPHARdb [38]. Again, a prerequisite for storing properly all these affinity data for targets is the construction of a target name thesaurus that allows for collapsing all target terms found in the various sources to a target dictionary composed of a unique term per target [39]. In our own

implementation of this task [16], of those drugs for which side-effect data are available, a total of 7,979 interactions were found between 1,153 drugs and 1,867 targets, of which 5,316 can be considered as "active" (pActivity > 6). Note that the total number of drug-target interactions for which experimental affinity is publicly known constitutes only the 0.37% of all possible affinities in the complete drug-target interaction matrix (1,153 x 1,867) and thus emphasises the fact that currently available drug-target interaction data are far from being complete [14,15].

Focusing on the selected network path described previously containing 14 drugs connected by side-effect similarity, one recognises five drug pair links that were identified earlier to have a probability of sharing a target higher that 25% and side-effect similarity P value below 0.1 (values in italics in Figure 1) [21]. Four of them were considered to involve drugs known to share some targets already but the link between zaleplon and mirtazapine was identified as one of the potentially interesting drug pairs not known to share any target at the time [21]. In this respect, the strategy followed in the work by Campillos et al. [21] consisted basically on transferring the primary target of one of the drugs in the pair to the other drug. This way, since the primary target of mirtazapine is the histamine H_1 receptor (pK_i = 9.1), zaleplon was tested on that target and low micromolar affinity was indeed found (pK_i = 4.6). Another of the links identified in that work was between donepezil and venlafaxine, two drugs involved also in our network path. In this case, the primary target of venlafaxine is the sodium-dependent serotonin transporter (pK_i = 7.4) and testing of donepezil on that target confirmed low affinity for it (pK_i = 5.1).

Figure 2 exemplifies how the strategy of transferring primary targets between drug pairs connected by side-effect similarity is applied to the last four drugs of our network path. Nefazodone is an antidepressant acting primarily as a potent antagonist of the serotonin receptor 5-HT_{2A} (pK_i = 8.1) and thus it is expected to transfer affinity for this target to its path neighbours. In the network path (Figure 1), nefazodone is linked to fluvoxamine which indeed shows low micromolar affinity for 5-HT_{2A} (pK_i = 4.9). In turn, fluvoxamine is also an antidepressant which primarily acts (pK_i = 8.4) on the sodium-dependent serotonin transporter (SERT). Under this transferring approach, its path neighbours are expected to have some affinity for SERT. Unfortunately, based on publicly available interaction data, we could only confirm at this stage that nefazodone shows submicromolar affinity for SERT (pK_i = 6.7). Following on the path, the primary target for the anti-Parkinson drug pergolide is the dopamine D_3 receptor (pK_i = 8.4) and

thus, its path neighbours should retain some affinity for that target. Most interestingly, we recover in this case the prediction that fluoxetine should have affinity for D_3, another one of the experimentally confirmed interactions from the work of Campillos et al. (pK_i = 5.7) [21]. Overall, this path analysis illustrates nicely the use of a primary target transferring strategy to identify common targets between drugs connected by side-effect similarity with no obvious structural similarity.

Given the currently known polypharmacology of drugs [16], one may argue at this stage that perhaps the transferring of a single target between neighbours in the side-effect network is an over simplification of the approach. As pointed out earlier [21], side effects rarely occur independently and different side effects may be caused through essentially distinct mechanisms of action. In this respect, an analysis of the publicly available affinity profiles of the 14 drugs in our network path under study reveals that those drug pairs share far more than one target. In particular, five targets are identified to have some level of affinity (<50 μM) for over 50% of the drug set, namely, histamine H_1 receptor (10), serotonin receptor 5-HT_{2A} (9), serotonin transporter (9), serotonin receptor 5-HT_{1A} (8), and adrenergic receptor α_{1A} (8). Acknowledging that publicly available data on drug-target interactions are far from complete, one may expect that many more common interactions between those drug pairs are yet to be identified [14,15].

To have a glimpse at this assumption, Figure 3 allows for comparing the affinity profiles on a selected panel of targets for the four drugs presented in Figure 2, including the five most frequent targets mentioned above complemented with an additional set of five targets, namely, serotonin receptor 5-HT_{2C}, adrenergic receptor α_{2A}, dopamine D_3 receptor, and the dopamine (DAT) and noepinephrine (NET) transporters, for which public interaction data on those four drugs was available. Interestingly, five of the targets in this list (namely, 5-HT_{1A}, 5-HT_{2A}, DAT, NET, and SERT) appeared grouped together in a recent study examining the coinvestigation frequency of medicines [40]. As can be observed, all drugs have biologically active interactions (pActivity > 5.0) with many targets other than the respective primary targets (grey bars). Note that, among the different affinity profiles, a few drug-target interactions remain unknown (marked with an arrow in Figure 3), namely, nefazodone and fluvoxamine with the dopamine D_3 receptor, and pergolide with the three neurotransmitter transporters DAT, NET, and SERT. Since the majority of the targets involved in these profiles seem to get

transferred between drugs linked by side-effect similarity, one may anticipate that these five missing interactions are likely to be confirmed experimentally at some point in the near future. Therefore, a multiple target transferring strategy between drug neighbours in a side-effect network could represent an attractive complement to molecular similarity approaches to complete the affinity profile of drugs.

4. Expert opinion and conclusions

The observation that drugs connected by side-effect similarity share affinities for multiple targets provides an indication that comparative pharmacology may be used to identify drugs having similar side-effect profiles and thus help establishing a link between preclinical and clinical drug-induced effects [21-23,40]. There are however several limitations to the current approaches to assessing the similarity between drug side-effect profiles, on one hand, and affinity profiles, on the other hand, that may limit their scope for anticipating drug side effects.

With respect to assessing the similarity between drug side-effect profiles, refined methods would need to take into consideration data on the relative risk (a binary approach is currently being used), drug concentration (all side effects are assumed to occur at the recommended dose), and side-effect co-incidence (a unique drug side-effect profile is considered, whereas they represent in fact an ensemble of independent effects associated to the same drug). Compilation and availability of this type of data are the main issues at the moment. In this respect, the recent publication of a dedicated resource that capture drug side effects and their frequency in patients relative to placebo represents a step forward in this direction [41]. In addition, large coordinated initiatives aiming at developing techniques for data mining of electronic health records across different countries are also expected to make available in the public domain some of these data for a selected list of ADRs [7]. In spite of the assumptions being currently made, side-effect similarity has emerged as an interesting approach to predict novel targets for old drugs beyond the applicability domain of methods based on molecular similarity [21].

With respect to assessing the similarity between affinity profiles, given the current low levels of completeness of experimental data [14,15], advances in the development of methods that can reliably perform *in silico* target profiling of small molecules will be key to this aspect [42,43]. Providing an estimation of the complete affinity profile of a drug may become of utmost importance to understand the relative frequency of ADRs

in specific populations [24,44]. For example, an elder population metabolises substances more slowly than a young population, which might have a direct effect on the concentration of drug in plasma and thus, become a potential liability for drug side effects due to affinities to off-targets that would otherwise be considered negligible assuming normal plasma levels. In this respect, one should also take into consideration that in some cases drug metabolites can be the actual species responsible for the side effects produced and thus target profiling of these metabolites should also be performed. The association of cardiopulmonary side effects to fenfluramine through activation of the 5-HT_{2B} receptor by its N-deethylated metabolite norfenfluramine represents a good example of this type of potential scenario [45]. Accordingly, the interest on metabolite prediction and identification has increased significantly in recent years [46] in connection with the role of P450 enzyme polymorphisms in the pathogenesis of ADRs [47].

Biological systems are intrinsically robust [48]. Accordingly, selectively interacting with one single target might not be the most efficient strategy to have therapeutic efficacy, as the system may find other ways to compensate for the perturbation introduced. The alternative is to interact with multiple targets, so making more difficult for the system to compensate for all. The result is therapeutic efficacy but it does not come for free, as decompensation in some particular systems may translate into adverse drug reactions. At present, we can link a certain pattern of interactions to the likely incidence of some side effects in a relatively small portion of the patient population. However, this still does not provide an answer to the key question as to why a small percentage of human systems are less robust than the rest to certain patterns of interactions. To address this question, one will need to integrate chemical structures, target affinities, biological pathways, and individual genomic data into a systems approach to drug action [49,50]. Efforts in this direction might pave the way for anticipating drug side effects at a personalised level [51].

Acknowledgments. This research was supported by the Spanish Instituto de Salud Carlos III and the Ministerio de Ciencia e Innovación (project BIO2008-02329). Funding was received from the European Community's 7th Framework Programme (FP7/2007-2013), under grant agreement no. 215847 (EU-ADR project), and the Innovative Medicines Initiative, under grant agreement no. 115002 (eTOX project).

Bibliography

1. EDWARDS IR, ARONSON JK: Adverse drug reactions: definitions, diagnosis, and management. *Lancet* (2000) **356**:1255-1259.
2. GIACOMINI KM, KRAUSS RM, RODEN DM *et al.*: When good drugs go bad. *Nature* (2007) **446**:975-977.
3. ASHBURN TT, THOR KB: Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* (2004) **3**:673-683.
4. HAUBEN M, BATE A: Decision support methods for the detection of adverse events in post-marketing data. *Drug Discov. Today* (2009) **14**:343-357.
5. HÄRMARK L, VAN GROOTHEEST AC: Pharmacovigilance: methods, recent developments and future perspectives. *Eur. J. Clin. Pharmacol.* (2008) **64**:743-752.
6. AJAYI FO, SUN H, PERRY J: Adverse drug reactions: a review of relevant factors. *J. Clin. Pharmacol.* (2000) **40**:1093-1101.
7. TRIFIRÒ G, PARIENTE A, COLOMA PM *et al.*: Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol. Drug Saf.* (2009) **18**:1176-1184.
8. KREJSA CM, HORVATH D, ROGALSKI SL *et al.*: Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discov. Devel.* (2003) **6**:470-480.
9. ROTH BL, SHEFFLER DJ, KROEZE WK: Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* (2004) **3**:353-359.
10. WHITEBREAD S, HAMON J, BOJANIC D, URBAN L: In vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov. Today* (2005) **10**:1421-1433.
11. SANGUINETTI MC, TRISTANI-FIROUZI M: hERG potassium channel and cardiac arrhythmia. *Nature* (2006) **440**:463-469.
12. ROTH BL: Drugs and valvular heart disease. *N. Engl. J. Med.* (2007) **356**:6-9.
13. CASES M, MESTRES J: A chemogenomic approach to drug discovery: focus on cardiovascular diseases. *Drug Discov. Today* (2009) **14**:479-485.
14. MESTRES J, GREGORI-PUIGJANÉ E, VALVERDE S, SOLÉ RV: Data completeness: the Achilles heel of drug-target networks. *Nat. Biotechnol.* (2008) **26**:983-984.
15. MESTRES J, GREGORI-PUIGJANÉ E, VALVERDE S, SOLÉ RV: The topology of drug-target interactions: implicit dependence on drug properties and target families. *Mol. BioSyst.* (2009) **5**:1051-1057.
16. VOGT I, MESTRES J: Drug-target networks. *Mol. Inf.* (2010) **29**:10-14.
17. KEISER MJ, ROTH BL, ARMBRUSTER BN *et al.*: Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* (2007) **25**:197-206.
18. GREGORI-PUIGJANÉ E, MESTRES J: Ligand-based approach to mining the chemogenomic space of drugs. *Comb. Chem. High Throughput Screen.* (2008) **11**:669-676.
19. CAMPILLOS M, KUHN M, GAVIN A-C *et al.*: Drug target identification using side-effect similarity. *Science* (2008) **321**:263-266.

20. KEISER MJ, SETOLA V, IRWIN JJ *et al.*: Predicting new molecular targets for known drugs. *Nature* (2009) **462**:175-182.
21. CAMPILLOS M, KUHN M, GAVIN A-C, JENSEN LJ, BORK P: Drug target identification using side-effect similarity. *Science* (2008) **321**:263-266.
22. FLIRI AF, LOGING WT, THADEIO PF, VOLKMANN RA: Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat. Chem. Biol.* (2005) **1**:389-397.
23. BENDER A, SCHEIBER J, GLICK M et al.: Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* (2007) **2**:861-873.
24. VIDAL D, MESTRES J: In silico receptorome profiling of antipsychotic drugs. *Mol. Inf.* (2010) submitted.
25. HENEGAR C, BOUSQUET C, LILLO-LE LOUËT A, DEGOULET P, JAULENT M-C: Building an ontology of adverse drug reactions for automated signal generation in pharmacovigilance. *Comput. Biol. Med.* (2006) **36**:748-767.
26. World Drug Index: http://thomsonreuters.com/products_services/science/science_products/a-z/world_drug_index
27. BODENREIDER O: The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* (2004) **32**:D267-D270. http://www.nlm.nih.gov/research/umls/
28. MeSH: http://www.ncbi.nlm.nih.gov/mesh
29. Dailymed: http://dailymed.nlm.nih.gov
30. Drug Information Online: http://www.drugs.com
31. MedlinePlus: http://medlineplus.gov
32. Medscape: http://www.medscape.com
33. ToxNet: http://toxnet.nlm.nih.gov
34. WHO Collaborating Centre for Drug Statistics Methodology, Guidelines for ATC classification and DDD assignment, 2010. Oslo, 2009. http://www.whocc.no
35. ChEMBL database (http://www.ebi.ac.uk/chembldb)
36. Psychoactive Drug Screening Program: http://pdsp.med.unc.edu
37. LIU T, LIN Y, WEN X, JORRISEN RN, GILSON MK: BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* (2007) **35**:D198-D201.
38. HARMAR AJ, HILLS RA, ROSSER EM et al.: IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Res.* (2009) **37**:D680-D685.
39. GARCIA-SERNA R, URSU O, OPREA TI, MESTRES J: iPHACE: integrative navigation in pharmacological space. *Bioinformatics* (2010) **26**:985-986.
40. FLIRI AF, LOGING WT, VOLKMANN RA: Drug effects viewed from a signal transduction network perspective. *J. Med. Chem.* (2009) **52**:8038-8046.
41. KUHN M, CAMPILLOS M, LETUNIC I, JENSEN LJ, BORK P: A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* (2010) **6**:343.

42. EKINS S, MESTRES J, TESTA B: In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br. J. Pharmacol.* (2007) **152**:9-20.
43. ROGNAN D: Structure-based approaches to target fishing and ligand profiling. *Mol. Inf.* (2010) **29**:176-187.
44. YANG L, LUO H, CHEN J, XING Q, HE L: SePreSA: a server for the prediction of populations susceptible to serious adverse drug reactions implementing the methodology of a chemical-protein interactome. *Nucl. Acids Res.* (2009) **37**:W406-W412.
45. SETOLA V, ROTH BL: Screening the receptorome reveals molecular targets responsible for drug-induced side effects: focus on 'fen-phen'. *Expert Opin. Drug Metab. Toxicol.* (2005) **1**:377-387.
46. ANARI MR, BAILLIE TA: Bridging cheminformatic metabolite prediction and tandem mass spectrometry. *Drug Discov. Today* (2005) **10**:711-717.
47. PIRMOHAMED M, PARK BK: Cytochrome P450 enzyme polymorphisms and adverse drug reactions. *Toxicology* (2003) **192**:23-32.
48. KITANO H: A robustness-based approach to systems-oriented drug design. *Nat. Rev. Drug Discov.* (2007) **6**:202-210.
49. SCHEIBER J, CHEN B, MILIK M *et al.*: Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J. Chem. Inf. Model.* (2009) **49**:308-317.
50. TATONETTI NP, LIU T, ALTMAN RB: Predicting drug side-effects by chemical systems biology. *Genome Biol.* (2009) **10**:238.
51. GINSBURG GS, WILLARD HF: Genomic and personalized medicine: foundations and applications. *Transl. Res.* (2009) **154**:277-287.

Affiliation

Ricard Garcia-Serna[†] PhD & Jordi Mestres[‡] PhD

[†] Current address

Chemotargets SL, Passeig de Circumval.lació 8, 08003 Barcelona, Catalonia, Spain

[‡] Author for correspondence

Chemogenomics Laboratory, Research Program on Biomedical Informatics (GRIB), Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain

Tel: +34 93 3160540; Fax: +34 93 3160550

E-mail: jmestres@imim.es

FIGURE CAPTIONS

Figure 1. Network of drugs connected by side-effect similarity, with focus on a particular network path composed of 14 drugs (left). For five drug-pair connections, the probability of sharing a target higher than 25% and side effect similarity P value below 0.1, as reported in reference [21], is also shown (values in italics).

Figure 2. The process of transferring the primary target of a drug to its side-effect neighbours in the network path.

Figure 3. The bioactivity profiles over ten targets of four drugs connected by side-effect similarity (see Figure 2). The dashed line indicates the activity level of 10 μM. Arrows mark current missing data for which low micromolar affinity is predicted.

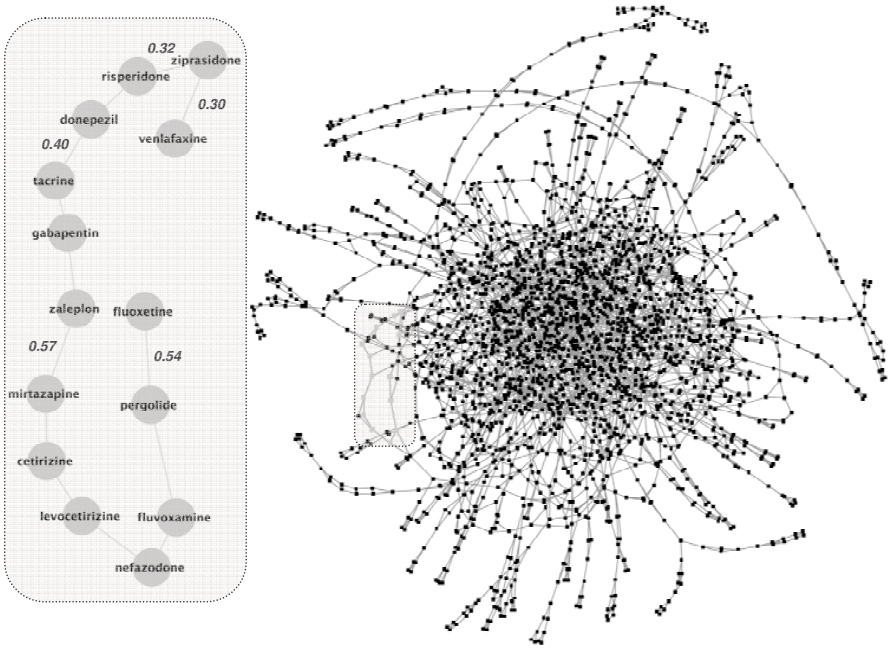

0.32
ziprasidone
risperidone
0.30
donepezil
venlafaxine
0.40
tacrine
gabapentin
zaleplon
fluoxetine
0.57
0.54
mirtazapine
pergolide
cetirizine
levocetirizine
fluvoxamine
nefazodone

FIGURE 1

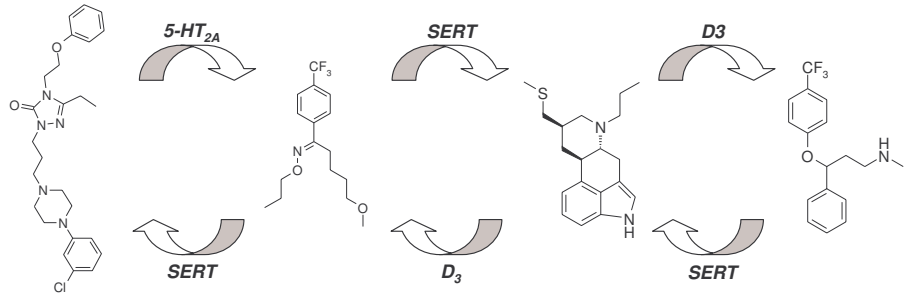

5-HT2A
SERT
SERT
D3
SERT
D3
nefazodone
5-HT2A=8.1
SERT=6.7
fluvoxamine
SERT=8.4
5-HT2A=4.9
D3=?
pergolide
D3=8.4
SERT=?
fluoxetine
SERT=8.2
D3=5.7 (from ref. [21])

FIGURE 2

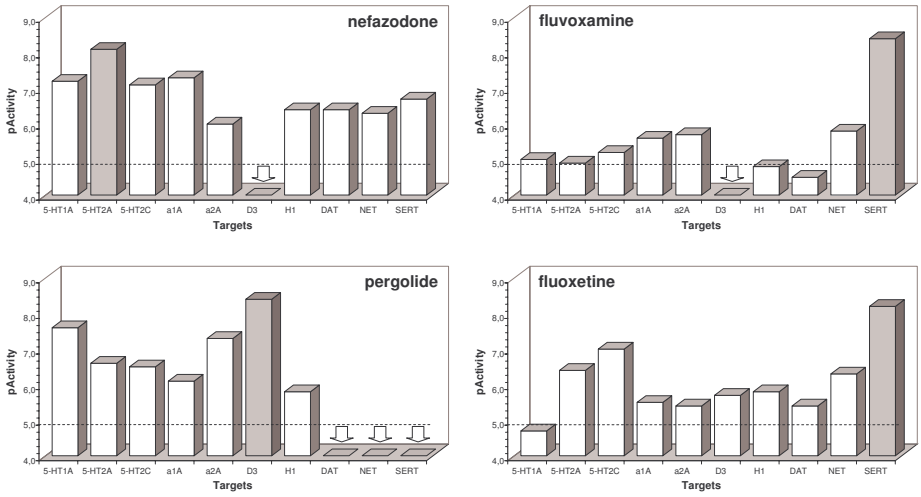

nefazodone
fluvoxamine
pergolide
fluoxetine
pActivity
9,0
8,0
7,0
6,0
5,0
4,0
5-HT1A
5-HT2A
5-HT2C
a1A
a2A
D3
H1
DAT
NET
SERT
Targets

FIGURE 3

Chapter IV.6

Chemical probes for biological systems

Garcia-Serna R, Mestres J

Drug Discov. Today 2010 (submitted)

Chemical probes for biological systems

Ricard Garcia-Serna and Jordi Mestres

Chemogenomics Laboratory, Research Program on Biomedical Informatics (GRIB), Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Parc de Recerca Biomèdica, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain

Abstract

According to the latest definition in use by the NIH Molecular Libraries Screening Centers Network, a compound to be nominated as a chemical probe should have, on one hand, an affinity below 100 nM for the primary target and, on the other hand, at least 10-fold selectivity against related targets. Even though the main objective of chemical biology initiatives are not to deliver clinically useful compounds, it is highly expected that the chemical probes being identified are then optimised by chemists to translate basic research discoveries into therapeutics. However, a detailed analysis of drugs reveals that only 14.4% of them would actually qualify as chemical probes. Most interestingly, it was found that the percentage of drugs that qualify as chemical probes is severely reduced as more information on the affinity profile over multiple targets is available. The main conclusion that can be drawn from this analysis is that using the current criteria we might be overlooking many compounds with potential therapeutic interest. The change from probing a single biological target to probing an intrinsically robust biological system

Corresponding author: Mestres, J. (jmestres@imim.es)

might require adjusting current criteria on affinity and selectivity for qualifying small molecules as chemical probes.

Defining a chemical probe

The identification of a small-molecule modulator for each individual function of all human proteins has been proposed as one of the grand challenges for chemical biology in the years to come [1]. Such an ambitious goal can only be achieved through large-scale coordinated initiatives involving chemical and screening centers at a multinational level. Recently, centres from 12 countries have agreed to join efforts in assembling an European Infrastructure of Open Screening Platforms for Chemical Biology (EU-Openscreen) that is currently halfway through its preparatory phase (2009-2011) [2]. In a much more advanced stage, the US National Institutes of Health (NIH) just completed the pilot phase (2004-2008) of its Molecular Libraries and Imaging (MLI) initiative comprising ten high-throughput screening (HTS) centres known as the Molecular Libraries Screening Centers Network (MLSCN) [3,4]. Within these five years, a Molecular Libraries Small Molecule Repository (MLSMR) was created and screened against an impressive total number of 691 assays, covering 171 targets and 29 phenotypic screens [5]. All the interaction data between small molecules and biochemical assays generated during this period has been deposited and made publicly accessible in PubChem [6].

The focus of these chemical biology initiatives is the identification of probe molecules to be used in basic research of biological systems. In this respect, the screening centres of the MLSCN have collectively nominated already 64 bioactive small molecules as chemical probes, most of which were considered to be of medium to high confidence by a panel of medicinal chemistry and drug discovery experts [5]. However, taking a decision on the exact range of values for the physicochemical and pharmacological properties that a small molecule should have to be considered a useful research tool to probe biology is difficult and, consequently, the definition has evolved naturally over the last few years. According to the latest definition in use by the MLSCN, compounds to be

nominated as chemical probes should comply with the following criteria: affinity below 100nM (pAffinity > 7.0) for the primary target and at least 10-fold selectivity (pSelectivity > 1.0) against related targets [5]. These potency and selectivity criteria are arguably the optimal ones but they seek to establish a certain level of confidence that the biological response observed upon using those chemical probes is due to the interaction with the assigned primary target rather than additional, often unsuspected, interactions with multiple other proteins [7].

Even though the aim of chemical biology initiatives is not to deliver clinically useful compounds, it is highly expected that the chemical probes being identified are then optimised by chemists to translate basic research discoveries into therapeutics [3]. With these expectations in mind, it is important to consider that there is mounting evidence that therapeutic efficacy is better attained via modulation of multiple proteins rather than through selective interaction with a single target [8-12]. Therefore, in the context of drug discovery, current potency and selectivity criteria for the nomination of chemical probes may need to be reassessed mainly to ensure that compounds with affinity profiles of potential therapeutic interest are not overlooked.

Do drugs qualify as chemical probes?

Drugs constitute the privileged minute portion of chemical space that has been thoughtfully optimised to attain therapeutic efficacy and thus, they can be considered the most representative set of small molecules that act as a chemical perturbation on a protein target in the context of a biological system. Accordingly, it would be interesting to explore to which extent currently known drugs fulfil current potency and selectivity criteria for chemical probes.

To investigate this aspect, a set of 2,548 drugs was compiled from four major public resources, namely, ChEMBLdb [13], PDSP [14], BindingDB [15],

and IUPHARdb [16], from which 19,250 drug-target interactions covering 1,243 individual targets were extracted. It is worth stressing that these repositories contain mainly drug-target interaction data originally generated in a large variety of laboratories that ultimately reported them in multiple bibliographic sources and thus, heterogeneity and consistency of interaction data is an issue. In this respect, if a drug had different values of the same interaction type for exactly the same target interaction (either within the same database or across databases), an average interaction value was assigned. Four interaction types were considered, namely, pKi, pKd, pIC50, and pEC50. A systematic analysis of the variations found in compounds with multiple interaction data of the same type for the same target revealed an average standard deviation of ca. 0.5 log units, irrespective of the value range. In the end, a total of 3,618 drug entries with consistent interaction data over one or multiple targets were compiled, comprising 1,633 entries with consistent pKi data, 1,609 with pIC50 data, 331 with pKd data, and 45 with pEC50 data, one drug having the possibility of being represented by more than one drug entry of consistent interaction data.

Among the 3,618 entries, 1,365 corresponded to drug entries from 890 drugs (34.9%) for which only interaction data of some type on a single target was available and thus selectivity criteria could not be applied to them. Therefore, focus was given to the remaining 2,253 drug entries from 1,658 drugs for which bioactivities for multiple targets were available in the public domain. From an interaction type perspective, they include 1083 (48.1%), 1030 (45.7%), 126 (5.6%), and 14 (0.6%) entries composed of consistent pKi, pIC50, pKd, and pEC50 data, respectively; from a target coverage perspective, they contain 572 (25.4%), 856 (38.0%), and 825 (36.6%) entries with consistent interaction data for 2, 3 to 5, and more than 5 targets, respectively.

The next step was to filter out all drug entries not having among their interaction data a value of pAffinity > 7 for at least one target. A total of 1,293

drug entries from 970 drugs remained, with almost 55% of them (706) being composed entirely of pKi data. With respect to completeness, they contain 273 (21.1%), 440 (34.0%), and 580 (44.9%) entries with consistent interaction data for 2, 3 to 5, and more than 5 targets, respectively. Finally, applying the filter of pSelectivity > 1 between their primary target and any other target for which interaction data is available, one is left with 414 drug entries representative of 368 drugs, with over 57% of them (237) containing pKi data only. From a completeness perspective, they contain 153 (37.0%), 156 (37.7%), and 105 (25.4%) entries with consistent interaction data for 2, 3 to 5, and more than 5 targets, respectively.

In summary, starting with a total number of 2,548 drugs for which interaction data was available from public sources, only 1658 drugs (65.1%) contained data for more than one target, of which 970 drugs (38.1%) fulfilled current potency criteria and 368 (14.4%) complied with both potency and selectivity criteria for chemical probes. Most interestingly, it was found that the percentage of drugs that qualify as chemical probes is severely reduced as more information on affinity data is available. As can be observed in Figure 1, 26.7% of drugs with known affinity for 2 targets would be nominated as chemical probes, whereas this percentage is reduced to just 12.7% for drug with known affinity for more than 5 targets. In addition, imposing having pAffinity > 7 for at least on target, 56.0% of drugs with known affinity for 2 targets would then qualify as chemical probes, whereas this value is drastically reduced to 18.1% for drugs with known affinity for more than 5 targets. These results emphasise the relevance of data completeness and thus, the need to perform extensive screenings on multiple targets for chemical probe nomination [7].

A detailed analysis of the 368 drugs fulfilling the current chemical probe criteria for potency and selectivity revealed that they can be classified in two different classes. The largest class is composed of 248 drugs that are characterised by having pAffinity < 7 for any target other than the primary

target. Escitalopram is a representative example of this class of drug chemical probes. As illustrated in Figure 2, this antidepressant has a strong affinity (pKi=8.78) for the sodium-dependent serotonin transporter (SERT) and shows high selectivity over the rest of targets for which interaction data is available, the largest affinity among those (pKi=5.91 for the muscarinic acetylcholine receptor M1) being clearly below current potency criteria for chemical probes.

In addition, a set of 120 drugs constitutes another class that share the property of having pAffinity > 7 for one or more secondary targets. Nortriptyline is a representative example of this second class of drug chemical probes. As can be observed in Figure 2, this second-generation tricyclic antidepressant presents strong affinity (pKi=8.85) for the sodium-dependent norepinephrine transporter (NET) and despite showing over 10-fold selectivity against the rest of targets for which interaction data is available, the affinities for SERT (pKi=7.71), serotonin receptor 5-HT_{2A} (pKi=7.59), and histamine H_1 receptor (pKi=7.22) are still above the affinity threshold defined currently as chemical probe criteria. This example highlights the fact that besides potency and selectivity criteria, one may need to define additional criteria for the maximum affinity on secondary targets as, in the context of a biological system, those affinities can be highly relevant and ultimately influence the biological response observed.

The distribution of the 414 drug entries qualified as chemical probes in the space defined by current potency and selectivity criteria is presented in Figure 3. Each circle represents a drug entry and varies with size and color: size is related to the amount of information available for each drug entry, with small, medium, and large circles marking drug entries with consistent interaction data for 2 targets, 3 to 5, and more than 5 targets, respectively; color is associated with the major protein families of therapeutic relevance, with red, blue, and yellow identifying drug entries for which the primary target is an enzyme, a G protein-coupled receptor (GPCR), or neither of both. The dashed line separates

the two different classes of drug chemical probes defined above, those being below the line corresponding to the 124 drug entries from 120 drugs that, despite complying with the selectivity criteria, have pAffinity > 7 for a protein other than the primary target.

With respect to the amount of information available, it is observed that, while 70.5% of the large circles are located within the range of low pSelectivity values between 1.0 and 2.0, 83.1% of the circles above a pSelectivity value of 2.0 correspond to small- and medium-sized circles. These results emphasise again that data completeness and drug selectivity are somehow related. With respect to protein families, almost half of the drug chemical probes (44.7%) have as primary target a GPCR protein. However, its distribution across the affinity-selectivity plane does not follow any particular trend. With the caution that data completeness imposes, based on information currently available from public sources one could conclude that the myth that drugs targeting enzymes are more selective than those targeting GPCRs or proteins from other families seems not to apply and examples of selective drugs from the different families can be found.

Beyond probing a single target

The application of the current potency and selectivity criteria for nominating chemical probes aims at identifying small molecules with high affinity for one target and clear selectivity over any other protein. Accordingly, such target-directed chemical probes will be hereafter referred to as *single probes* (Box 1). As discussed above, based on currently available public interaction data, a total of 368 drugs can be qualified as single probes, representing only 14.4% of the total number of drugs considered in this study. Close inspection of the remaining 85.6% revealed that 890 drugs (34.9%) have known interaction data for one target only, 653 drugs (25.6%) have interaction data for more than one

target but none with pAffinity > 7, 385 drugs (15.1%) have interaction data for more than one target, pAffinity > 7 for at least one target, but could not meet the selectivity criteria, and a final set of 252 drugs (9.9%) have interaction data for more than one target, pAffinity > 7 for more than one target as well, do not meet the selectivity criteria within any of the targets for which pAffinity > 7, but do meet the selectivity criteria then against any remaining target. This latter set of drugs will be referred to as *multiple probes* (Box 1).

An example of a multiple probe drug is trifluprom azine. As illustrated in Figure 4, this antipsychotic has a strong affinity (pKi=8.68) for the dopamine D2 receptor. For trifluprom azine to qualify as a single probe, the interaction with any other target should be less than 7.68 (dotted line). However, it shows also high potency (pKi=8.4) for the serotonin receptor 5-HT_{2A} (HTR2A) well within the selectivity window and thus, this target is added to the list of targets probed by this drug. At this stage, for trifluprom azine to qualify as a multiple probe, the interaction with any remaining target should be less than 7.4 (dashed line). Indeed, among the additional interaction data known at present, the most potent affinity for a target is below that threshold (pKi=7.28 for the histamine H_1 receptor, HRH1). Therefore, trifluprom azine would be finally nominated as a multiple probe of the probing profile defined by the targets HTR2A and HRH1. Since the number of probing targets would be two, trifluprom azine would be referred to as a multiple probe of level 2.

Extending the analysis beyond drugs, our interest turned then into identifying all chemical probes present in the four major public resources of interaction data [13-16], as a means to assess the current coverage of probing profiles. In total, 34,460 small molecules qualified as chemical probes, of which 27,459 are single probes for 527 targets and 7,001 are multiple probes for 959 distinct target profiles. The distribution of the number of target profiles currently covered at each probe level is illustrated in Figure 5. Contrary to the expected combinatorial explosion of possible profiles upon increasing the

number of targets, current coverage of probing profiles decreases rapidly as the probing level increases. Thus, while multiple probes for 439 probing profiles of 2 targets could be identified, only 61 probing profiles of 5 targets are currently covered. This may emphasise the traditional focus towards generating/collecting highly potent and selective compounds rather than pluripotent compounds over combinations of multiple targets.

In addition, probing profiles were assigned to major protein families on the basis of their constituent targets. In the case that all probing targets are enzymes or GPCRs, the probing profile is assigned to enzymes or GPCRs, respectively. All other probing profiles contain probing targets that belong to any of the other major therapeutic families (e.g., ion channels or nuclear receptors) or are simply a combination of them. As can be observed, enzyme probing profiles appear to be more populated than GPCR and other probing profiles at probe levels from 1 to 5, whereas probing profiles containing combinations of targets belonging to different families seem to be the most common option for probing profiles composed of more than 5 targets. Again, it should be stressed that the high degree of incompleteness of public interaction data [7] advices to take the present conclusions with caution, as improving data completeness may promote some small molecules to chemical probes but could disqualify others.

Towards a complete probing chemome

The analysis presented above on current coverage of probing profiles points to the fact that, beyond chemical probes selective for single targets, small molecules exist that probe multiple targets in a selective manner over the rest and thus, paving the way towards the systematic probing of biological targets at a systems level. The coordination of this enormous challenge can be effectively addressed by focusing on achieving complete probing of segments of biological systems, such as all members of a protein family or all proteins of a

biochemical pathway. As an illustrative example, Figure 6 contains the current probing status of the histamine receptor family.

Histamine receptors are a class of GPCRs composed of four members, namely, H1, H2, H3, and H4 [17]. Accounting for all possible combinations, complete probing of this entire class would require the identification of 14 chemical probes: 4 single probes, having high affinity for each of the individual members and selectivity over the rest, and 10 multiple probes, covering all probing profiles that can be generated from targeting several receptors at a time. By exploring the current information contents of the major public repositories of interaction data, small molecules covering 10 out of the 14 probing profiles could be identified, 5 drugs being among them. The structures and affinity profiles of these molecules are collected in Figure 6. In the interaction map provided, each row corresponds to a different combination of probing targets and the potency of small molecules for each receptor is reflected by a color gradation, black being highly potent, light grey being weekly potent, and white indicating interaction data for which no information is currently available in the public domain.

In all cases for which a representative compound is provided, drugs were given priority to other small molecules that could fit the potency and selectivity criteria for probing a given profile. Among them, cyproheptadine, tiotidine, cipralisant, and histamine were selected as the single probe representatives of the H1, H2, H3, and H4 probing profiles, respectively. In addition, impromidine was found to fit the potency and selectivity criteria as multiple probe for the probing profile defined by the H2, H3, and H4 receptors. It ought to be stressed that all molecules selected as chemical probes of the respective target profiles were identified on the basis of their affinities for the histamine receptors. Some of them were found to have gaps of interaction data for all four receptors (such as probes 1, 2, and 4 for the interaction on H2, H4, and H4, respectively) and others may have affinities also for additional proteins other than histamine

receptors. The main purpose of the current selection is to illustrate that it is conceptually possible to generate a complete set of small molecules that can either be used directly or as starting points for and optimisation process to fully probe a particular
segment of a biological system. It was recently shown that the application of *in silico* target screening to an academic chemical library allowed for identifying novel antagonists for all four members of the adenosine receptor family [18]. In this respect, the development of novel methodologies for the *in silico* target screening of molecules [19-21] is expected to have a significant impact towards assembling a complete probing chemome.

Conclusions

Biological systems are implicitly robust and selectively acting on one particular target may not be the most efficacious way of modulating or interfering with that system [22]. Indeed, recent evidences indicate that most drugs attain their *in vivo* efficacy through modulation of multiple targets rather than selective interaction on a single target. In addition, drugs represent the ultimate product of a long optimisation process in which potency and selectivity, among other pharmacokinetic and pharmacodynamic properties, are improved. Any chemical point less advanced in this process will show less optimal potency and selectivity criteria. Accordingly, if only 14.4% of drugs qualify as chemical probes, a much lower percentage is expected for the starting chemical points of drug discovery projects. Therefore, these findings are not only questioning the relevance of the nominated chemical probes as starting points for drug discovery but, most importantly, emphasising the fact that those would only be a low percentage of the number of compounds with therapeutic potential that may have come out of the screening campaigns.

While in recent years drug discovery has gradually shifted away from the

one chemical – one protein paradigm, chemical biology seems to be still very much anchored in it. Many therapeutically useful chemical probes could be missed in the process if polypharmacology is not adequately considered in the qualification criteria for small molecule tools to probe biological systems.

Acknowledgements

This research was supported by the Spanish Instituto de Salud Carlos III and the Ministerio de Ciencia e Innovación (project BIO2008-02329).

References

1 Schreiber, S. L. (2005) Small molecules: the missing link in the central dogma. *Nat. Chem. Biol.* 1, 64-66.
2 Hibert, M. F. (2009) French/European academic compound library initiative. *Drug Discov. Today* 14, 723-725.
3 Austin, C. P. *et al.* (2004) NIH molecular libraries initiative. *Science* 306, 1138-1139.
4 Kaiser, J. (2008) Industrial-style screening meets academic biology. *Science* 321, 764-766.
5 Oprea, T. I. *et al.* (2009) A crowdsourcing evaluation of the NIH chemical probes. *Nat. Chem. Biol.* 5, 441-447.
6 Wang, Y. *et al.* (2010) An overview of the PubChem BioAssay resource. *Nucl. Acids Res.* 38, D255-D266.
7 Mestres, J. *et al.* (2008) Data completeness: the Achilles heel of drug-target networks. *Nat. Biotechnol.* 26, 983-984.
8 Roth, B. L. *et al.* (2004) Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* 3, 353-359.
9 Morphy, R. *et al.* (2004) From magic bullets to designed multiple ligands. *Drug Discov. Today* 9, 641-651.
10 Hopkins, A. L. *et al.* (2006) Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* 16, 127-136.
11 Cavalli, A. *et al.* (2008) Multi-target-directed ligands to combat neurodegenerative diseases. *J. Med. Chem.* 51, 347-372.
12 Hopkins, A. L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4, 682-690.
13 ChEMBL database. http://www.ebi.ac.uk/chembldb.

14 Jensen, N. H. and Roth, B. L. (2008) Massively parallel screening of the receptorome. *Comb. Chem. High Throughput Screen.* 11, 420-427. Psychoactive Drug Screening Program. http://pdsp.med.unc.edu.
15 Liu, T. *et al.* (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* 35, D198-D201. http://www.bindingdb.org.
16 Harmar, A. J. et al. (2009) IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Res.* 37, D680-D685. http://www.iuphar-db.org.
17 Hill, S. J. et al. (1997) International Union of Pharmacology. XIII. Classification of histamine receptors. *Pharmacol. Rev.* 49, 253-278.
18 Areias, F. M. et al. (2010) In silico directed chemical probing of the adenosine receptor family. *Bioorg. Med. Chem.* 18, 3043-3052.
19 Ekins, S. et al. (2007) In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br. J. Pharmacol.* 152, 9-20.
20 Bajorath, J. (2008) Computational analysis of ligand relationships within target families. *Curr. Opin. Chem. Biol.* 12, 352-358.
21 Rognan, D. (2010) Structure-based approaches to target fishing and ligand profiling. *Mol. Inf.* 29, 176-187.
22 Kitano, H. (2007) A robustness-based approach to systems-oriented drug design. *Nat. Rev. Drug Discov.* 6, 202-210.

Box 1. Definitions

Chemical probe. A small molecule with the ability to perturb one or multiple components of a protein system giving rise to a unique biological response. Under a systems perspective, polypharmacology balances the relevance of potency and selectivity when deciding whether a small molecule qualifies as a chemical probe. In this respect, multiple pharmacological profiles may actually converge into the same biological response and thus they may all be considered redundant chemical probes; in contrast, similar pharmacological profiles with different relative affinities may result in essentially distinct biological responses.

Single probe. A small molecule with affinity $pA_1 > a$ for its primary target and selectivity $pA_1 - pA_i > s$ for any other protein $i \neq 1$, with ideally $pA_{i \neq 1} \leq a$. According to the latest definition in use by the MLSCN, $a = 7$ and $s = 1$ [5].

Multiple probe. A small molecule with affinity $pA_{\{n\}} > a$ for a set of $\{n\}$ targets and selectivity $\min(pA_{\{n\}}) - pA_i > s$ for any other protein $i \neq \{n\}$, with ideally $pA_{i \neq \{n\}} \leq a$. To be consistent with the current definition of a single probe, the values of the parameters are also $a = 7$ and $s = 1$.

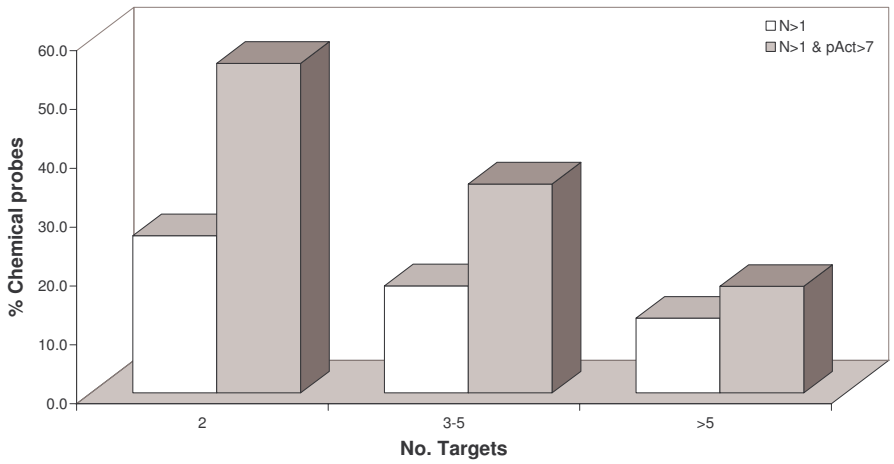


Figure 1. Probability of drugs to qualify as chemical probes depending on information known about their target profile

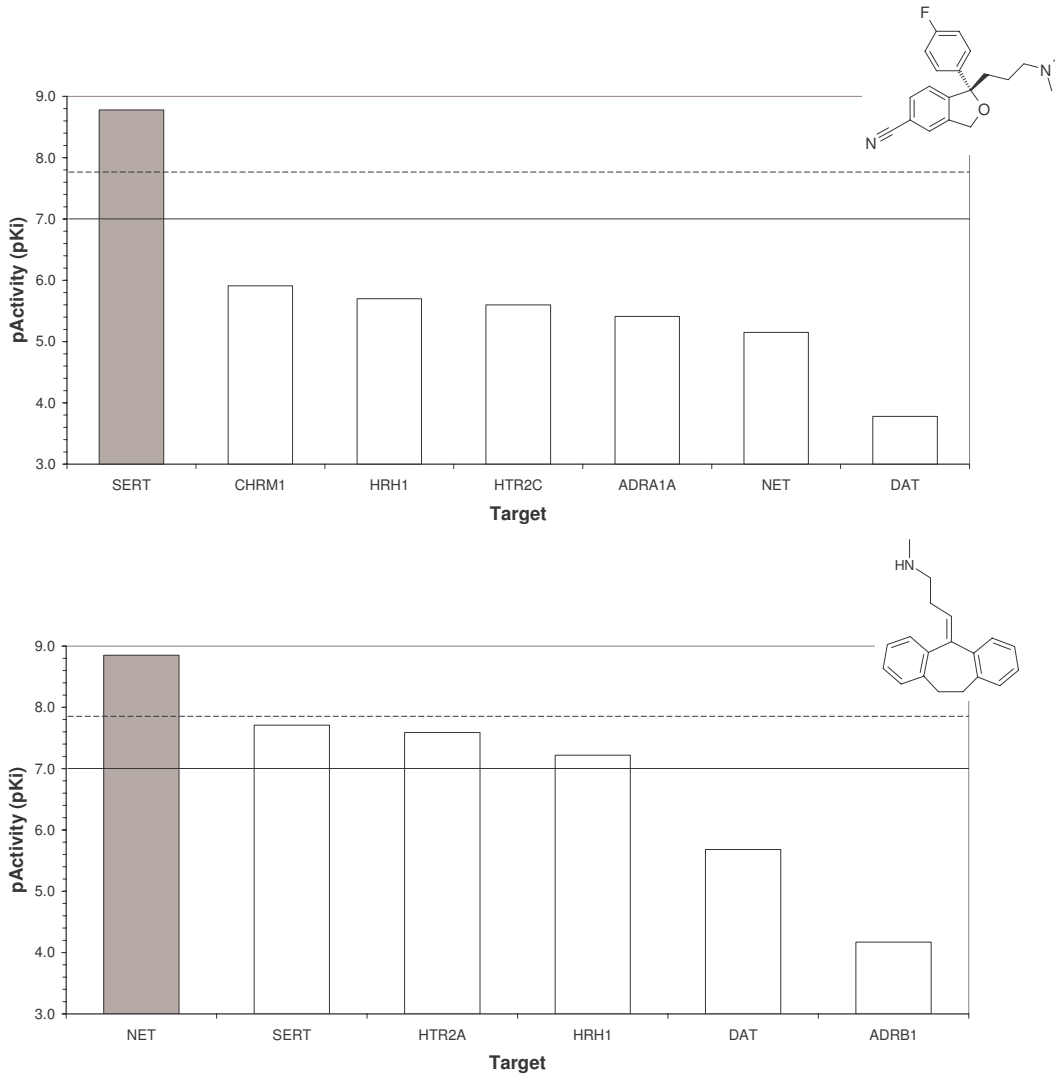


Figure 2. Escitalopram (top) and nortriptyline (bottom), two examples of drugs acting as chemical probes

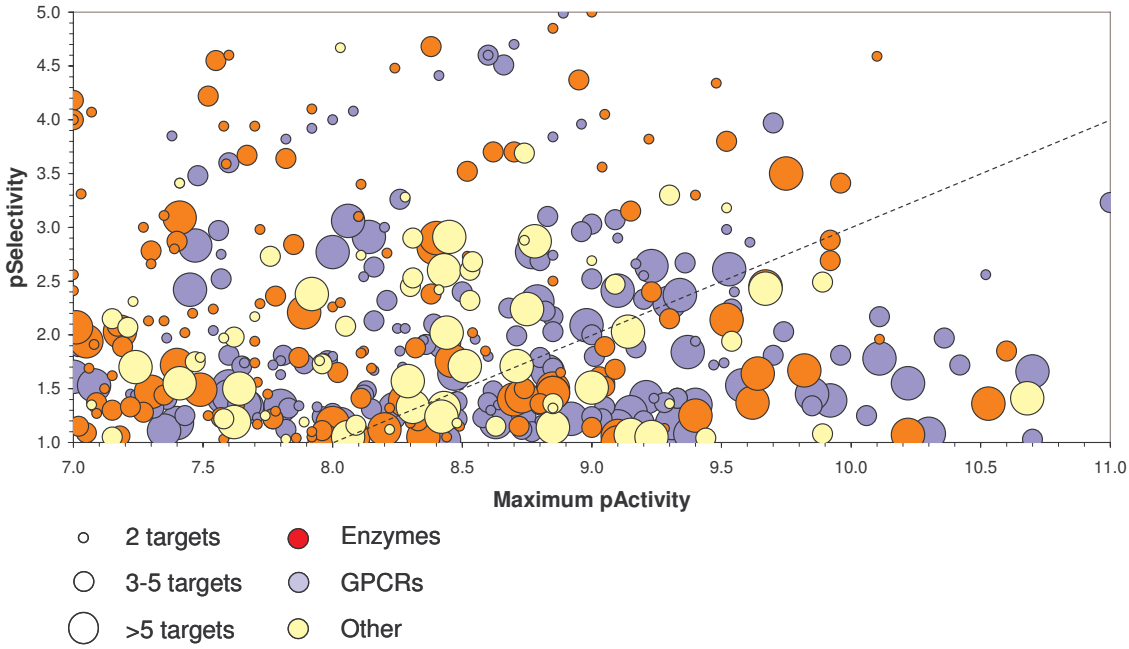


Figure 3. Distribution of chemical probes in the space defined by current potency and selectivity criteria. Size of circles denote degree of information available. Color of circles related to the major protein family associated with the primary target. The dashed line separates the two classes of chemical probes (see text for details).

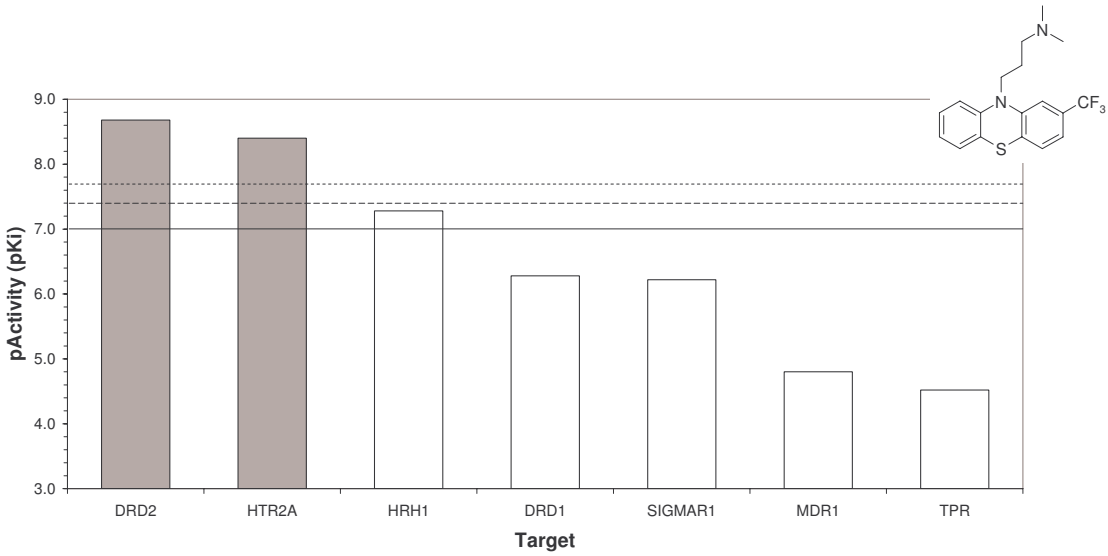


Figure 4. Triflupromazine, an example of a drug acting as a multiple probe.

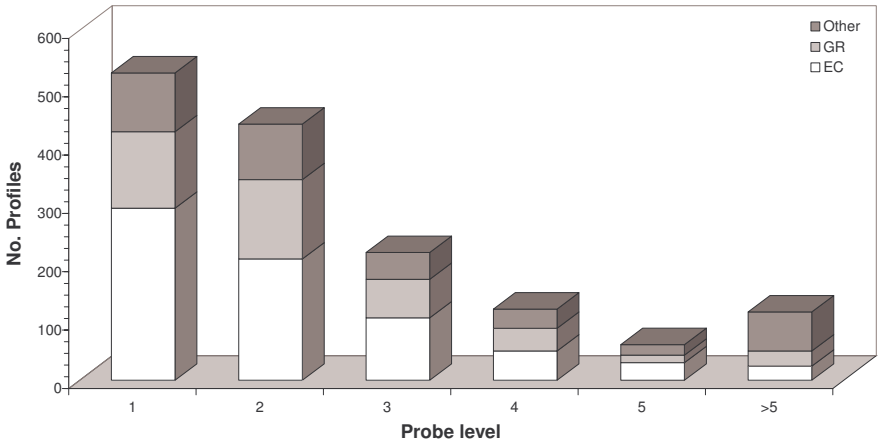


Figure 5. Current profile coverage by chemical probes. Probe level indicates the number of targets involved in the probing profile

0

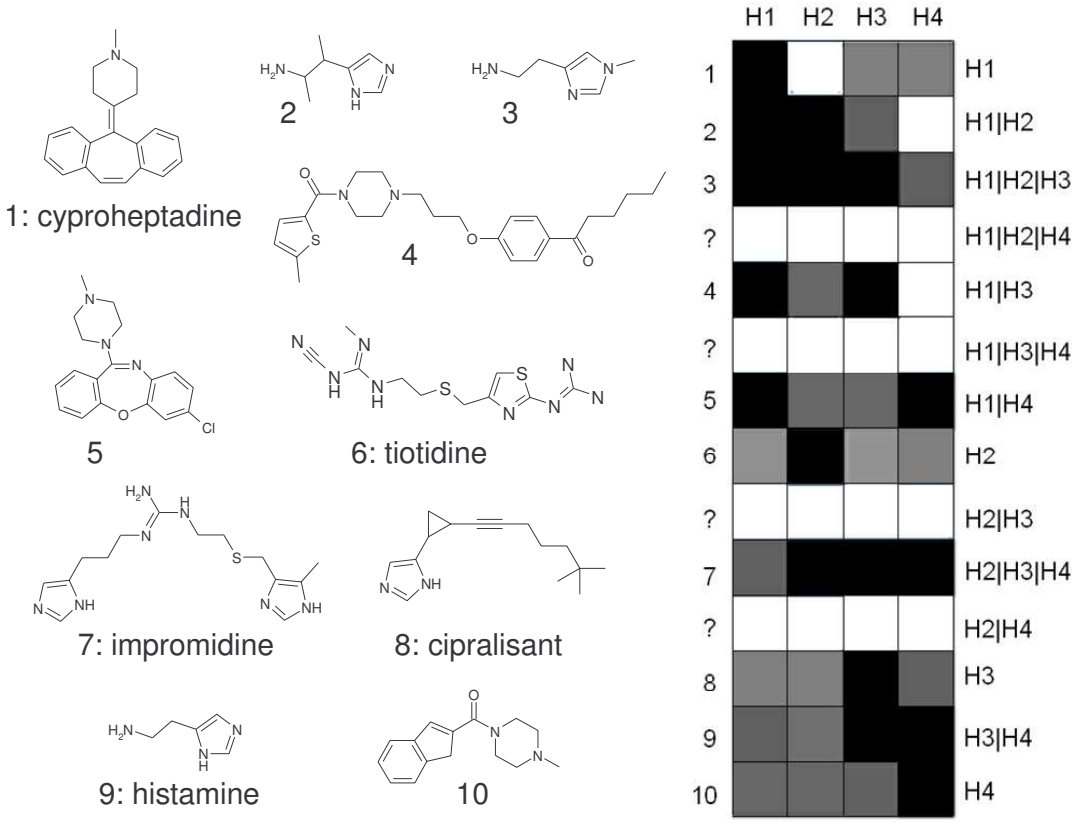


Figure 6. Current status towards a complete chemical probing of the family of histamine receptors

Part V - Conclusions

1. The contribution to the construction of an annotated compound library directed to the family of nuclear receptors and the integration of hierarchical classification schemes for both ligands and targets led to the identification of highly promiscuous scaffolds with activity over specific groups of these proteins. These scaffolds could be used as the basis for further synthesis of compound libraries to probe orphan nuclear receptors.

2. The nuclear receptor chemical library and other public repositories of ligand-target interaction data have been parsed and added to an integrated database currently containing 824.000 interaction values between 240.000 unique ligands and more than 4.000 protein targets. This database has been used as the basis for the analysis of the polypharmacology of drugs and the development of ligand-based methods for *in silico* target profiling.

3. An analysis of these ligand-target interaction data reveals that the completeness issue is a key limitation in computational drug discovery. A strategy to overcome this problem has been posited and a set of complete interaction matrices from public sources has been extracted to be used as benchmark sets for *in silico* target profiling methodologies.

4. An analysis of the currently known polypharmacology of drugs emphasizes that less than 15% of drugs would actually qualify as chemical probes under the current criteria for potency and selectivity. If chemical probes are to be used as starting points for drug discovery process both criteria are proposed to be revised.

5. Three two-dimensional descriptors, namely, SHED, FPD, and PHRAG, have been successfully combined in a new ligand based approach to *in silico* target profiling that exploits the ligand-protein interaction data contained in the integrated database. Recent applications of these approaches to *in silico* pharmacology have provided ample evidence of the key impact that these

computational methods are having in both chemical biology and drug discovery.

6. The development of FCP was conceived as a web-based tool to facilitate the graphical and quantitative analysis of the current state and trends in the functional coverage and bias of the solved structures deposited in the Protein Data Bank, by making use of the substantial efforts made by a number of researchers in our laboratory in developing and curating classification schemes for proteins belonging to different families.

7. The development of iPHACE was designed to provide a deeper understanding on the polypharmacology of drugs and the cross-pharmacology of targets, through navigation of selected annotated chemical libraries containing highly curated drug target interaction data by meas of a purposely designed visualization framework that integrates all chemical and biological data in a simple and efficient manner.

8. The creation of a network linking drugs depending on their assigned side effects profile, and its integration with the information contained in the previously generated ligand-target interaction database, shed light into the relations between drug adverse reaction profiles and their corresponding target profiles. The fact that drugs connected in the side effect network are found to have similar target profiles suggests that preclinical comparative pharmacology may represent an interesting strategy for anticipating drug adverse reactions.

Part VI - References

1. Hiroaki Kitano: **Systems Biology: A Brief Overview. Science** 2002, **295**:1662-4

2. Eric E. Schadt Æ Bin Zhang Æ Jun Zhu**: Advances in systems biology are enhancing our understanding of disease and moving us closer to novel disease treatments**. Genetica 2009, **136**:259–269

3. Lingchong You: **Toward Computational Systems Biology**. Cell Biochemistry and Biophysics 2004, **40**(2):167-184

4. N. M. Luscombe, D. Greenbaum, M. Gerstei**: What is Bioinformatics? A Proposed Definition and Overview of the Field**. Method Inform Med 2001, **40**:346–58

5. Westerhoff HV, Kolodkin A, Conradie R, Wilkinson SJ, Bruggeman FJ, Krab K, van Schuppen JH, Hardin H, Bakker BM, Moné MJ, Rybakova KN, Eijken M, van Leeuwen HJ, Snoep JL.: **Systems biology towards life in silico: mathematics of the control of living cells**. J. Math. Biol. 2009, **58**:7–34

6. Peter V. Coveney, Philip W. Fowler: **Modelling biological complexity: a physical scientist's perspective**. J. R. Soc. Interface 2005, **2**:267–280

7. Searls DB.: **Data integration: challenges for drug discovery**. Nat Rev Drug Discov. 2005, **4**(1):45-58.

8. Nils Gehlenborg, Seán I O'Donoghue Nitin S Baliga, Alexander Goesmann, Matthew A Hibbs, Hiroaki Kitano, Oliver Kohlbacher, Heiko Neuweger, Reinhard Schneider, Dan Tenenbaum & Anne-Claude Gavin: **Visualization of omics data for systems biology**. Nature methods supplement 2010, **7**(3);56-69

9. Carole Goble, Robert Stevens: **State of the nation in data integration for bioinformatics**. Journal of Biomedical Informatics 2008, **41**:687–693

10. Stephen P.Gardner: **Ontologies and semantic data integration**. Drug Discovery Today 2005, **10**(14):1001-1007

11. Michael Ashburner, Catherine A. Bal, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry,Allan P. Davis,Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin & Gavin Sherlock: **Gene ontology: tool for the unification of biology**. Nature Genetics 2000, **25**:25–9.

12. Daniel L. Rubin, Nigam H. Shah and Natalya F. Noy: **Biomedical ontologies: a functional perspective**. Briefings in bioinformatics 2007, **9**(1):75-90

13. Milne, G.M: **Pharmaceutical productivity – the imperative for new paradigms**. Annu. Rep. Med. Chem. 2003, **38**:383–396

14. Tae Yong Kim, Hyun Uk Kim, Sang Yup Lee: **Data integration and analysis of biological networks**. Current Opinion in Biotechnology 2010, **21**:78–84

15. T. Slater, C. Bouton, E. Huang: **Beyond data integration**. Drug Discovery Today 2008, **13**(13):584-589

16. Valdur Saks, Claire Monge and Rita Guzun: **Philosophical Basis and Some Historical Aspects of Systems Biology: From Hegel to Noble** - Applications for Bioenergetic Research. Int. J. Mol. Sci. 2009, **10**:1161-1192

17. Chuming Chen, Peter B.McGarvey, Hongzhan Huang, Cathy H.Wu: **Protein Bioinformatics Infrastructure for the Integration and Analysis of Multiple High-Throughput "omics" Data**. Advances in Bioinformatics 2010, ID 423589, 19 pages

18. Sangchul Rho, Sungyong You, Yongsoo Kim, Daehee Hwang: **From proteomics toward systems biology: integration of different types of proteomics data into network models**. BMB reports 2008, **41**(3):184-193

19. Angelo Nuzzo1, Alberto Riva, Riccardo Bellazzi: **Phenotypic and genotypic data integration and exploration through a web-service architecture**. BMC Bioinformatics 2009, **10**(12):S5

20. Stefano Bianchi, Anna Burla, Costanza Conti, Ariel Farkash, Carmel Kent, Yonatan Maman, Amnon Shabo: **Biomedical Data Integration – Capturing Similarities while Preserving Disparities**. Conf Proc IEEE Eng Med Biol Soc. 2009, **2009**:4654-4657.

21. M. Edelstein a; F. Buchwald a; L. Richter a;S. Kramer: **Integrating background knowledge from internet databases into predictive toxicology models**. SAR and QSAR in Environmental Research 2010, **21**:21–35

22. W. Patrick Walters, Matthew T. Stahl and Mark A. Murcko: **Virtual screening - an overview**. Drug Discovery Today 1998, **3**(4):160-178

23. Bohacek, R. S., McMartin, C. & Guida, W. C.: **The art and practice of structure-based drug design: a molecular modelling perspective**. Med. Res. Rev. 1996, **16**:3–50.

24. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ.: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings**. Adv Drug Deliv Rev. 2001, **46**(1-3):3-26.

25. Susumu Goto Yasushi Okuno, Masahiro Hattori, Takaaki Nishioka, Minoru Kanehisa: **LIGAND: database of chemical compounds and reactions in biological pathways**. Nucleic Acids Research 2002, **30**(1):402-404

26. Christopher Lipinski & Andrew Hopkins: **Navigating chemical space for biology and medicine**. Nature 2004, **432**(7019):855-61.

27. Cao Y, Jiang T, Girke T.: **A maximum common substructure-based algorithm for searching and predicting drug-like compounds**. Bioinformatics 2008, **24**(13):366-74.

28. Willett P. Br: **Similarity-based virtual screening using 2D fingerprints**. Drug Discov Today 2006, **11**(23-24):1046-53

29. D. Rognan: **Chemogenomic approaches to rational drug design**. J Pharmacol. 2007,**152**(1): 38–52

30. Bemis GW, Murcko MA.: **The properties of known drugs. 1. Molecular frameworks**. J Med Chem. 1996, **39**(15):2887-93.

31. Marcus A. Koch, Ansgar Schuffenhauer, Michael Scheck, Stefan Wetzel, Marco Casaulta, Alex Odermatt, Peter Ertl, Herbert Waldmann: **Charting biologically relevant chemical space: A structural classification of natural products (SCONP)**. PNAS 2005, **102**(48):17272–17277

32. Böcker A.: **Toward an improved clustering of large data sets using maximum common substructures and topological fingerprints**. J Chem Inf Model. 2008, **48**(11):2097-107.

33. Robert P. Sheridan, Simon K. Kearsley: **Why do we need so many chemical similarity search methods?** Drug Discovery Today 2002, **7**(17): 903-11

34. Li JW, Vederas JC**.: Drug discovery and natural products: end of an era or an endless frontier?** Science 2009, **325**(5937):161-5.

35. Alan L. Harvey: **Natural products in drug discovery**. Drug Discovery Today 2008, **13**(19/20):894-901

36. Kristina Grabowski, Karl-Heinz Baringhausb, Gisbert Schneider: **Scaffold diversity of natural products: inspiration for combinatorial library design**. Nat. Prod. Rep. 2008, **25**:892–904

37. Thomas Scior, Philippe Bernard, José Luis Medina-Franco, Gerald. M. Maggiora: **Large Compound Databases for Structure-Activity Relationships Studies in Drug Discovery**. Mini-Reviews in Medicinal Chemistry 2007, **7**:851-860

38. Bondensgaard K, Ankersen M, Thøgersen H, Hansen BS, Wulff BS, Bywater RP.: **Recognition of privileged structures by G-protein coupled receptors**. J. Med. Chem. 2004, **47**:888–899

39. Stephen A Hitchcock: **Blood–brain barrier permeability considerations for CNS-targeted compound library design**. Current Opinion in Chemical Biology 2008, **12**:318–323

40. E. Jacoby, A. Mozzarelli: **Chemogenomic Strategies to Expand the Bioactive Chemical Space**. Current Medicinal Chemistry

2009, **16**:4374-4381

41. Elisabet Gregori-Puigjané, Jordi Mestres: Coverage **and bias in chemical library design**. Current Opinion in Chemical Biology 2008, **12**:359–365

42. Christoph M. Huwe: **Synthetic library design**. Drug Discovery Today 2006, **11**(15/16):763-767

43. Jared T. Shaw: **Naturally diverse: highlights in versatile synthetic methods enabling target- and diversity-oriented synthesis**. Nat. Prod. Rep. 2009, **26**:11–26

44. Valerie J Gillet: New **directions in library design and analysis**. Current Opinion in Chemical Biology 2008, **12**:372–378

45. Mireille Krier, Guillaume Bret, Didier Rognan: **Assessing the Scaffold Diversity of Screening Libraries**. J. Chem. Inf. Model. 2006, **46**:512-524

46. Richard J. Spandl, Andreas Benderb, David R. Spring: **Diversity-oriented synthesis; a spectrum of approaches and results**. Org. Biomol. Chem. 2008, **6**:1149–1158

47. N. Baurin, R. Baker, C. Richardson, I. Chen, N. Foloppe, A. Potter, A. Jordan, S. Roughley, M. Parratt, P. Greaney, D. Morley, and R. E. Hubbard Vernalis: **Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds**. J. Chem. Inf. Comput. Sci. 2004, **44** (2):643–651

48. website: http://www.ebi.ac.uk/chebi/

49. Paula de Matos, Rafael Alcántara, Adriano Dekker, Marcus Ennis, Janna Hastings,Kenneth Haug, Inmaculada Spiteri, Steve Turner, Christoph Steinbeck: **Chemical Entities of Biological Interest: an update**. Nucleic Acids Research 2010, **38**(Database issue):249-254

50. website: http://pubchem.ncbi.nlm.nih.gov/

51. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH.: **PubChem: a public information system for analyzing bioactivities of small molecules**. Nucleic Acids Res. 2009, **1**:37

52. Konstantin V. Balakin, Alexander V. Kozintsev, Alex S. Kiselyov and Nikolay P. Savchuk: **Rational Design Approaches to Chemical Libraries for Hit Identification**. Current Drug Discovery Technologies 2006, **3**:49-65

53. Nikolay P Savchuk_, Konstantin V Balakin, Sergey E Tkachenko: **Exploring the chemogenomic knowledge space with annotated chemical libraries**. Current Opinion in Chemical Biology 2004, **8**:412–417

54. Harmar AJ, Hills RA, Rosser EM, Jones M, Buneman OP, Dunbar DR, Greenhill SD, Hale VA, Sharman JL, Bonner TI, Catterall WA, Davenport AP, Delagrange P, Dollery CT, Foord SM, Gutman GA, Laudet V, Neubig RR, Ohlstein EH, Olsen RW, Peters J, Pin JP, Ruffolo RR, Searls DB, Wright MW and Spedding M.: **IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels**. Nucl. Acids Res. 2009, **37**(Database issue):680-685

55. website: http://www.embl.org

56. website: http://www.ebi.ac.uk/chembl/

57. website: http://pdsp.med.unc.edu/indexR.html

58. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Oprea TI: **WOMBAT: world of molecular bioactivity**. Chemoinformatics in Drug Discovery. Edited by Wiley-VCH; 2004:223-239.

59. website: http://www.prous.com/product/electron/mddr.html

60. website: http://www.iupac.org/inchi/

61. website: http://openbabel.org

62. website: http://www.chemaxon.com/

63. Levitt M.: Nature **of the protein universe**. Proc Natl Acad Sci U S A. 2009, **106**(27):11079-84

64. The UniProt Consortium: **The Universal Protein Resource (UniProt)**. Nucleic Acids Research 2009, **37**(Database issue):D169–D174

65. Paul D. Dobson, Yu-Dong Cai, Benjamin J. Stapley and Andrew J. Doig: **Prediction of Protein Function in the Absence of Significant Sequence Similarity**. Department of Current Medicinal Chemistry 2004, **11**:2135-2142

66. Russell L. Marsden, Juan A. G. Ranea, Antonio Sillero, Oliver Redfern, Corin Yeats, Michael Maibaum, David Lee, Sarah Addou, Gabrielle A. Reeves, Timothy J. Dallman, Christine A. Orengo: **Exploiting protein structure data to explore the**

evolution of protein function and biological complexity. Phil. Trans. R. Soc. B 2006, **361**:425–440

67. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes**. J Mol Biol. 2001, **310**(2):311-25

68. Christine A. Orengo, Janet M. Thornton: **Protein families and their evolution: a structural perspective**. Annu. Rev. Biochem. 2005, **74**:867–900

69. Greene,L.H., Lewis,T.E., Addou,S., Cuff,A., Dallman,T., Dibley,M., Redfern,O., Pearl,F., Nambudiry,R., Reid,A., Sillitoe,I., Yeats,C., Thornton,J.M., Orengo,C.A.: **The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution**. Nucleic Acids Research 2007, **35**:291-297

70. R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunesekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman: **The Pfam protein families database**. Nucleic Acids Research 2010 Database Issue **38**:D211-222

71. Sleator RD, Walsh P.: **An overview of in silico protein function prediction**. Arch Microbiol. 2010, **192**(3):151-5

72. James D Watson, Roman A Laskowski and Janet M Thornton: **Predicting protein function from sequence and structural data**. Current Opinion in Structural Biology 2005, **15**:275–284

73. Roy D. Sleator, Paul Walsh: **An overview of in silico protein function prediction**. Arch Microbiol 2010, **192**:51–155

74. **Enzyme Nomenclature 1992** [Academic Press, San Diego, California, ISBN 0-12-227164-5 (hardback), 0-12-227165-3 (paperback)] with Supplement 1 (1993), Supplement 2 (1994), Supplement 3 (1995), Supplement 4 (1997) and Supplement 5 (in Eur. J. Biochem., 1994; **223**:1-5; Eur. J. Biochem., 1995; **232**:1-6; Eur. J. Biochem., 1996: **237**:1-5; Eur. J. Biochem., 1997; **250**:1-6, and Eur. J. Biochem., 1999; **264**:610-650; respectively)

75. Access the complete list at http://www.iuphar.org/nciuphar_arti.html

76. Paul D Thomas, Huaiyu Mi, Suzanna Lewis: **Ontology annotation: mapping genomic regions to biological function**. Current Opinion in Chemical Biology 2007, **11**:4–11

77. Mi, H.Y. et al.: **Assessment of genome-wide protein function classification for Drosophila melanogaster**. Genome Res. 2003, **13**:2118–2128

78. S.C.E. Tosatto, S. Toppo: **Large-Scale Prediction of Protein Structure and Function from Sequence**. Current Pharmaceutical Design 2006, **12**:2067-2086

79. Wu CH, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu ZZ, Ledley RS, Lewis KC, Mewes HW, Orcutt BC, Suzek BE, Tsugita A, Vinayaka CR, Yeh LS, Zhang J, Barker WC.: **The Protein Information Resource: an integrated public resource of functional annotation of proteins**. Nucleic Acids Res. 2002, **30**(1):35-7.

80. Kanehisa M.: **The KEGG database**. Novartis Found Symp. 2002, **247**:91-101

81. Andriy Kryshtafovych, Krzysztof Fidelis: **Protein structure prediction and model quality assessment**. Drug Discovery Today 2009, **14**(7-8):386-93

82. Yang Zhang: **Protein structure prediction: when is it useful?** Current Opinion in Structural Biology 2009, **19**:145–155

83. Eugene V. Koonin, Yuri I. Wolf & Georgy P. Karev: **The structure of the protein universe and genome evolution**. Nature. 2002, **420**(6912):218-23.

84. Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM, Wilson IA, Godzik A**.: Exploration of uncharted regions of the protein universe**. PLoS Biol. 2009, **7**(9):e1000205

85. Mestres J. : **Representativity of target families in the Protein Data Bank: impact for family-directed structure-based drug discovery**. Drug Discov Today 2005, **10**(23-24):1629-37

86. Melanie A. Adams, Michael D. L. Suits, Jimin Zheng, Zongchao Jia: **Piecing together the structure–function puzzle: Experiences in structure-based functional annotation of hypothetical proteins**. Proteomics 2007, **7**:2920–2932

87. Babu A. Manjasetty, Andrew P. Turnbull, Santosh Panjikar, Konrad Büssow, Mark R. Chance: **Automated technologies and novel techniques to accelerate protein crystallography for structural genomics**. Proteomics 2008, **8**:612–625

88. Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer E.F., Brice M.D., Rodgers J.R., Kennard O., Shimanouchi T, Tasumi M: **Protein data bank - computer-based archival file for macromolecular structures**. Journal of Molecular Biology 1977, **112**(3):535-542

89. Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, Philip E. Bourne: **The Protein Data Bank**. Nucleic Acids Research 2000, **28**:1235-242

90. Martin Caffrey and Vadim Cherezov: **Crystallizing Membrane Proteins Using Lipidic Mesophases**. Nat Protoc. 2009, **4**(5):706–731

91. Bourne PE, Allerston CK, Krebs W, Li W, Shindyalov IN, Godzik A, Friedberg I, Liu T, Wild D, Hwang S, Ghahramani Z, Chen L, Westbrook J**.: The status of structural genomics defined through the analysis of current targets and structures**. Pac Symp Biocomput 2004, 375-86.

92. Toni Gabaldón: **Large-scale assignment of orthology: back to phylogenetics?** Genome Biology 2008, **9**:235

93. David Eisenberg, Edward M. Marcotte, Ioannis Xenarios & Todd O. Yeates: **Protein function in the post-genomic era**. Nature 2000, **405**(6788):823-6.

94. Chothia C, Lesk AM.: **The relation between the divergence of sequence and structure in proteins**. Embo J 1986, **5**:823-6.

95. Roman L. Tatusov, Eugene V. Koonin, David J. Lipman: **A Genomic Perspective on Protein Families**. Science 1997, **278**(5338):631-7

96. Elijah Roberts, John Eargle, Dan Wright, Zaida Luthey-Schulten: **MultiSeq: unifying sequence and structure data for evolutionary analysis**. BMC Bioinformatics 2006, **7**:382

97. Tatsuya Fukuda, Jun Yokoyama, Toru Nakamura, In-Ja Song, Takuro Ito, Toshinori Ochiai, Akira Kanno, Toshiaki Kameya, Masayuki Maki: **Molecular phylogeny and evolution of alcohol dehydrogenase (Adh) genes in legumes**. BMC Plant Biology 2005, **5**:6

98. G. Manning, D.B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam: **The Protein Kinase Complement of the Human Genome**. Science 2002, **298**:1912-1934

99. Novac N, Heinzel T.: **Nuclear receptors: overview and classification**. Curr Drug Targets Inflamm Allergy. 2004, **3**(4):335-46.

100. Nuclear Receptors Nomenclature Committee: **A unified nomenclature system for the nuclear receptor superfamily**. Cell. 1999, **97**(2):161-3.

101. Saier MH, Tran CV, Barabote RD: **TCDB: the Transporter Classification Database for membrane transport protein analyses and information**. Nucleic Acids Res. 2006, **34**(Database issue):181-6

102. Saier MH Jr, Yen MR, Noto K, Tamang DG, Elkan C.: **The Transporter Classification Database: recent advances**. Nucleic Acids Res. 2009, **37**(Database issue):274-8

103. Violaine Pillet, Marc Zehnder, Alexander K. Seewald, Anne-Lise Veuthey, Johann Petrak: **GPSDB: a new database for synonyms expansion of gene and protein names**. Bioinformatics 2005, **21**:1743-1744.

104. Johannes Goll, Robert Montgomery, Lauren M. Brinkac, Seth Schobel, Derek M. Harkins, Yinong Sebastian, Susmita

Shrivastava, Scott Durkin, Granger Sutton: **The Protein Naming Utility: a rules database for protein nomenclature**. Nucleic Acids Research 2010, **38**(Database issue):D336-9

105. **The UniProt Consortium: The Universal Protein Resource (UniProt) in 2010**. Nucleic Acids Research 2010, **38**(Database issue):142-8

106. Cathy H. Wu, Hongzhan Huang, Lai-Su L. Yeh, Winona C. Barker: **Protein family classification and functional annotation**. Computational Biology and Chemistry 2003, **27**:37-47

107. Marianne A. Grant: **Protein Structure Prediction in Structure-Based Ligand Design and Virtual Screening**. Combinatorial Chemistry & High Throughput Screening, 2009, **12**(10):940-60.

108. Cavasotto CN, Phatak SS: **Homology modeling in drug discovery: current trends and applications**. Drug Discovery Today 2009, **14**(13-14):676-83

109. Yaniv Loewenstein, Domenico Raimondo, Oliver C Redfern, James Watson, Dmitrij Frishman, Michal Linial, Christine Orengo, Janet Thornton, and Anna Tramontano: **Protein function annotation by homology-based inference**. Genome Biol. 2009, **10**(2): 207.

110. Scheeff ED, Bourne PE: **Application of protein structure alignments to iterated hidden Markov model protocols for structure prediction**. BMC Bioinformatics 2006, **7**:410.

111. Dunbrack RL Jr.: **Sequence comparison and protein structure prediction**. Curr Opin Struct Biol. 2006, **16**(3):374-84.

112. Kundrotas PJ, Lensink MF, Alexov E: **Homology-based**

modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles. Int J Biol Macromol. 2008, **43**(2):198-208.

113. M I Sadowski, D T Jone: **The sequence–structure relationship and protein function prediction**. Current Opinion in Structural Biology 2009, **19**:357–362

114. Andrzej Joachimiak: **High-throughput crystallography for structural genomics**. Current Opinion in Structural Biology 2009, **19**:573-584

115. Kamal Azzaoui, Jacques Hamon, Bernard Faller, Steven Whitebread, Edgar Jacoby, Andreas Bender, Jeremy L. Jenkins, Laszlo Urban**: Modeling Promiscuity Based on in vitro Safety Pharmacology Profiling Data**. ChemMedChem 2007, **2**(6):874-880

116. Mestres J, Gregori-Puigjané E, Valverde S, Solé RV.: **The topology of drug-target interaction networks: implicit dependence on drug properties and target families**. Mol Biosyst. 2009, **5**(9):1051-7.

117. Zahra Pourpak, Mohammad R. Fazlollahi, Fatemeh Fattahi: **Understanding Adverse Drug Reactions and Drug Allergies: Principles, Diagnosis and Treatment Aspects**. Recent Patents on Inflammation & Allergy Drug Discovery 2008, **2**:24-46

118. Emma C. Davies, Christophe F. Green, David R. Mottram, Munir Mohamed: **Adverse Drug Reactions in Hospitals: A Narrative Review**. Current Drug Safety 2007, **2**:79-87

119. Liebler DC, Guengerich FP.: **Elucidating mechanisms of drug-induced toxicity**. Nat Rev Drug Discov. 2005, **4**(5):410-20.

120. Guengerich FP, MacDonald JS.: **Applying mechanisms of chemical toxicity to predict drug safety**. Chem Res Toxicol. 2007, **20**(3):344-69

121. Eric A.G. Blomme, Yi Yang, Jeffrey F. Waring: **Use of toxicogenomics to understand mechanisms of drug-induced hepatotoxicity during drug discovery and development**. Toxicology Letters 2009, **186**:22–31

122. FO Ajayi, H Sun and J Perry: **Adverse drug reactions: a review of relevant factors**. J. Clin. Pharmacol. 2000; **40**:1093-101

123. Thomas J. Moore, AB; Michael R. Cohen, RPh, MS, ScD; Curt D. Furberg, MD: **Serious Adverse Drug Events Reported to the Food and Drug Administration**. 1998-2005. Arch Intern Med. 2007, **167**(16):1752-1759

124. Schuster, D., Laggner, C., Langer, T.: **Why drugs fail—a study on side effects in new chemical entities**. Curr. Pharm. Des. 2005, **11**:3545–3559

125. Scheiber J, Jenkins JL, Sukuru SC, Bender A, Mikhailov D, Milik M, Azzaoui K, Whitebread S, Hamon J, Urban L, Glick M, Davies JW.: **Mapping adverse drug reactions in chemical space**. J Med Chem. 2009, **52**(9):3103-7.

126. Richard M. LoPachin and Anthony P. DeCaprio: **Protein Adduct Formation as a Molecular Mechanism in Neurotoxicity**. Toxicological sciences 2005, **86**(2), 214–225.

127. Shufeng Zhou, Ph.D. and Eli Chan, Ph.D.: **Drug bioactivation, covalent binding to target proteins and toxicity relevance**. Drug Metabolism Reviews 2005, **1**:41–213.

128. Hartmut Jaeschke and Mary Lynn Bajt: **Intracellular Signaling Mechanisms of Acetaminophen-Induced Liver Cell Death**. Toxicological sciences 2006, **89**(1):31–41.

129. Yuan Gao, Ricky D. Holland, Li-Rong Yu: **Quantitative proteomics for drug toxicity**. Brief Funct Genomic Proteomic. 2009, **8**(2):158-66

130. Feng Ge & Qing-Yu He: **Genomic and proteomic approaches for predicting toxicity and adverse drug reactions**. Expert Opin Drug Metab Toxicol. 2009, **5**(1):29-37

131. B. Alex Merrick and Frank A. Witzmann: **The role of toxicoproteomics in assessing organ specific toxicity**. EXS. 2009, 99:367–400.

132. Florian Nigsch, NJ Maximilan Macaluso, John BO Mitchell, Donatas Zmuidinavicius: **Computational toxicology: an overview of the sources of data and of modelling methods**. Expert Opin. Drug Metab. Toxicol. 2009, **5**(1):1-14

133. Wolfgang Muster, Alexander Breidenbach, Holger Fischer, Stephan Kirchner, Lutz Müller, Axel Pähler**: Computational toxicology in drug development**. Drug Discovery Today 2008, **13**(7/8):303-10

134. Thomas Hartung: **Toxicology for the twenty-first century**. Nature. 2009, **460**(7252):208-12

135. Luis G. Valerio Jr**.: In silico toxicology for the pharmaceutical sciences**. Toxicology and Applied Pharmacology 2009, **241**:356–370

136. Carolyn J. Mattingly: **Chemical databases for environmental health and clinical research**. Toxicology Letters 2009, **186**:62–65

137. Zhu F, Han BC, Pankaj Kumar, Liu XH, Ma XH, Wei XN, Huang L, Guo YF, Han LY, Zheng CJ, Chen YZ.: **Update of TTD: Therapeutic Target Database**. Nucleic Acids Res. 2010, **38**(Database issue):787-91

138. Fonger GC: **Hazardous substances data bank (HSDB) as a source of environmental fate information on chemicals**. Toxicology 1995, **103**(2):137-45.

139. Sangkuhl K, Berlin DS, Altman RB, Klein TE.: **PharmGKB: understanding the effects of individual genetic variants**. Drug Metab Rev. 2008, **40**(4):539-51.

140. website: http://dailymed.nlm.nih.gov/dailymed/about.cfm

141. website: http://www.nlm.nih.gov/medlineplus/

142. website: http://www.medscape.com/

143. website: http://www.drugs.com/

144. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ, Schneider R, Skoblo R, Russell RB, Bourne PE, Bork P, Preissner R.: **SuperTarget and Matador: resources for exploring drug-target relationships**. Nucleic Acids Res. 2008, **36**(Database issue):919-22.

145. Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, Peer Bork: **A side effect resource to capture phenotypic effects of drugs**. Molecular Systems Biology 2010, **6**:343

146. Montserrat Cases, Jordi Mestres: **A chemogenomic approach to drug discovery: focus on cardiovascular diseases**. Drug Discovery Today 2009, **14**(9-10):479-85

147. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M.: **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. Nucleic Acids Res. 2008, **36**(Databaseissue):901-6

148. Rubin DL, Thorn CF, Klein TE, Altman RB.: **A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge**. J Am Med Inform Assoc. 2005, **12**(2):121-9.

149. Gary H. Merrill: Concepts **and Synonymy in the UMLS Metathesaurus**. Journal of Biomedical Discovery and Collaboration 2009, **4**:7

150. website: http://www.nlm.nih.gov/mesh/