5

Unifying Bioinformatics and Chemoinformatics for Drug Design

J.B. Brown and Yasushi Okuno
Kyoto University Graduate School of Pharmaceutical Sciences
Department of Systems Bioscience for Drug Discovery
Japan

1. Overview

Until relatively recently, the field of drug design and development was disconnected from advances in computing and informatics. Without even considering the concept of computing, previous generations of scientists and clinicians had great motivation to examine the symptoms of ill or injured patients, infer from sufficient observation data about the causes of their symptoms, and search for chemical remedies that could cure or somewhat allieviate a person's ailment. Today, remedies come from sources such as herbal medicines, a high-quality nutritional diet, or human-designed medicines developed in research laboratories.

However, there are a great number of afflictions where existing natural remedies are insufficient, and intervention using computation can be beneficial. Around the same time the central dogma of molecular biology was proposed in the 1950s, computing technology was being born in vacuum tubes. For the next 10 years, molecular biology and computing each advanced in their own spectacular ways, yet applying computing to problems in molecular biology was still a novelty.

By the end of the 1960s, computing had reached a stage mature enough to be applicable to biochemical problems of limited scope, and the first generation of bioinformatics and chemoinformatics was born. Continuing into the next decade, evolutionary trees were one bioinformatics topic (Waterman et al., 1977), and chemoinformatics topics such as the efficient representation of chemicals for searchable databases were explored (Wipke & Dyott, 1974). Computing technology was slowly becoming a useful tool to explore the theoretical underpinnings of the information representing the mechanisms of life.

Both bioinformatics and chemoinformatics have emerged independently in parallel (Jacoby, 2011), much like computing and molecular biology did at first. Their synergy was largely ignored, not for lack of interest, but rather because the computing power necessary to examine and solve large chemical biology problems that impact drug design was still insufficient. (Note the difference between biochemistry, which is biology-centric and focuses on molecule function, versus chemical biology, which focuses on chemical compounds and their biological effects.) Furthermore, from the perspective of pharmaceutical companies, why would they need to consider changing the laboratory techniques which founded their industry in the first place? Fast forward from the 1970s to the present. Over the past decade computing technology has and continues to become cheaper, to the point where it is now possible to equip individual researchers with multi-processor computing workstations that can compute systems applicable to drug design and optimization in a realistic amount of time.

With the recent boost in computing power, the era has come where clinicians, wet-lab scientists, and informaticians can collaborate in inter-disciplinary research for the advancement of drug design to improve the quality of life. Clinicians provide tissue samples of patients exhibiting particular symptoms, bench scientists observe tissue behavior and identify critical molecules of interest, and informaticians provide ways to extract information from the tissue sample that can be useful in retroactively understanding the mechanism which brought about the symptoms. *However, the power of informatics is not only in its retrospective analysis capability.* Depending on the illness or condition under investigation, the objective of a collaborative research project will be for design of a new pharmaceutical, either a retro-active drug (suppress a symptom after it has come about) or a pro-active drug (suppress a symptom before it has come about). Such design is now appreciably impacted by informatics because of the scale of data and precision required for proper understanding of a phenomenon.

In this chapter, we provide several examples of this collaborative drug design process that incorporates informatics for chemical biology and their translational experimental impact. To begin, we review some of the previous methods in which pharmaceuticals have been developed and analyzed *in-silico*. Next, we focus on a recent public database that represents one effort to unify the bioinformatic and chemoinformatic aspects needed for analysis of a major class of proteins that are drug targets. Our attention then shifts to how to mine this information in a useful way and uncover new knowledge that can be and is infact tested at a lab bench. Finally, we provide a glimpse into ongoing research in algorithms that can be incorporated into existing interaction analysis methods, with the end goal of boosting the performance of virtual pharmaceutical design beyond what has been achieved thus far.

2. In-silico development of pharmaceuticals

2.1 Pharmacology basics

In pharmacology, two fundamental molecules in biochemical reactions are **target proteins**, and chemical compounds that attach to proteins, often called **ligands**. When a ligand binds to its target, it triggers a cascade of signals such as a transfer of small groups of atoms or physical modification of receptor structure. The signal transduction affects how a cell, and ultimately a living organism, functions. Therefore, the larger question in modern *in-silico* pharmacological research is how to construct models which accurately correlate experimental observations and protein-ligand information to activity, and therefore provide a way to prospectively (computationally) evaluate the potential efficacy of a newly-designed drug molecule.

2.2 A brief survey of existing virtual screening methods

Virtual screening (VS) is the process of evaluating a library of compounds using a computational model in order to rank, and thus screen for, molecules that exhibit desired characteristics. For pharmaceuticals, the main characteristic is its **bioactivity** or efficacy, which is the amount of a compound needed to trigger a desired physiological effect, typically in at least half of the target. For drugs to be advanced beyond a laboratory experiment stage, bioactivity with micromolar (μM) activity is minimally required, though nanomolar (nM) activity is often a criterion used by pharmaceutical companies to be considered for advanced clinical trials necessary for final product approval before manufacturing and distribution.

The reason why VS is important is simple - the size of chemical space is estimated to be on the order of 10^{60} (Dobson, 2004), and therefore it is economically and logistically unrealistic to perform **assays** (bioactivity tests) for every chemical compound with every protein in a biological system. Estimates for the current development costs of a single drug molecule to reach the market are close to USD $500,000,000. How to explore the enormous chemical space and reduce such development cost is an open topic addressed to some degree in this chapter.

To date, two major classes of VS methods have been created. The first is structure-based virtual screening (SBVS). In SBVS, biophysical and phyisochemical models are applied to estimate binding energies of ligands with a target protein, and the most energy-favorable compounds can be interpreted as the ligands most likely to exhibit bioactivity. A key requirement of SBVS methods is that they require knowledge of a target protein. As long as the three-dimensional structure of the target is known, SBVS can be useful, since molecular shape complementarity and physical properties important for specific binding can be utilized (Schneider & Fechner, 2005). SBVS was a contributor to the development of the first generation of cyclic urea HIV protease inhibitor drugs (Lam et al., 1994). Most SBVS approaches include the use of force fields, a topic that will be discussed later in the chapter.

The second major class of VS methods is ligand-based virtual screening (LBVS). In LBVS, a set of ligands known to bind to a protein is used to construct a model that correlates ligand to characteristics to observable properties. Note that for LBVS methods, knowledge about a target protein is not required. An argument in favor of LBVS is that one can probe and build hypotheses about an uncharacterized cellular system without knowing any of the proteins that the cell line contains. LBVS contributed to ligands that effect human T-cell activation (Schneider & Fechner, 2005).

More thorough surveys of the wealth of SBVS and LBVS methods created to date can be found in the literature (Jacoby, 2011; Schneider & Fechner, 2005; Schneider et al., 2009).

2.3 Problems with heterogeneity

Above, we have given a sample of VS algorithms and their contributions to drug design. However, in most situations, the targets being researched and the target properties used for evaluation of the VS methodology are different from study to study. Therefore, it is often difficult to directly compare the ability of different VS studies. It has been estimated that the ratio of VS approaches to applications is close to one (Schneider, 2010).

The problem of heterogeneity arises from the fact that the goals of past experimental (wet) researches, and consequently the data available for *in-silico* method development, were dependent on the target organism and physiological effect being investigated. For example, one study may seek the IC_{50} concentrations of a chemical library which can further vary based on the assay equipment used, while another study evaluates results using the inhibition constant K_i. The $\mathbf{IC}_p$ metric is the concentration required for p% inhibition of a specific cellular activity. For example, blood clotting drugs for people with hemophilia would be considered more effective (at clotting) as the IC_{50} value (inhibition of bleeding activity) becomes smaller. Being able to reduce IC_{50} values means having the ability for a drug candidate to be equally effective at lower concentrations.

How can heterogeneity issues be overcome? One option is to reduce the range of values possible for a particular metric being used. For a compound library primary screening stage, one may not need the strength of interactions (change in free energy) or the IC_{50} concentration beyond a μM level. Instead, one may simply need a "yes/no"-type of information in order to build a hypothesis. In the blood clotting example, we may only need to build a screening model that predicts if a clotting drug has μM efficacy or better. In the process of drug lead discovery, "yes/no" is often sufficient, with a numerical value for bioactivity or binding affinity to be the criterion in the optimization phase.

Here, we provide a few examples of 2-class interaction data useful for drug design research. The Database of Useful Decoys (DUD) provides over 2000 known ligand-receptor interactions (Huang et al., 2006). For each interacting ligand, over 30 similar but non-interacting "decoy" ligands are also provided. Another public database is the NCI anticancer activity dataset, which logs the bioactivity of thousands of ligands against 60 types of cancer cell lines. For each cell line, each ligand receives a $\{+1, -1\}$ label indicating if it does or does not inhibit the cell line's growth. The NCI-60 dataset is available from the ChemDB (Chen et al., 2005). 2-class labels will be utilized in Sections 3 and 4, and bioactivity values will be the focus in Section 5.

2.4 Machine learning

In-silico development of pharmaceuticals is greatly aided by the use of **machine learning**, an active research field which develops algorithms to extract statistically meaningful information from large datasets. The resulting information models can then be applied to clustering, ranking, or inferfence about unseen data. For those unacquainted with machine learning, it is easy to think of how a human child learns to distinguish colors or shapes, after which they can cluster objects of "similar" color or shape together. The concept of similarity is somewhat of a philosophical argument, and machine learning, much like human learning, can be adjusted through the definition of "similar". For drug lead discovery and optimization, machine learning is the tool that helps us navigate chemical and interaction spaces.

Recently, major contributions to the machine learning field are being achieved through **kernel methods**. Kernel methods can be thought of as being comprised of two separate parts (Shawe-Taylor & Cristianini, 2004): data "recoding" to create patterns representable by linear functions, and efficient linear pattern analysis algorithms applied to the recoded data. Along with good theoretical properties, kernel methods have three special features which account for their recent focus.

First, they can be applied to non-vectorial data that does not have a natural notion of similarity defined, such as chemical graphs or receptor-ligand interactions. Second, for both vectorial and non-vectorial data, the calculation of similarity is equivalent to having explicitly mapped each original data point $x \in X$ into a higher, possibly infinite-dimensional feature space F and using the inner product in F to measure similarity, e.g. a kernel function $K(x, y) = \phi(x) \cdot \phi(y)$, for some transformation $\phi : X \rightarrow F$. This second feature is critical because it represents the "recoded" similarity value *without actually performing the explicit transformation* ϕ. Similarity of feature vectors that grow exponentially or are infinitesimal in length, and hence are difficult or otherwise impossible to compute, can still be analyzed via kernel methods as long as an efficient algorithm to compute the kernel function $K : X \times X \rightarrow \Re$ exists. Third and finally, in light of the previous two reasons, kernel functions can replace the inner product in pattern analysis algorithms. A simple (though uninteresting) kernel function is the basic inner product $K(x, y) = x \cdot y$; more interesting kernel functions, more of their properties, and manipulations on them are abound in the references. The pattern analysis algorithm used in this chapter is the Support Vector Machine (SVM) (Cristianini & Shawe-Taylor, 2000).

Machine learning is now widely used in image analysis including facial, text, and license plate recognition, vital data clustering for clinical applications, and weather and geological condition prediction. In the remainder of this chapter, we will demonstrate how machine learning that includes kernel methods is applied to receptor-ligand analyis, inference of novel protein-ligand binding, and prediction of bioactivity using atomic partial charge information.

3. Bioinformatics and chemoinformatics for GPCR ligand analysis

3.1 GPCR ligands as drug targets

G protein-coupled receptors (**GPCR**s) are a type of transmembrane receptor found in eucaryotes. Their physical structure consists of seven transmembrane helices connected by six extracellular and intracellular loops (EL-{1,2,3}, IL-{1,2,3}). The N-terminus of a GPCR is extracellular, while the C-terminus is intracellular. Once bonded to by peptide or small organic ligands, they activate signal transduction pathways inside a cell, and thus, extracellular ligands which bind to GPCRs affect a cell's internal downstream signaling. GPCR ligands may be broadly classified into **agonists** which increase the amount of signalling that occurs after binding, or **antagonists** which nullify the effect of agonists and return a cell to normal signalling levels.

GPCRs are involved in an amazing number of cellular processes, including vision, smell, mood and behavioral regulation, immune system activity, and automatic nervous system transmission. It suffices to say that loss of function in GPCRs or regain of function by agonistic or antagonistic drugs directly affects the health of an organism. The number of GPCRs in the human genome is more than 1000, with at least 400 of therapeutic interest. In contrast to such a number of potential therapeutical GPCRs, drugs currently available on the market address less than 10% of them (Okuno et al., 2008). For a number of GPCRs, the only ligand known is its endogenous (natural) ligand, and for a considerable number of cases, some GPCRs are *orphaned*, meaning that no ligand is known for which binding occurs. This is the entry point to GPCR *in-silico* research, requiring a unification of bioinformatics (for GPCRs) and chemoinformatics (for ligands). Successful design of agonists and antagonists aided by virtual screening powered through machine learning holds considerable consequence on the future of pharmaceuticals.

3.2 GPCR-ligand data

The GPCR LIgand DAtabase (GLIDA) represents a major effort in using protein and chemical similarity informatics techniques independently as well as synergystically (Okuno et al., 2008). As discussed above, the majority of drugs available on the market address only a small fraction of GPCRs. The amount of GPCR-ligand interaction space explored is still minimal. Therefore, exploration of new regions in the interaction space represents the potential for a number of new GPCR ligands. GLIDA is a database to chart and further navigate such space. GLIDA utilizes proteome data from the human, mouse, and rat genomes. Drug development for humans is an obvious motivation, and mouse and rat genomes are selected because they are frequently used in experimental trials. The interaction data for GLIDA has been assembled from both public and commercial sources, including DrugBank (Wishart et al., 2006), PubMed and PubChem (Wheeler et al., 2007), the Ki Database (Roth et al., 2004), IUPHAR-RD (Foord et al., 2005), and MDL ISIS/Base 2.5.

The total number of GPCRs in GLIDA is roughly 3700, which has remained stable over the past five years. What has increased over the lifetime of the GLIDA database is the number of ligand entries available for analysis and interaction. Since the database's initial public release containing 649 compounds, the ligand database has grown to contain over 24,000 agonists and antagonists. In parallel to the explosion in the number of ligands available, the number of GPCR-ligand interactions catalogued has also swelled from 2000 to 39,000.

In theory, a naive bioassay system could test the activity of all $3700 * 24000 \propto 10^7$ GPCR-ligand pairs. This has at least two disadvantages. First, the exorbitant costs associated with such an approach are prohibitive. Second, machine learning methods which incrementally use data from said theoretical assay will encounter problems with model construction in later phases due to data points which are inconsistent with models of prior generations, and the computational time cost and efficiency of machine learning when inputting $\geq 10^7$ data points is poor. Therefore, the importance of virtual screening is clear.

3.3 Making sense of such quantities of data

With mechanisms in place for storing all of the protein sequences, chemical structures, and interaction data, the major informatics question is how to extract meaningful information from the GLIDA database. GLIDA provides a number of analysis services for mining its information.

First, consider the case when a GPCR has a set of known ligands, and one wishes to search for other similar GPCRs to see if they share the same target ligands. GLIDA provides two types of search services for this scenario. First, protein-protein similarity using primary sequence analysis can be done. For this task, the standard BLAST algorithm is used. The result of such a search is a collection of proteins which exhibit sequence similarity. GLIDA also offers an alternative GPCR search strategy that uses gene expression patterns in tissue origins. In addition to those two services, GPCRs registered in the database are organized hierarchically using the organization scheme from the GPCRDB project (Horn et al., 2003). These bioinformatics tools allow an investigator to look for similar GPCRs with increased efficiency.

Next, consider the second case when a ligand has a target receptor, and one wishes to query (assay) the same receptor for activity by using similar ligands. In this case, KEGG atom type (Hattori et al., 2003) frequency profiles are used to represent molecules, and ligands with similar frequency patterns are returned by a similarity search. The idea that similar frequency patterns results in similar molecules is based on the concept in linear algebra that vectors with minimum distance between them in a space have similar component vectors. For this task, principal component analysis (PCA), a methodology used in bioinformatics as well, is applied. The database search also provides links to external databases such as PubChem and DrugBank, allowing users to investigate similarity in other chemical properties provided by those sources. For example, an investigator may want to evaluate the molecular weight and number of stereocenters of the 10 ligands most similar to the endogenous ligand of a particular GPCR. GLIDA provides the links to each ligand's external information, so the user need only to follow the links provided. Linking multiple chemoinformatics resources makes GLIDA a useful tool for exploring unknown areas in GPCR-ligand interaction space. Readers interested in learning more about clustering of compound libraries can consult additional chapters in this book.

Above, we mentioned the idea of assaying a GPCR for similar ligands. A natural question to ask next is: what about dissimilar ligands? A current and important topic in drug lead design is **scaffold hopping**. Scaffold hopping is the occurrence of a pair of ligands which both exhibit bioactivity for a target protein, but with completely different core structures (scaffolds). The topic of how data in GLIDA can be used to prospectively evaluate dissimilar ligands for a GPCR is addressed later in the chapter.

The next tool that GLIDA provides is a visual GPCR-ligand interaction matrix. This is a graphical version of the GPCR-ligand assay data assembled through bioactivity assay experiments. It allows one to get a quick visual inspection of the areas in interaction space which have been explored and remain to be explored. Such interaction maps are quickly becoming the cornerstone of how to explore interaction spaces in not only GPCRs but also in many other types of protein classes which can be perturbed. At each cell in the interaction matrix, three states are possible for a GPCR-ligand pair: (partial or full) known agonist, (partial or full) known antagonist which in GLIDA includes inverse agonists, or an unknown/non-interacting state.

The interaction matrix unifies each of the bioinformatics and chemoinformatics algorithms that GLIDA employs. For a particular cell in the matrix, neighboring columns are the result of protein similarity calculations. Neighboring rows indicate ligand similarity after applying the ligand clustering algorithm described above. As a result, the algorithm unification and resulting visualization gives drug designers key clues for planning future sets of bioactivity assay experiments and refining drug lead scaffold design.

3.4 Applied examples of GLIDA

The utility of GLIDA can be demonstrated through the following two examples.

In Figure 1, an example interaction matrix is shown. Located between two human alpha adrenoceptors ADA-1A and ADA-1B lies a similar GPCR named Q96RE8. However, the ligation status of Q96RE8 is unknown. Looking at the interaction space of neighbors ADA-1A and ADA-1B, we see that they share a number of common ligands with similar bioactivity. Therefore, performing laboratory assays of Q96RE8 using the known ligands of similar proteins would be expected to discover several novel ligand-receptor pairs. This demonstrates the power of interactome visualization provided by GLIDA.

Second, let us make a small investigation of a GPCR critical to normal human function. The dopamine receptor is a type of GPCR involved in a variety of functions, such as the control of blood pressure and heart rate, certain aspects of visual function, and control of movement (Strange, 2006). With such a wide variety of behaviors, it is of little surprise that the family of dopamine receptors (modernly subdivided into five receptors D_1-D_5) are liganded by many compounds. Using GLIDA, we first look at the ligands L110 (apomorphine) and L084 (7-OH DPAT), and notice that they are both agonists. GLIDA provides structural images of the two compounds as well, shown in Figure 2. Observing the pair, we see that both contain a hydroxyl (-OH) moiety attached to an aromatic ring. They both also contain a nitrogen atom with methyl chains attached that is located near or in the rigid ring system. These common features may be of use in designing new ligands to attach to the dopamine D_1 receptor. In Section 5, we will further discuss atom charges and rigidity of these two ligands.

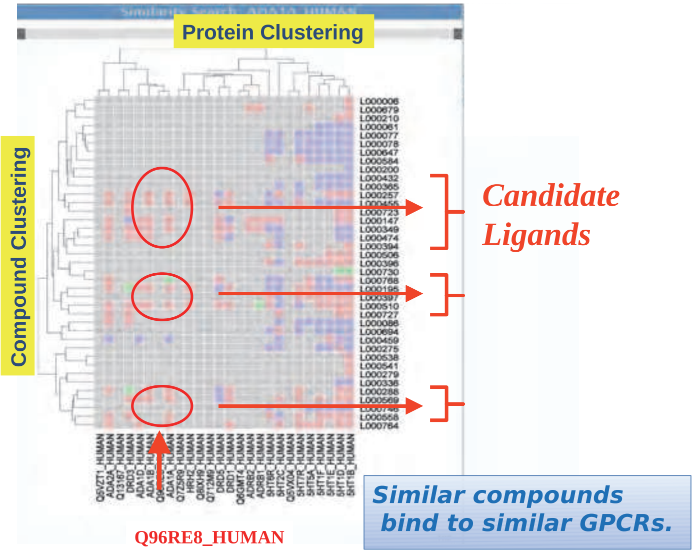


Fig. 1. An example of using the GLIDA database to visually inspect the GPCR-ligand interaction space. The GPCR Q96RE8 is similar to GPCRs ADA-1A and ADA-1B. The interaction matrix suggests that ligands of ADA-1{AB} may also be ligands of Q96RE8.

4. Unified informatics for prospective drug discovery: chemical genomics

4.1 GLIDA in retrospect

The GLIDA database provides a considerable amount (39,000 pairs) of GPCR-ligand interaction data. That data is provided by established research resources, such as the KEGG, IUPHAR, PubChem, and Drugbank databases. The experimental interaction data has been published in peer-reviewed journals.

The next big step in unifying bioinformatics and chemoinformatics for GPCR ligand discovery and pharmaceutical development is how to incorporate the information in GLIDA in a way that can not only analyse data of past trials, but also provide reliable inference of

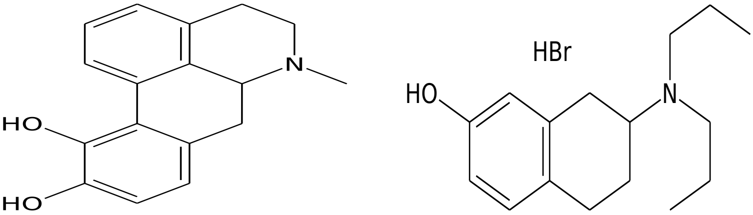


Fig. 2. A pair of agonists for the dopamine D_1 receptor that contain overlapping substructures.

novel GPCR-ligand interaction pairs. We can generalize the situation further by stating the following goal:

> Given a protein-ligand interaction database, build a model which can accurately and compactly express the patterns existing in interacting complexes (protein-ligand molecule pairs), and which also has sufficient predictive performance on experimentally-verifiable novel interactions.

The pharmaceutical motivation behind this next step is the understanding of the mechanisms of **polypharmacology** and **chemical genomics**. Polypharmacology is the phenomenon that a single ligand can interact with multiple target proteins. Chemical genomics is the other direction: that a single protein can interact with multiple ligands. Since polypharmacology is one of the underlying reasons for the side-effects of drugs, and since chemical genomics helps explain the nature of signalling networks, it is critical to advance our understanding of the two mechanisms.

In the early days of molecular biology, it was thought that a single protein could be influenced by a single compound, much like a unique key for a specific lock. Of course, we know that a series of locks can have a single master key (e.g., used by hotel cleaning staff), and that a series of keys can all open a single specific lock (e.g., apartment complex entrance). Polypharmacology and chemical genomics are the replacement of the 1-to-1 protein-ligand concept with the analogous ideas of master keys or a generic lock that can be opened by many keys. Both one ligand binding to multiple receptors (MacDonald et al., 2006) and multiple ligands binding to the same receptor (Eckert & Bajorath, 2007) have been demonstrated experimentally. Also, a quick look at Figure 1 demonstrates polypharmacology in rows and chemical genomics in columns of the interaction matrix. The 1-to-1 concept of binding has had to be replaced with a systems biology approach that considers interaction space as a network of nodes and edges, where nodes are proteins and their ligands, and bonds are drawn between nodes when two molecules interact (bond). What makes informatics special is the ability to incorporate both polypharmacology and chemical genomics.

Recent advances in high-throughput screening have created an enormous amount of interaction data. Facilities such as the NIH's Chemical Genomics Center employ automation technologies that let researchers test thousands of ligands at various concentrations on multiple cell types. This non-linear explosion of interaction information requires new methods for mining of the resulting data. Additionally, to become more economically efficient, it is important to reduce the numbers of ligands being tested at facilities like the CGC to those which are more likely to interact with the target proteins of a specific cell type. This new type of virtual screening has been termed Chemical Genomics-Based Virtual Screening (CGBVS).

4.2 Reasoning behind Chemical Genomics-Based VS

The interaction matrix provided in the GLIDA database is precisely the motivation for development of CGBVS techniques. Earlier in the chapter, some of the merits of LBVS and SBVS were discussed. However, the two methodologies, which have been the principle VS methods of drug lead design research to this point, have their own drawbacks as well, which we discuss here.

Since the LBVS methods use no information about the target protein, the ability to investigate the network of polypharmacology is hampered. For example, let us assume we have ligand dataset L_1 that uses IC_{50} values based on bioactivity assays with receptor R_1, and dataset L_2

uses cell count values measured after ligation to R_2 ($\neq R_1$). LBVS models built using L_1 cannot be applied to screen L_2, because the models have been constructed under the assumption that R_1 was the target. Evaluating L_2 on the basis of the LBVS model contstructed using L_1 is proverbially "comparing apples to oranges". The same argument applies for testing L_1 using a LBVS model built from L_2. One of the other known issues, especially with graph kernel-based LBVS (see Section 5.2), is that both cross-validated training performance and prediction performance on unseen ligands containing scaffold hopping are poor. Graph kernel QSARs (Brown et al., 2010), a type of LBVS methodology, have depended on the frequency of common topology patterns in order to derive their models, which hence rank non-similar scaffolds lower.

The SBVS methods require knowledge of the crystal structure of the target protein, which means that the problem just mentioned for development via LBVS is not a concern. Unfortunately, SBVS has its own set of limitations. First, the force fields used in SBVS techniques are constantly undergoing revision. It has been argued that because force fields and free energy methods are unreliable, they have contributed little to actual drug lead development (Schneider et al., 2009). Second, the amount of free parameters present in force fields and molecular dynamics simulations make them extremely difficult to comprehend and accurately control. We experienced this firsthand when using these methods to investigate HIV protease cyclic urea inhibitor drugs. Third, the amount of computation involved in SBVS methods is enormous, even for small peptide ligands of 5-10 residues. Consequently, SBVS cannot be applied to large libraries such as the 24,000 compounds stored in the GLIDA database. Last but not least, SBVS is completely unapplicable when the target protein crystal structure is unavailable, which is frequently the case when a new cold virus or influenza strain emerges in a population.

Hence, we arrive at the need to create a new generation of informatics algorithms which overcome the difficulties of LBVS and SBVS. In this section of the chapter, we consider the development of a first generation CGBVS-style analysis for exploring target-ligand interaction space. Additionally, the connection between computational chemogenomics (Jacoby, 2011) and real experimental verification is critical for advancement of *in-silico* drug lead design. Also, for the drug lead discovery phase, we wish to search for novel scaffold ligands of a target protein, rather than explore the amount of lead optimization possible. A graphical depiction of the various concepts and direction of research is shown in Figure 3. We will return to the topic of LBVS and its role in drug lead optimization later in the chapter. Though several research projects have investigated new receptors for existing drugs (polypharmacology), the work below is the first to perform the opposite (chemical genomics): discovery of new bioactive scaffold-hopping ligands for existing receptors (Yabuuchi et al., 2011).

4.3 Computational and wet experiments performed

As stated above, an objective of CGBVS is to obtain reliable prediction of ligands previously unknown to bind to target proteins, and experimentally assay those predictions. In Table 1, a complete set of dry and wet experiments performed is summarized.

As Table 1 shows, the connection between theory (dry) and reality (wet) is tested extensively. Critics of informatics argue that its results often do not correlate well with field tests. However, as Table 1 and Section 4.6 show, such a case study answers such criticisms.

Chemo- and bio-informatics are still in their infancies as established fields of study. One obvious observation since their inception is the difficulty for informatics labs wishing to field

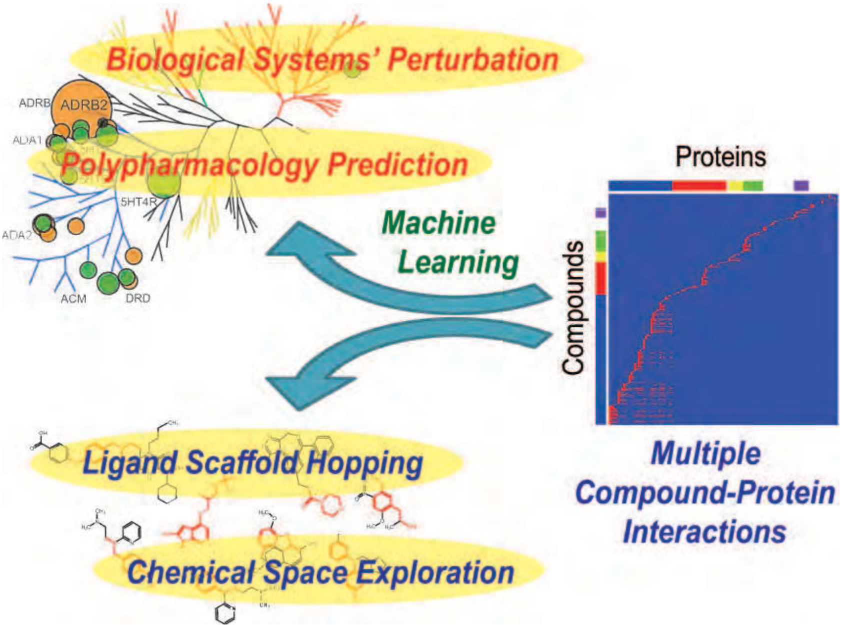

Biological Systems' Perturbation
ADRB
ADRB2
ADA1
Polypharmacology Prediction
5HT4R
ADA2
ACM
DRD
Machine Learning
Proteins
Compounds
Ligand Scaffold Hopping
Chemical Space Exploration
Multiple Compound-Protein Interactions

Fig. 3. The concepts and objectives of Chemical Genomics-Based Virtual Screening (CGBVS).

test their models. It is hoped that studies such as this one will either encourage laboratories to engage in both dry and wet research, or will foster increased collaborative efforts between dry and wet research laboratories.

Type	Test	Purpose	Target	Compounds	Method
Dry	CPI prediction	CGBVS vs. LBVS	all GPCRs	GLIDA	cross-validation
Dry	β2AR binding	CGBVS vs. SBVS	β2AR (GPCR)	GLIDA	hit rate
Wet		CGBVS vs. LBVS/SBVS	β2AR		binding assay
Dry	β2AR binding	non-GPCR ligand test	β2AR	Bionet	prediction score
Wet					cell-based assay
Dry	NPY1R binding		NPY1R		prediction score
Wet					cell-based assay
Dry	CPI prediction	CGBVS vs. LBVS/SBVS	EGFR CDK2	GVK - training DUD - test	prediction accuracy
Wet	EGFR binding CDK2 binding	Kinase inhibitor test	EGFR CDK2	Bionet - test	cell-based assay

Table 1. Summary of dry and wet experiments performed to evaluate the effectiveness of chemical genomics-based virtual screening (CGBVS). CPI: compound-protein interaction

4.4 Informatic analysis - methods

In mining interaction data, there are three pieces of information per interaction sample: the protein sequence, the ligand two-dimensional structure, and whether or not interaction occurs. The CGBVS strategy employs the SVM pattern analysis algorithm with a kernel function that uses explicit feature vectors. In this case study, the strength of interaction is not considered, because the various databases providing the interaction information use different metrics to indicate interaction, as noted in Section 2.3. Therefore, we must restrict the interaction domain (the output value) to a binary value. The remaining input to the SVM is the unification of biological and chemical data.

For proteins in interacting pairs, their sequence data is transformed into a vectorial representation using the well-known mismatch kernel (see Shawe-Taylor & Cristianini (2004)), which we will denote by $\phi_M(P)$ for protein sequence P. The mismatch kernel outputs the frequency of subsequences of fixed length in an input sequence; in particular, the mismatch kernel can be specified to allow a maximum number of mismatches in the subsequence being counted. Details of the parameters of the (2,1)-mismatch kernel used for CGBVS can be found elsewhere (Yabuuchi et al., 2011).

In chemoinformatics, many researches have produced various types of chemical descriptors. Examples are the Extended Connectivity Fingerprint descriptors, the PubChem descriptors, and the DragonX descriptors. Each type of descriptor takes a chemical structure as input and produces a fixed length feature vector that represents characteristics of the molecule. The characteristics may describe topology patterns, connectivity frequencies, electrostatic properties, or other measureable types of information. For CGBVS studies, the DragonX descriptors are employed, which we will denote by $\phi_D(L)$ for ligand L.

If protein-ligand interaction (binding, no binding) is represented by the value $i \in \{B, NB\}$, then each training data interaction in CGBVS can be represented using the feature vector

$$FV(P, L, i) = [\ \phi_M(P), \phi_D(L), i\ ] \quad . \tag{1}$$

Test data simply has one less dimension, since the binding interaction value is unknown.

In experiments, $\phi_M(P)$ is a 400-dimensional vector, and the dimensionality of $\phi_D(L)$ is 929. Therefore, the SVM builds predictive models using a total of 1330 dimensions.

One of the key differences between LBVS methods and CGBVS methods is the absence of receptor information in LBVS. The feature vectors for LBVS simply do not have the $\phi(P)$ element.

4.5 Computational experiments

The first computational test done is comparison of CGBVS to existing LBVS. The details of the LBVS technique used can be found in Yabuuchi et al. (2011). Using the initial release of interaction data in the GLIDA database (317 GPCRs - 866 ligands - 5207 interactions), 5-fold cross-validation was repeated multiple times to assess the average predictive performance of the three techniques. The CGBVS method outperformed LBVS by more than 5%, reaching more than 90% accuracy in predicting GPCR-ligand interactions. Such results indicate the extra performance gained by inclusion of the receptor protein in feature vector information.

Next, using the β2-adrenergic receptor (β2AR), a fairly well characterized GPCR with known crystal structure, retrospective testing of SBVS and CGBVS was done. In this round of testing, β2AR ligands were available in the GLIDA interaction set, so they were eliminated from the training set and used as positive control data. In this test as well, CGBVS provided a higher enrichment rate than SBVS, meaning more of the highest ranking compounds predicted by CGBVS were known β2AR ligands than those ranked by SBVS. As opposed to SBVS, CGBVS considers other protein-ligand interaction, which improves its predictive power.

Having verified CGBVS performance using GLIDA's data in a retrospective way, the next test for validating the usefulness of CGBVS was to vary the ligand dataset while holding the target receptor constant (β2AR). For this purpose, the Bionet chemical library consisting of 11,500 compounds was used. As no type of cross-validation could be performed for this library, an alternative measure of "goodness" was used: the aforementioned ability to scaffold hop. For a number of top-ranking predictions (ligands) experimentally assayed for bioactivity, scaffold hopping was observed. The same process that was used for testing the Bionet library against the β2AR receptor was repeated using the neuropeptide Y-type 1 receptor (NPY1R), with similar successful results.

The next aspect of testing performed was to remove the restriction on the target protein domain. Instead of GPCRs, protein kinases, molecules whose transfer of phosphate groups extensively impact cellular signalling, were used as the target protein. A kinase inhibitor interaction dataset made available by GVK Biosciences was divided into training and test sets, after which CGBVS, LBVS, and SBVS were trained in the same manner as before. Kinase-inhibitor interaction prediction accuracy rates again showed that CGBVS was more effective in mining the interaction space because of its ability to consider multiple interactions as well as its ability to incorporate both bioinformatic and chemoinformatic aspects into its interaction complex representation.

4.6 Laboratory assay experiments

For bioinformatics and chemoinformatics to live up to their promise in drug discovery, it is critical that their predictions be verifiable at the laboratory bench. In this case study for CGBVS, assay experiments were also performed in order to test computational predictions. As the focus of this book is on informatics, details of the bioassays will be very brief.

Among the top 50 β2AR prediction scores, those commercially available and not already identified in the literature as known β2AR ligands were tested in assays. It is also worth noting that of those ligands that were commercially available, some were known only to be ligands for different protein domains. This finding provided further evidence of polypharmacology. Compounds such as granisetron were found to have effective concentration (EC_{50}) values in the mid-μM range.

For testing the Bionet chemical library with β2AR , 30 compounds were assayed. The power of CGBVS and informatics to explore interaction space and contribute novel drug leads was confirmed, as nine of 30 compounds had EC_{50} or IC_{50} values in the nM-μM range. Compared to the hit rates of typical high-throughput screenings where thousands of compounds are assayed, the hit rate of CGBVS is impressive. Finally, using similar assay techniques, novel ligands were also found for the EGFR and CDK2 receptors. The structures of novel ligands and their assay details are published in Yabuuchi et al. (2011).

4.7 Future directions for CGBVS

There are many interesting directions that CGBVS can be continued in.

First, as the amount of ligands available in a chemical library grows, so too does the interaction space. However, the interaction space is so large that blindly inserting all interactions into a machine learning algorithm is inefficient. The development of new techniques to efficiently sample the interaction space while maintaining the ability to discover novel drug leads prospectively is a very open topic.

Second, CGBVS has shown that it is successful in identifying ligands that exhibit scaffold hopping. It therefore reasons that CGBVS can be embedded as part of a ligand generation algorithm. Development of ligand generation using the Particle Swarm Optimization class of algorithms is another area of ongoing research (Schneider et al., 2009).

Third, one of the largest hurdles in evaluating protein-ligand interaction prediction techniques is the availability of non-interacting data. Most scientific journal articles publish results indicating successful ligand design and interaction strength. However, for the advancement of *in-silico* screening techniques, *the public availability of datasets of non-interacting pairs is equally important*. In the CGBVS study, non-interacting pairs were defined as randomly selected protein-ligand pairs not existing in an interaction database. It is easily possible that such data contains a fraction of false negatives.

5. Boosting CGBVS through improved LBVS methods

5.1 Why return to LBVS?

Earlier in the chapter, we showed how CGBVS outperformed LBVS in terms of target-ligand interaction prediction performance. Even futher, the predicted interactions were tested in wet laboratory experiments, and results showed that CGBVS was superior in prospective interaction studies.

In this final section, we discuss new techniques for optimizing prediction of ligand properties, such as binding affinity or bioactivity. The techniques fall under the framework of LBVS. It may seem contradictory that, despite showing the superior performance of CGBVS over LBVS, we return the discussion to recent advancements in LBVS methods. However, the motivation for pushing the state of the art in LBVS is at least four-fold:

- There are many cell lines which are completely uncharacterized, and therefore, no information about receptors and other proteins exists. In this situation, no specific protein-ligand interaction information is available, but it is still possible to observe and catalog perturbations to cell lines through the supply of various ligands, and build predictive models for screening the next generation of ligands to test on the cell line. For example, such perturbation modeling is useful for deciding on a subsequent selection of chemical libraries to apply to chemical genomics assays.
- Even in cases where a few target proteins and resulting interactions are known, it may be an insufficient amount of interaction data to build effective predictors. For example, one amine receptor and one peptide receptor are hardly enough to characterize the entire interactome of mice or rats.
- As the CGBVS method used a combination of a protein sequence kernel subsystem and a chemical descriptor feature vector subsystem, any improvement in the chemical similarity subsystem can contribute to enhanced CGBVS performance.
- An important distinction exists between the roles of CGBVS and LBVS. The CGBVS process is responsible for the drug lead screening and discovery process. Once the set of potential drug molecules to search through has been reduced via CGBVS to the neighborhood of a

newly discovered drug lead and a target protein, the optimization process can be handed off to more focused LBVS methodologies.

Given these contexts, it is therefore worth continuing the investigation into new LBVS methods.

5.2 Graph kernels for target property prediction

Kernel methods feature the convenient property that kernel functions can be designed as compositions of other kernel functions. Therefore, the CGBVS method can also use any other chemical kernel function, in combination with or as a replacement for the DragonX descriptors used in the GPCR, polypharmacology, and chemical genomics studies.

Most LBVS approaches are used to describe Quantative Structure-Activity/Property Relationships (QSAR/QSPR), which attempt to correlate quantities of ligand structural features to properties, typically agonistic bioactivity. In recent years, chemical property prediction via graph topology analysis, a type of QSAR, has received attention. Graph kernel functions (hereafter "graph kernels") transform a molecule's atoms and bonds into respective graph vertices and edges. Initial researches into graph kernels created a series of random walks on a chemical graph, in order to explore the topology space of input molecules. The resulting path frequencies were then used to assess the similarity of chemical structures.

The second generation of graph kernels expanded the type of subgraph being used for kernel function calculation. Rather than walks on chemical graphs, Mahé and Vert expanded the subgraph space to use subtrees existing in compound substructures (Mahé & Vert, 2009). For subtree space $T = \{t_1, t_2, \ldots\}$, a weight $w(t)$ that evaluates the amount of subtree branching and size complexity is assigned to each tree t, and the function $\psi_t(G)$ counts the frequency of occurrence of tree-pattern t in a molecular graph. The original subtree graph kernel for two molecules M_1 and M_2 is:

$$K(M_1, M_2) = \sum_{\alpha_1, \alpha_2 \in M_1, M_2} K_t(t_{\alpha_1}, t_{\alpha_2}) \quad , \tag{2}$$

where α_i is an atom in a molecule, and t_{α_i} is the subtree rooted at α_i. Based on the idea of convolution kernels, the existing graph kernels have generally been designed for chemical structures by computing a kernel for trees or paths. In other words, the graph kernels are defined by incorporating a more fundamental kernel (similarity) $K_S(s_1, s_2)$ between substructures $s_1, s_2 \in S$ existing inside of graphs. Removing coefficients and constants, and labelling the chemical graph of molecule M_i as G_i, the graph kernels are essentially $K(G_1, G_2) = \sum_{s_1, s_2 \in G_1, G_2} K_S(s_1, s_2)$. This (K_s) is precisely the meaning of K_t in the definition above. The subtrees were found to improve the performance when predicting ligand anti-cancer activity in multiple types of cell lines using the NCI-60 dataset described in Section 2.3.

A recent third generation of graph kernels next addressed the unresolved problem of chirality (Brown et al., 2010). The constraints for matching subtrees were extended to enforce matching atom and bond stereo configurations, meaning the calculation of $K_t(t_1, t_2)$ was altered to check for stereochemistry, and performance evaluation using a set of human vitamin D receptor (hVDR) ligands with a large number of stereoisomers demonstrated a clear performance boost in predicting bioactivity. hVDR ligands are being considered as therapeutic drugs because of their ability to induce differentiation in leukemia cells and additional ability to suppress transcription in cells of cancerous tumors.

5.3 State of the art: pushing graph kernels further

Here, we describe a new generation of graph kernels that are being actively researched by the authors. One of the key drawbacks with existing graph kernels being applicable to large scale drug optimization is that the graph information alone contains no electrostatic information which is essential for optimizing protein-ligand binding. Molecular dynamics simulations and docking programs often make large use of electrostatics in order to estimate binding free energies and other properties; however, we have stated above that such programs are unfeasible for scanning large chemical libraries. Another problem with existing graph kernel QSAR methods is their inability to extract patterns from datasets that contain large amounts of scaffold hopping. Therefore, a goal in the next generation of graph-based molecule kernel methods is to incorporate electrostatic information in a two-dimensional graph kernel in such a way that it can better describe the protein-ligand binding space and more accurately correlate ligands to their properties.

5.4 The partial charge kernel

5.4.1 Motivations

The partial charge kernel is built using the following motivations:

- Substructures such as esters (-C(C=O)OC-) and thioesters (-C(C=O)SC-) contain the exact same topology. Therefore, even if graph mismatching similar to the sequence mismatch kernel used in Section 4 were introduced into existing graph kernel methods in order to maximize subtree overlap, the distribution of atomic charge would still be different in the two molecules. A mechanism for capturing this difference in atom charges is important. For example, consider the figures of GPCR ligands clozapine and chlorpromazine in Figure 4. They contain structural similarity, but their charge distribution is considerably different. This difference in information is imporant for the effectiveness of machine learning algorithms.
- The rigidity of molecules or their substructures directly impact their binding affinity for a particular receptor. The existing graph kernels do not take rigidity into account. For example, the structures of apomorphine and 7OH-DPAT shown in Figure 2 are largely rigid structures, but there is a critical difference in the flexibility of the methylenes (-CH_2-) and methyls (-CH_3) attached to the nitrogen atoms. Similarly, the 1,4-methyl-dinitrocyclohexane ring (-NCCN(C)CC) in clozapine (Figure 4) is more rigid than the antenna (-CCCN(C)(C)) of chlorpromazine.
- Stereochemistry plays a critical role in the efficacy of drug molecules (Lam et al., 1994). As with the chiral graph kernels recently developed, stereochemistry must be addressed. Therefore, the partial charge kernel also contains a stereochemistry factor.
- Without sufficient path lengths in graph kernels, identical substituents in remote parts of molecules anchored off of scaffolds are ignored. In previous studies, the walk and subtree path lengths considered were typically no more than 6 bonds (7 atoms). However, this is insufficient for molecules such as steroids typically composed of joined rings, in which a path length of 6 bonds cannot "see" multiple distant substituents accurately. This is

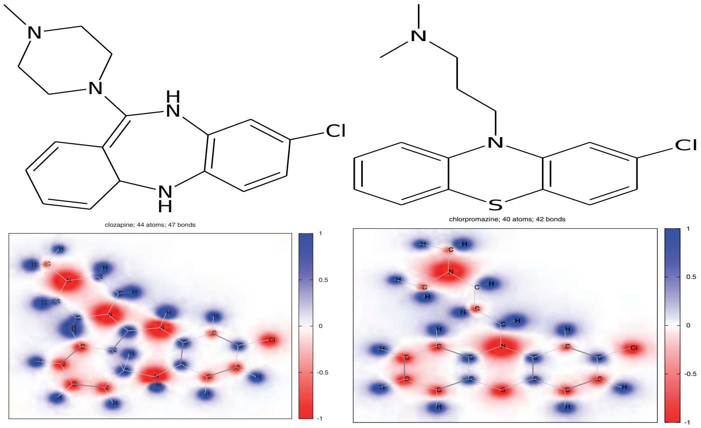

N
N
H
N
Cl
N
H
N
N
Cl
S
clozapine; 44 atoms; 47 bonds
chlorpromazine; 40 atoms; 42 bonds
1
0.5
0
-0.5
-1
1
0.5
0
-0.5
-1

Fig. 4. The dopamine D_1 and D_5 ligands clozapine (left) and chlorpromazine (right), shown with their structures on top and a distribution of their atomic charges on bottom. The ligands are agonists for the D_5 receptor, but antagonists for the D_1 receptor.

also the case when considering a path of length six from the chloride atoms in Figure 4. Therefore, a molecule partitioning algorithm is formulated into the charge kernels.

5.4.2 Design concepts

We will abbreviate many of the mathematical details of the partial charge kernels being actively invesigated, and will instead provide descriptions of the concepts that they are meant to address.

First, since the tree kernels were computationally difficult to apply for tree depths greater than 6 or 7, an initial idea is to apply maximum common substructure (MCS) algorithms to look at the global maximum overlap and consider the atom-pairwise difference in charges over the two molecules. This strategy suffers from the idea that core structure substituents in datasets are highly varied, and the MCS will therefore erroneously discard these portions of molecules. Therefore, let us define a componentization function $CF(M)$ that divides a molecule. An example of a well-known componentization function is the RECAP rule set for retrosynthetic analysis (Lewell et al., 1998). For experiments in this chapter, CF breaks a molecule into its ring systems and remaining components. Components consisting of a single hydrogen atom are eliminated. Label the resulting set of components $CF(M) = C = \{c_1, c_2, ...c_n\}$.

The partial charge kernel's general idea is to evaluate the electrostatic differences in molecule components. For input molecules M_1 and M_2, the similarity of components is summed:

$$K_{SUMCOMP}(M_1, M_2) = \sum_{c_1 \in CF(M_1)} \sum_{c_2 \in CF(M_2)} K_{MOL}(c_1, c_2) \tag{3}$$

Next, we proceed with the design of the molecule component similarity function $K_{MOL}(c_1, c_2)$. The MCS is employed here in two ways. First, it provides the mapping of atoms in one component to another, such that their difference in atomic charge can be evaluated. Second, the ratio of the MCS and molecule sizes provides a quick measure of the significance of the overlap computed. This ratio has a free parameter attached that lets one control how quickly to diminish the similarity of molecules based on the difference in their size and MCS. Such a ratio is a modifying function, $mf_s(c_1, c_2)$, of a more fundamental similarity calculation using the atom charges.

$$mf_s(c_1, c_2) = \Big(\frac{2 * MCS_A(c_1, c_2)}{N_A(c_1) + N_A(c_2)} \Big)^{\phi_s} \tag{4}$$

Finally, the molecule components and their charge distributions will be the information used to build a fundamental kernel $K_{FK}(c_1, c_2)$. The partial charge kernel is designed using convolution through its fundamental kernel $K_{FK}(c_1, c_2)$, much like the predecessor tree kernels formulated in equation (2). Though only the size scaling modifier function has been presented, any number of modifiers providing a multiplicative or additive effect could be attached to the fundamental kernel. The molecule kernel is thus defined as

$$K_{MOL}(c_1, c_2) = mf_1 \circ mf_2 \circ \ldots \circ mf_m(K_{FK}(c_1, c_2)) \quad . \tag{5}$$

We will abbreviate the further details of $K_{FK}(c_1, c_2)$ necessary to incorporate the various motivations given above. In experimental results below, componentization, stereochemistry, molecule size ratio, and rigidity are all formulated into K_{MOL}.

5.5 Computational experiment performance

In computational experiments, we evaluate the ability of the partial charge kernels to calculate the bioactivity of ligands of three different receptors in two different organisms. The first type, ecdysteroids, are ligands that are necessary for shedding in arthropods. A set of 108 ligands containing 11 stereoisomer groups was used. The second type of data used is human vitamin D receptor ligands, whose benefits have been discussed above. Including the endogenous hVDR ligand, a total of 69 ligands containing 18 stereoisomer groups were used. Finally, a well known dataset of 31 human steroids which bind to corticosteroid binding globulin (CBG) is evaluated. The dataset contains two pairs of stereoisomers. More details about all three datasets can be found in Brown et al. (2010).

For each dataset, the training and testing dataset are randomly divided using a 70%/30% split, and this randomized split process is repeated five times. First, internal cross-validation tests are done on the training set. Then, using the entire training dataset (of a particular split), a predictive model is built, and bioactivity is predicted for each ligand in the split's test dataset. To independently evaluate the training set cross-validation and test set prediction performances, two correlation metrics are used. The training dataset uses the q^2 metric:

$$q^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad , \tag{6}$$

where y_i is sample (compound) i's known experimental value (activity level or target property), $\hat{y}_i$ is its value output by a predictor during cross-validation, and $\bar{y}$ is the known experimental average value. The test dataset uses the R metric:

$$R = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}} \quad , \tag{7}$$

where $\bar{y}$ is the average of the predicted values. The maximum values of each of the two correlation metrics are 1. The correlation metrics can take on negative values as well, which suggests that a model has poor predictive ability.

To evaluate the partial charge kernels, we consider four criteria:

- $q^2 \geq 0, R^2 \geq 0$
- $q^2 \geq 0.5, R^2 \geq 0$
- $q^2 \geq 0, R^2 \geq 0.6 \quad (R \geq 0.774)$
- $q^2 \geq 0.5, R^2 \geq 0.6$

The first criterion is a very simple measure to ensure that a model has some reasonable ability to correlate molecules to their bioactivities. The second and third criteria are more strict measures that have been recommended in order for a QSAR model to be applicable to drug development at an industrial scale. The fourth criterion enforces both the second and third criteria.

Results of partial charge kernel bioactivity prediction experiments on the human CGB steroids are highly impressive. Several thousand models satisfied the second requirement of training data cross-validation performance, and a number of those had R values over 0.85, satisfying the fourth set of requirements as well. Though the performance is impressive, optimistic caution must be exercised because the amount of data available is rather small compared to other datasets. Results on the ecdysteroid dataset, over three times as large as the CGB steroid dataset, demonstrate the point. Experiments from the ecdysteroid dataset (using random train-test splits) produce many models with performance that satisfy both the second and third requirements, but the number of models which satisfy both requirements is limited. Still, the prediction performances obtained are better than the graph kernels previously reported. The use of atomic charge information and localized analysis (via componentization functions) in the kernel function results in prediction improvement. Finally, experiments done using the hVDR ligand dataset, which contains a rigid core structure, show that accounting for differences in partial charges and rigidity in components is important for *in-silico* drug optimization. For three different hVDR dataset train-test splits tested, partial charge kernel QSARs built achieve $q^2 \geq 0.7$ performance. Some of those QSAR models come close to meeting the fourth criterion, such as a QSAR we could derive with performance $(q^2 = 0.69, R = 0.745)$. This is considerably better performance than the chiral graph kernels we previously developed, which achieved predictions in the range of $(q^2 = 0.5, R = 0.6)$.

5.6 Ongoing developments in partial charge kernels

The partial charge kernels have shown improvement in prediction performance over the basic graph kernels. A number of designs and tests are being planned to bolster prediction performance.

First is the idea of polarity distribution. If the variance of average component charge is large, then there are more likely to be multiple sites in the ligand responsible for its activity and polypharmacology. The distance between components and their average charge must be correlated somehow. Second, a hybrid of the graph kernels and partial charge kernel has been proposed. In this kernel function scheme, the graph kernels are employed as in their original design (Brown et al., 2010), but instead of using a 0/1 value when calculating the kernel function for a pair of atoms, their difference in charge is used. Finally, as most of the individual design concepts used in the partial charge kernel each contain one or two parameters, investigation of optimal parameter sets which form the pareto optimal is important in order to bound the number of parameter sets applied to new datasets for predictive model construction.

In terms of datasets, the hVDR ligands and ecdysteroids provide a nice starting point of investigating a single particular receptor. By using the known ligands for each GPCR in the GLIDA database, we can construct a chemical genomics-type of predictive model which could be applied for screening molecules with optimum bioactivity. Though the hVDR and ecdysteroid ligands contain a wide variety of bioactivities and structures, the number of compounds available is relatively small compared to some other databases. In this respect, it is important to validate the partial charge kernel's ability to show similarly good performance on larger data sets.

6. Conclusion

In this chapter, we have considered a number of issues and developments centered around *in-silico* design of drug molecules. We demonstrated how the unification of bioinformatics and chemoinformatics can produce a synergistic effect necessary for the mining of protein-ligand interaction space. Yet, development of algorithms in each of bioinformatics and chemoinformatics must continue in order to address life science informatics problems of larger scale. Chemoinformatic algorithm advancement through the partial charge kernels, in planning for incorporation into the CGBVS framework demonstrated, is an example of such algorithm advancement. We hope that the survey provided here has provided stimulation to the reader to investigate and contribute to the complex yet extremely exciting field of *in-silico* drug design.

7. References

Brown, J., Urata, T., Tamura, T., Arai, M., Kawabata, T. & Akutsu, T. (2010). Compound analysis via graph kernels incorporating chirality, *J. Bioinform. Comp. Bio.* S1: 63–81.

Chen, J., Swamidass, S., Dou, Y., Bruand, J. & Baldi, P. (2005). Chemdb: a public database of small molecules and related chemoinformatics resources., *Bioinformatics* 21: 4133âĂŞ4139.

Cristianini, N. & Shawe-Taylor, J. (2000). *Support Vector Machines and other kernel-based learning methods*, Cambridge University Press: Cambridge, U.K.

Dobson, C. (2004). Chemical space and biology, *Nature* 432: 824–828.

Eckert, H. & Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches., *Drug. Discov. Today* 12: 225–233.

Foord, S., Bonner, T., Neubin, R., Rosser, E., Pin, J., Davenport, A., Spedding, M. & Harmar, A. (2005). International union of pharmacology. xlvi. g protein-coupled receptor list, *Pharmacol. Rev.* 57: 279–288.

Hattori, M., Okuno, Y., Goto, S. & Kanehisa, M. (2003). Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways, *J. Am. Chem. Soc.* 125: 11853–11865.

Horn, F., Bettler, E., Oliveira, L., Campagne, F., Cohen, F. & Vriend, G. (2003). Gpcrdb information system for g protein-coupled receptors, *Nuc. Acids Res.* 31: 294–297.

Huang, N., Shoichet, B. & Irwin, J. (2006). Benchmarking sets for molecular docking, *J. Med. Chem.* 49: 6789âĂŞ6801.

Jacoby, E. (2011). Computational chemogenomics, *Comput. Mol. Sci.* 1: 57–67.

Lam, P., Jadhav, P., Eyermann, C., Hodge, C., Ru, Y., Bacheler, L., Meek, J., Otto, M., Rayner, M., Wong, Y., Chong-Hwan, C., Weber, P., Jackson, D., Sharpe, T. & Erickson-Viitanen, S. (1994). Rational design of potent, bioavailable, nonpeptide cyclic ureas as hiv protease inhibitors, *Science* 263: 380–384.

Lewell, X., Judd, D., Watson, S. & Hann, M. (1998). Recap–retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry, *J. Chem. Inf. Comput. Sci.* 38: 511–522.

MacDonald, M., Lamerdin, J., Owens, S., Keon, B., Bilter, G., Shang, Z., Huang, Z., Yu, H., Dias, J., Minami, T., Michnick, S. & Westwick, J. (2006). Identifying off-target effects and hidden phenotypes of drugs in human cells., *Nat. Chem. Bio.* 2: 329–337.

Mahé, P. & Vert, J. (2009). Graph kernels based on tree patterns for molecules, *Mach. Learn.* 75: 3–35.

Okuno, Y., Tamon, A., Yauuchi, H., Niijima, S., Minowa, Y., Tonomura, K., Kunimoto, R. & Feng, C. (2008). Glida: Gpcr-ligand database for chemical genomics drug discovery - database and tools update, *Nuc. Acids Res.* 36: D907–912.

Roth, B., Lopez, E., Beischel, S., Westkaemper, R. & Evans, J. (2004). Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for cns drug discovery, *Pharmacol. Ther.* 102: 99–110.

Schneider, G. (2010). Virtual screening: an endless staircase?, *Nat. Rev. Drug Disc.* 9: 273–276.

Schneider, G. & Fechner, U. (2005). Computer-based *de novo* design of drug-like molecules, *Nat. Rev. Drug Disc.* 4: 649–663.

Schneider, G., Hartenfeller, M., Reutlinger, M., Y., T., Proschak, E. & Schneider, P. (2009). Voyages to the (un)known: adaptive design of bioactive compounds, *Trends in Biotech.* 27: 18–26.

Shawe-Taylor, J. & Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*, Cambridge University Press.

Strange, P. (2006). *Introduction to the principles of drug design and action*, 4 edn, Taylor and Francis Group: Boca Raton, Florida, chapter Neurotransmitters, Agonists, and Antagonists, pp. 523–556.

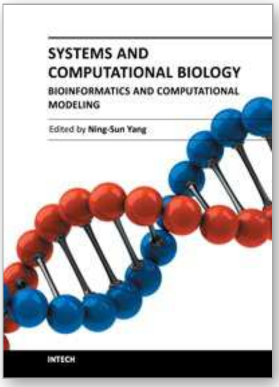
Waterman, M., Smith, T., Singh, M. & Beyer, W. (1977). Additive evolutionary trees, *J. Theor. Biol.* 64: 199–213.

Wheeler, D., Barrett, T., Benson, D., Bryant, S., Canese, K., Chetvernin, V., Church, D., DiCuccio, M., Edgar, R., Federhen, S., Geer, L., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D., Madden, T., Maglott, D., Ostell, J., Miller, V., Pruitt, K., Schuler, G., Sequeira, E., Sherry, S., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R., Tatusova, T., Wagner, L. & Yaschenko, E. (2007). Database resources of the national center for biotechnology information, *Nuc. Acids Res.* 35: D5–D12.

Wipke, W. & Dyott, T. (1974). Simulation and evaluation of chemical synthesis. computer representation and manipulation of stereochemistry, *J. Am. Chem. Soc.* 96: 4825–4834.

Wishart, D., Knox, C., Guo, A., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. & Woolsey, J. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration, *Nuc. Acids Res.* 34: D668–D672.

Yabuuchi, H., Niijima, S., Takematsu, H., Ida, T., Hirokawa, T., Hara, T., Ogawa, T., Minowa, Y., Tsujimoto, G. & Okuno, Y. (2011). Analysis of multiple compound-protein interactions reveals novel bioactive molecules., *Molecular Systems Biology* 7: 472.



Systems and Computational Biology - Bioinformatics and Computational Modeling
Edited by Prof. Ning-Sun Yang

ISBN 978-953-307-875-5
Hard cover, 334 pages
Publisher InTech
Published online 12, September, 2011
Published in print edition September, 2011

Whereas some â€œmicroarrayâ€ or â€œbioinformaticsâ€ scientists among us may have been criticized as doing â€œcataloging researchâ€, the majority of us believe that we are sincerely exploring new scientific and technological systems to benefit human health, human food and animal feed production, and environmental protections. Indeed, we are humbled by the complexity, extent and beauty of cross-talks in various biological systems; on the other hand, we are becoming more educated and are able to start addressing honestly and skillfully the various important issues concerning translational medicine, global agriculture, and the environment. The two volumes of this book present a series of high-quality research or review articles in a timely fashion to this emerging research field of our scientific community.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

J.B. Brown and Yasushi Okuno (2011). Unifying Bioinformatics and Chemoinformatics for Drug Design, Systems and Computational Biology - Bioinformatics and Computational Modeling, Prof. Ning-Sun Yang (Ed.), ISBN: 978-953-307-875-5, InTech, Available from: http://www.intechopen.com/books/systems-and-computational-biology-bioinformatics-and-computational-modeling/unifying-bioinformatics-and-chemoinformatics-for-drug-design

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821