Computational multitarget drug discovery

Jeremy A. Horst[1,2,3], Adrian Laurenzi[1], Brady Bernard[1,4], Ram Samudrala[1,2*]

[1] Department of Microbiology, School of Medicine, University of Washington
[2] Department of Oral Biology, School of Dentistry, University of Washington
[3] Department of Orofacial Sciences, University of California at San Francisco
[4] Institute for Systems Biology
*corresponding author: ram@compbio.washington.edu

1. Introduction

Pharmaceutical substances have been discovered by means ranging from serendipitous observation (Fleming, 1929; Fleming et al., 1950) to specific engineering (Schneider and Fechner, 2005). The purpose is nearly always to combat one particular disease, and the approach is most often trial and error. The efficiency of these pharmaceutical hunts has been improved greatly by high throughput pharma platforms, but the requirement of physical experiment makes these screens scale in expense linearly at best. The expense of discovering a new chemical entity is estimated at US$0.5B to US$2B (DiMasi et al., 2003; Adams and Brantner, 2006).

Recent successes in computational modeling of compound to protein docking open the possibility of nonphysical prelaboratory screens. In our experience this has vastly increased the success rate of bench experiments (Jenwitheesuk et al., 2008; Costin et al., 2010; Table 1). Computational modeling of protein ligand interactions has been applied to find pharmacologic targets in known drug-disease pairs (Jenwitheesuk and Samudrala, 2007; Keiser et al., 2009). The more obvious use of these docking methods is to guide discovery of a drug for a disease, as modeling enables design (Schneider and Fechner, 2005). Design does not need to be limited to one protein target. Searching for one compound for multiple targets in the same pathogen increases odds for successful inhibition of at least one target, and facilitates discovery of multitarget lead inhibitors [Note 1], which vastly decreases the probability of developing resistance (or habituation) and decreases toxicity via lowered effective dose (Rogawski, 2000; Nezami et al., 2003; Csermely et al., 2005; Jenwitheesuk et al., 2008; Table 1).

Thus far the search for multitarget inhibitors has focused on one organism at a time (Nezami et al., 2003; Jenwitheesuk et al., 2008; Keiser et al., 2009), but modeling multidisease effects has explained clinical patterns of elimination for two diseases by one drug (Samudrala and Jenwitheesuk, 2007). The advent of computational multidisease screens will enable access to the most accurate aspects of computational screening, bearing the possibility of vastly reducing barriers to drug development.

In this chapter we elaborate the conceptual framework underlying rational drug discovery, describe contemporary computational approaches, discuss emerging concepts, and introduce a pipeline to integrate the array of promising techniques and ideas which are already transforming drug discovery.

2. The pharmacologic hunt of yesteryear

2.a. Ethnopharmacy

Since before written history humans have sought available substances (mineral, animal, plant) to cure specific ailments. The hundreds of medicinal substances catalogued in the materia medica of various cultures before and during the time of Socrates (DeVos, 2010; Campbell et al., 2005; Manniche, 1989) demonstrates that the hunt for pharmacologic activity may predate the technology of the scientific method itself. Whether disproving hypotheses or embarking on fishing expeditions, experiences with curative and toxic substances may have conceptually secured the intuitive approach of trial and error investigation.

For thousands of years humans have applied trial and error experiments, separating out extracts of active agents to increase potency and remove unwanted properties. The earliest records describing pharmacologic safety include descriptions of animal models and progressive increases in dosage to test safety and efficacy (Huff, 2003). Nonetheless, technological improvements were limited to purifications and altering the design of the trial itself.

For two hundred years we have isolated specific pharmacologically active molecules (Hamilton and Baskett, 2000). For a century we have knowingly modified the chemical structure of natural compounds to tune desirable and undesirable effects. These attempts of drug discovery and design have led to one specific molecule at a time to combat microbial infection (Ehrlich, 1910; Lloyd et al., 2005) and noninfectious diseases (Strebhardt and Ullrich, 2008).

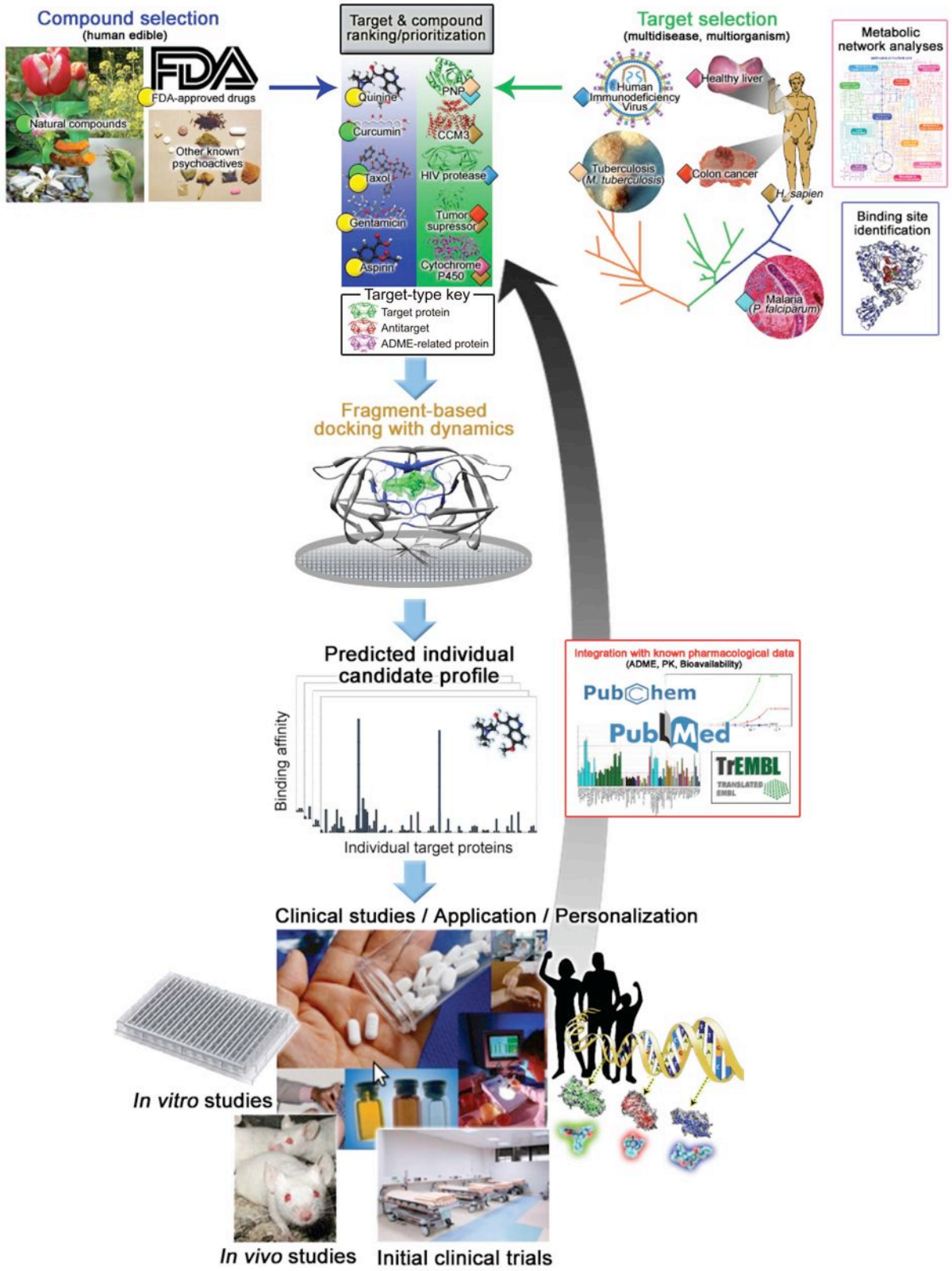

Figure 1. Computational multidisease multitarget screening pipeline. The schematic view of our computational multidisease multitarget screening pipeline relates emerging concepts and techniques described in this chapter, which are already transforming drug discovery. The contemporary weaknesses of computational modeling can be overcome to find pharmacologically active substances by careful selection of the protein and compound sets to be used in computational screening (shown on the sides at top). To maximize the chance of bioactivity and safety in humans, compounds to be considered for screening (top left) should be selected from existing drugs (section 6.c) or natural compounds (section 6.a). The selection of protein targets (upper right) that can be exploited to stop a disease is a nontrivial problem requiring extensive analysis (sections 2.b, 3.a). The probability of finding pharmacologically active compounds is heightened by targeting multiple proteins relating to a disease (section 6.j), which can be in the same signaling network (network targeting, section 5.d), or in different disease associated pathogens (section 6.k); screening against antitarget host proteins can also be performed to control off-target effects (section 6.d). The protein structure (section 5.c)

and binding sites (section 5.a) can be predicted using knowledge based methods (section 4.a). Next, target proteins are prioritized based on the susceptibility of the binding site (sections 3.b, 3.c), the accessibility of the subcellular location, and the similarity of physiologic substrates to the compound set (section 4.f). The set of potential pharmacologic compounds are then prioritized (top center) based on features of the target protein and disease site (sections 3a, 3b) and similarity to target substrates (sections 2.d, 4.f). Finally, the compounds are computationally docked to the active sites of the target proteins (upper middle; sections 4.b, 4.e) with small bursts of molecular dynamics (section 5.b), scored (section 4.c), and ranked with respect to each other (section 4.d). Initially a large compound set is evaluated, with subsequent cycling between directed fragment based optimization, and cycling back to evaluate many similar compounds, which mimics the bench process for discovery of a new chemical entity (section 2.e). The profiles of predicted binding affinities for each compound are compared to titrate selectivity and minimize untoward side effects (lower middle; section 6.j). The use of compounds of known human safety profiles comes to fruition when approaching validation (section 6.a): for diseases with no sufficient model system and no existing cure, existing pharmacologic agents may progress directly to initial clinical trials (center bottom; section 6.c). As well, the multitarget approach of using compounds which are predicted to be active against multiple pathogen proteins increases the odds of success: if a compound is predicted to inhibit six proteins, there is a good chance that it will actually inhibit at least one (section 6.j). As an extension, computational screens of targets for multiple diseases increases the odds of finding a target for the inhibitor; allowing the discovery process to drive disease selection enables access to the most accurate computational predictions (section 6.k). There are initial indications that computational simulations can be more accurate than high throughput screens, possibly because they model bioactivity in an explicitly physiological manner whereas the implicit physical interaction model of bench screens are susceptible to nonspecific aggregation, covalent bonding, and promiscuous binding (section 6.f). Meanwhile, sophisticated bench analysis techniques offer the pinnacle of accuracy, particularly the dissection of enthalpic and entropic contributions to the free energy of binding by isothermal titration calorimetry (section 6.e). Protein, whole pathogen, whole animal, and clinical analysis (center bottom) feeds back to improve the accuracy of simulations (large arrow) by integration with existing pharmacologic data (middle right). Modeling the impact of genetic variance on protein structure allows design of generalized inhibitors for rapidly mutating pathogens and cancers, and specification to individual human differences to control side effects (bottom right; section 6.h). Our group and others have demonstrated the early maturity of computational modeling of protein-ligand interactions by predicting compounds for desired pharmacologic activity and testing them in prospective experiments. A philosophy of freely available open source software has been embraced by many publicly funded groups (section 6.i). These methods not only save time and resources but are beginning to be more accurate than in vitro screening methods (section 6.f). The combination of computational multitarget drug discovery and stringent bench experimentation will lead a new era of effective selective drugs.

2.b. Protein targets

With the advent of molecular biology we found the key to rational drug discovery: inhibiting specific protein "targets" essential to the progression of the disease causing agent. Targets are carefully identified by the consensus of extensive experimentation verified by multiple independent research groups. Thus the major goal of pharmacologic development has emerged as discovering or designing compounds that demonstrate favorable therapeutic activity towards a specific protein target..

Under the current paradigm, an attractive target is a protein essential to the infection, onset, or replication of the disease causing agent, or a protein able to control one of these processes. The protein target should be different enough from homeostatic host proteins that a drug which inhibits its action would not kill the host. A target should be essential to the metabolism, growth, or reproduction of a pathogen or the progression of a neoplasm, and maximally different from all other antitarget human proteins.

2.c. Hitting the target

Tests of pharmacologic efficacy have been refined from observing the signs and symptoms of a disease, to growth of the disease causing agent (e.g. proliferation of pathogens or cancer cells), to functional assays of specific target proteins. Meanwhile the search for target protein inhibitors have always been governed by the same two approaches:

Random screens. Whether one at a time or run in parallel by brute force, many available substances are tested for efficacy. Often a wide net is cast by screening an enormous and diverse compound library (as many as 1.7 million compounds; Plouffe et al., 2008). There is a tendency to test only representatives from a given group of substances; an intelligent step to increase the efficiency of the pharmacologic hunt wherein the "hit" group is explored in further screening. But reduced screens increase the odds of missing subtle differences that might allow target binding by nonsampled members of the group. Thus where resources permit, large screens are conducted. From the 1960s to the 1980s "high throughput screens," enabled by extraneous technology such as assembly lines and robotics, permitted the pharmaceutical industry to blossom almost strictly based on the paradigm of vast screens (Strebhardt and Ullrich, 2008). This is still the most common approach used by the pharmaceutical industry today. Without deep understanding of the target chemistry, sampling nature's pharmacopeia may well be the most efficient approach to finding a starting place: a hit compound (Chong et al., 2006; Weisman et al., 2006; Jenwitheesuk et al., 2008; Table 1).

Directed exploration. Intuitively, the response to finding an agent that has any noticeable desired effect is to seek better effects by similar agents. Intelligent searches for pharmacologically active substances generally follow explorative sampling around successful compounds already discovered in random screens (Schreiber, 2000). Similar existing compounds can be tested for more desirable activity, or chemical modifications can be made by substituting, converting, and adding moieties (Abdi et al., 2010).

Those of us who develop computational techniques for drug discovery tend to consider targets from infectious and noninfectious diseases as the same, but the reality is that they are not. Generally, the goal is to inhibit targets of infectious diseases (increase the therapeutic index), but a human disorder that is not directly caused by a pathogen may be caused by the malfunction of a protein, so inhibition is not always the goal. For malfunctioning proteins, the goal may be to discover a drug that promotes the active conformation or overcomes the loss of effective signal activity. While computational drug discovery techniques are quite robust, molecular etiology must be considered to select the target and to specify the desired pharmacologic effect.

2.d. Similar active substances for rational selection

Sophistication in understanding the similarity of pharmacologic agents was first developed in the ancient processes of chemical extraction. Similar separation in organic solvents indicates similar polarity and hydrophilicity, and often foretells identical chemical moieties. Comparison of compounds with similar chemical properties to compounds with similar pharmacologic effects resulted in the concepts of pharmacophores (Cammarata and Menon, 1976) and quantitative structure activity relationships (QSAR; Hansch, 1969). These concepts enable intelligent exploration of the chemical and structural space around the natural substrate.

In a case for which the activity profile of a vast drug bank is known for a particular pathogen, analysis of similarly active compounds can facilitate understanding of the basis of molecular recognition between a small molecule and its protein target (Plouffe et al., 2008).

2.e. Cycling between random and directed searches

Directed exploration requires either identification of the physiologic substrate, a hit compound, or deep knowledge of the target (discussed later in this chapter). Sampling around successful compounds with similar active substances represents an additional round of screening, which can be iterated to attempt improvement. The process of following up an initial hit with rational design is termed "lead optimization," and is discussed further in section 3.c.

By modifying functional and structural groups to enhance targeting by initial hits, the pharmaceutical industry and the field of organic synthesis generally have massively exploded the available pharmacopeia (Cupido et al., 2007). Thus directed exploration can optimize a hit compound for a desired effect, and the process also feeds back more bioavailable compounds for random screens generally. Chaotically cycling between the two approaches for the gamut of medical purposes during the past century of drug discovery has clearly resulted in enormous productivity (Schneider and Fechner, 2005), and an evolution of the available pharmacopeia.

2.f. Screening in current Pharma

Very generally, the approach of major pharmaceutical companies is to run a large chemical compound library against target proteins of interest using a simple protein based in vitro reporter system, or simply: high throughput screening (HTS). The initial hits are then assessed in progressively complex and representative in vitro and in vivo model systems, whereupon active compounds are considered "leads" to a drug. Finally the long and arduous process of three phases of clinical trials is undertaken to obtain approval from a governing agency (FDA in the US).

The traditional cycling between random and directed searches is inefficient since the blinded screens result in a vast number of hits and leads that fail to be effective or safe in humans. The pharmaceutical industry (Pharma) sets prices to derive profit beyond the tremendous overhead (Adams and Brantner, 2006), and as a result therapeutics are often out of reach to those who need it most. For many infectious diseases there is little or no profit to be made, as the sole prevalence is in impoverished peoples. As a result many potential drug targets for these diseases are ignored by Pharma (Orti et al., 2009).

Although much of Pharma follows traditional methods, the economic opportunities within increasingly complex diseases have driven it to make some of the most significant advancements (Borisy et al., 2003; Becker et al., 2004; Plouffe et al., 2008; Natoli et al., 2010).

3. Established technological advancements

3.a. The exploitable niche

Many proteins have an enzymatic cleft relatively specific to its substrate(s) by patterns of charge, flexibility, and space (Jensen, 1974; Khersonsky et al., 2006). Metabolites enter the cleft and emerge with some chemical alteration. Reaction products have lower affinity for the active site, so they dissipate. The physiologic substrate will not bind to the enzyme irreversibly, as the purpose of the interaction is generally to modify ligand, target, or both, and thereafter distribute this change as a signal to the cell or environment. This requirement of physiologic ligand expulsion creates the quintessential exploitable niche for drug discovery.

The protein target is evolved to stabilize a thermodynamically unstable substrate ligand transition state. The protein might binds the ground state but it stabilizes the reaction intermediate, which decreases the activation energy for the reaction and thereby promulgates the ligand product state. Yet the protein is also evolved to favor egress of the product after the reaction. The protein is most fit to bind the intermediate (rather than ground or product states), but as this state is by its very nature transient it should be possible to find substances which are similar to the reaction intermediate but stable in this form. As the transition state is the thermodynamically least favored state, applying a ligand which is thermodynamically stable in a similar form will kinetically overwhelm the protein and thereby inhibit target protein activity (Keiser et al., 2009; Abdi et al., 2010).

3.b. Target dissection for inhibitor design

Proteins fold into complex structures. Some parts are evolved to stabilize the topologic fold, while others carry out physiologic interactions, and others yet do both (Horst et al., 2010). The chemical structure of the active cleft dictates the function and the range of adoptable structural conformations. Modeling the pattern of tolerated and optimal moieties across the active cleft enables design and virtual selection of pharmacologic inhibitors (Jenwitheesuk et al., 2005). The presentation of hydrophobicity, polarity, and charge across the surface dictates where complimentary functional groups should be placed.

Affinity can be understood as change in free energy upon binding, which represents the sum entropy and enthalpy changes for protein, ligand, and solvent. Significant conformational constraints can decrease entropy of the ligand and protein during binding. The protein attracts binders by the potential energy stored in the hydrated hydrophobic pocket. Matching any nitrogen, oxygen, or fluorine moieties with a hydrogen bond adds further enthalpic drive to the reaction, resulting in a more strongly binding and therefore a more effective inhibitor (see section 6.e. for further understanding of enthalpy and entropy in computational drug discovery). Thus knowledge of the three dimensional chemical structure of the target active site enables design of strong binders which might be used pharmacologically as inhibitors.

3.c. Rational design and optimization

As discussed above, the affinity of a hit compound can be improved by strengthening contacts identified by analysis of the active cleft of the protein structure (enthalpic improvement). Successful inhibitors bind a range of active site conformations, or induce a particularly stable conformation. The natural substrate of the target protein can be studied to understand the contacts which stabilize the physiologic interaction, but the chemical scaffold of the metabolite cannot often be used to design a stable inhibitor. In part for this very reason, a good inhibitor generally avoids covalent modification by the target protein, but the inhibitor may be modified by other proteins to increase affinity (e.g. partial breakdown during first pass metabolism, or phosphorylation by other enzymes in the targeted pathway).

The goal of optimization is to improve the therapeutic index: to increase activity (efficacy) and decrease toxicity (specificity). Optimization steps can increase affinity or specificity, but seldom improve both simultaneously. Goals for efficacy include outcompeting the physiologic ligand (metabolite), while the more complex goals for toxicity include minimizing other reactions (specificity) and producing a favorable absorption, distribution, metabolism, and excretion profile (ADME). To balance pharmacokinetic properties during lead optimization the ADME profile is considered in the context of the clinical indication (Ekins, 2005).

Possible modifications to optimize organic inhibitors are nearly infinite. They include adding any chemical group from a single carbon (methyl group) to a heterocyclic, tethering components to force a particular conformation, or swapping atoms to alter ionic or hydrogen bonding, or patterns of hydrophobicity. Changes made to bioactive peptides alone include multimerisation and additions of lipid, polyethylene glycol, or peptidomimetic features (Bellmann-Sickert and Beck-Sickinger, 2010).

Of course much of the understanding of protein ligand interactions comes from analysis using computational graphics programs. In accord, exploration of affinity optimization can be carried out "by hand" at the computer terminal, applying experience and intuition to fit specific chemical moieties to concavity forms and electrochemical contacts (Noble et al., 2004; Abdi et al., 2010). The optimization process can also be applied by cyclically testing alterations of virtual hits from computational docking (Becker et al., 2006). Alternatively, computational methods can be used to produce a group of virtual hits, of which enough compounds are tested at the bench to secure multiple submicromolar hit compounds for follow up animal experiments (Becker et al., 2004; Desai et al., 2004; Desai et al., 2006; Jenwitheesuk et al., 2008; Table 1). Improvements to the latter approach are the subject of sections 4-6 of this chapter.

While many examples of structure based drug optimization exist, a quintessential example of computationally guided optimization is found in the work of Becker and colleagues, in the production of PRX-00023 as a lead compound for major depressive disorder and generalized anxiety disorder (Becker et al., 2006; Table 1). The 1nM K_i hit arylpiperazinylsulfonamide (PRX-93009) was found using purely computational methods by modeling the 5-HT_{1A} GPCR (serotonin receptor 1A), docking a library of 40,000 compounds, and running 78 virtual hits in an in vitro reporter system (Becker et al., 2004; Table 1). While the magnitude of target activity demonstrated great success, the compound presented suboptimal selectivity and pharmokinetics. The same group ran the compound, arylpiperazinylsulfonamide, against fifty other GPCRs *in vitro*, modeled the experimentally derived interactions (α_1- and α_2-adrenergic receptors and hERG), and optimized selectivity for 5-HT_{1A} by removing or substituting moieties that strengthened off-target contacts, and by adding many compensatory on-target contacts (Becker et al., 2006). The resulting compound, PRX-00023, was sufficiently selective to 5-HT_{1A}, and presented a pharmacologic availability profile similar to existing drugs for the same indication. The entire process from the computational screen through entry into phase III clinical trials took only two years (Becker et al., 2006). Unfortunately while it was tolerated, the efficacy was not enough (de Paulis et al., 2007; Rickels et al., 2008). Nonetheless, this adventure demonstrates that computational methods can facilitate lead compound discovery and catalyze the process of getting to the question of real clinical efficacy.

3.d. Multitarget dosing

In many cases, no single drug is sufficiently effective in the therapeutic range to cure the disease, or even to reduce symptoms or recurrence effectively. Thus multiple drugs can be combined to heighten the effect. Simultaneous effects on multiple targets can decrease therapeutic doses, so that less efficacious and slightly more toxic compounds can be used safely. As well, pathogens often develop resistance to single drug therapy, but simultaneous occurrence of multiple resistant mutations are exponentially less prevalent. The "multitarget" concept of targeting more than one protein in a single dose emerged to address these issues.

Perhaps the most successful application of intentional multitarget drug administration is presented in dosing with inhibitors of HIV reverse transcriptase, protease, and integrase in the fight against HIV/AIDS (Hirschel and Francioli, 1998). Multidosing is titrated in a trial-and-error manner, using patient suffering as the error. Because of this undesirable situation, novel approaches have emerged to model synergistic effects of polypharmacology. For example, combinatorial effects have been tested in vitro using an automated robotics and informatics pipeline. Pairs of substances that display synergistic inhibition of *Candida albicans* growth, cytokine production, and tumor growth, exhibit complex efficacy patterns: highly nonlinear effects are observed in plots of the concentration of one compound versus the other. The complexity is evident of either single protein targeting by different inhibitors, or more likely: inhibition of multiple proteins involved in the same physiologic process (Borisy et al., 2003). Further examples, design, and benefits of polypharmacology are discussed throughout this book.

4. Computational drug discovery

The structure guided computational approach to evaluating protein-ligand interactions generally consists of three steps: (a) conformational sampling of the rotation, translation, and torsion angle degrees of freedom between the protein and ligand, (b) scoring the resulting interactions with a discriminatory function to identify native and near-native complexes from a set of incorrect conformations, and (c) ranking possible ligands to distinguish between strong, weak, and non-binders. Despite previous successes, limitations persist in structure-guided drug screening and design implementations to date. The principle disconnect between what computational drug discovery is hoped to be and the reality of what it provides, is that computational predictions *enrich* for compound-protein activity rather than *design* it. In the best reported cases there are still many false positives and false negatives (Table 1); structure guided discovery is a rational starting point, but does not yet provide a comprehensive view of biologic interactions.

4.a. Principles and data sources

As successful approaches to protein structure prediction do not model any part of the folding process, modeling the physiologic conformation of a bound ligand has little if anything to do with the actual physical process of binding. While the hypothetical situation of modeling the wave function for each atom in the system could produce a descriptive simulation of ligand binding, this approach is computationally intractable. Again analogous to the example of protein structure prediction (Moult, 2005), the methods most successful for modeling the stable end state conformations are those that directly consider many measurements of other end state conformations (Kitchen et al., 2004; Bernard and Samudrala, 2009). In essence, physical properties such as interatomic distances, repulsion, or attraction are taken to build models to estimate stability of the protein-ligand system. The strength of computational methods is in automating these analyses across enormous amounts of ligand to protein pairs.

4.b. Docking

The term "docking" describes placement of a ligand onto the molecular surface of a protein, in a manner that mimics the real physical interaction as closely as possible. The interaction of any two particles above absolute zero temperature are dynamic, so the protein-ligand physical interaction includes a distribution of conformations that may be clustered extremely tightly (<0.1 Å root mean squared deviation) or include significantly dynamic protein and ligand movements, as can be found in the range of holo PDB structures (Berman et al., 2000).

Docking can be performed in a manner to offer alternative molecules to an initial hit or known physiologic substrate, in which a base molecule provides a starting conformation. Our group recently demonstrated the utility of this approach to peptide inhibitor design, wherein we took as the starting conformation a strand from the physiologic substrate protein (in the PDB structure), and substituted alternate residue side chains, following a greedy search protocol (Costin et al., 2010; Table 1).

The complexity of the docking problem expands with the degrees of freedom of each ligand. Unfortunately, while proteins are often treated as rigid surfaces on which to dock a ligand, they are dynamic as well, including movements in response to ligand binding, termed "induced fit" (Koshland, 1958).

Translation. In the most simple case of docking, a roughly spherical ligand (such as a metal ion) is translated about the protein. The translation space sampled can be a grid, limited to a region of the protein or a defined space surrounding the protein, or can be continuous, in which case movements from a starting point must be guided by a scoring function. In either case it is tractable to sample within 0.1 Å of the binding site in a suitable model of the protein structure, and so selection of the real binding site is left to the scoring function (discussed in the next section).

Orientation. For the anisotropic case of all multiple atom ligands, orientation must be considered. The rigid ligand is rotated about the grid or starting point. To achieve the same 0.1 Å resolution as described above for the isotropic translation search, the requisite search space would be increased 51-fold for a hydrogen molecule (the number of nonredundant 0.1 Å square grid points on a 0.76 Å diameter hemisphere) and exponentially more for ligands of greater size. However, this search is still tractable, and has been applied in various attempts to break down more complex molecules into rigid fragments.

Bond rotation. Nonrigid ligands contain rotatable single (sigma) bonds that dramatically increase the sample space. Simplifications can be made to some rotatable bonds to decrease the impact on sample space, for example removing bond angles that produce eclipsing of large repulsive chemical groups. However, multiple rotatable bonds in a ligand generally breaks the tractability of the docking search, and heuristic strategies must be employed. The earliest versions of docking methods simplified flexible ligands as rigid (Kuntz et al., 1982), yet even now rotatable bonds not only increase the search space but decrease the accuracy of all docking methods (Kitchen et al., 2004; Plewczynski et al., 2010)!

Most docking methods combine the three types of movements (translation, orientation, rotation). The combined movement is generally guided by a scoring function, but the way they are applied can be very different (Ewing et al., 2001; Kitchen et al., 2004). For example the movements from one sampled conformation to the next might be decided by comparing scores for the first and a stochastic progression (Metropolis Monte Carlo approach), or the trajectory resulting from an estimate of forces in the system (molecular dynamics approach). Therefore, at the heart of the docking protocol is the scoring function.

4.c. Scoring and discriminatory functions

Functions for evaluating protein-ligand interactions are generally referred to as 'scoring functions.' Scoring functions applied to the problem of selecting the most realistic ligand conformation amongst a set of docked poses is a 'discriminatory function.' Protein-ligand scoring functions are categorized into molecular dynamics force fields, empirical functions, and knowledge-based functions. Force fields are commonly built to explicitly model physical forces (acceleration) of idealized gas phase enthalpy including electrostatics and van der Waals forces (shape complementarity; Lennard-Jones, 1924). Often left out are the contributions of entropy (e.g. torsional) and solvation, while heuristic considerations such as number of hydrogen bonds are most often included (Kitchen et al., 2004).

The assignment of the terms "knowledge-based" and "empirical" are historical; both use experimental data to build scores and

coefficients. Both perform statistical comparisons of the query case to many bench laboratory derived binding affinities and/or structural conformations. Generally, empirical functions combine physical terms by regression analysis of experimental binding data, whereas knowledge-based functions derive scores for ranges of spatial parameters (distance, torsion angles, or voxels) from experimentally derived structures without any attempt to divide the underlying physical forces (Kitchen et al., 2004; Bernard & Samudrala, 2009).

The molecular dynamics force field assisted model building with energy refinement program (AMBER) represents the flagship molecular dynamics function. AMBER models the potential energy of each conformation with a set of terms for covalent bonds, bond angles, torsion angles, electrostatics, and van der Waals energies (Weiner et al., 1984). AMBER has gone through continual updating by many contributors, to progressively incorporate physics-based models of diverse systems and optimize the coefficients of the formula for specific types of interactions (Ponder and Case, 2003; Case et al., 2005). Although molecular dynamics force field functions hypothetically have the capacity to direct ligand docking into the lowest energy conformation, using these functions to model an entire protein-ligand system has the tendency to result in models continually expanding out from the physiologically compact state; artificial constraints can be used to hold the model together, but these constraints represent a deviation from the goal of physics based modeling, are not generalizable, and the results are usually not predictive. Nonetheless, judicious use of a limited progression of molecular dynamics steps guided by these functions can be highly useful for modeling protein-ligand systems (Jenwitheesuk and Samudrala, 2003a).

Increased success in developing discriminatory functions have often arisen from specifying the type of protein target, with the presumption that different forces dominate ligand binding by proteins such as transmembrane receptors and transcription factors. However, our group recently developed a generalized knowledge based discriminatory function score to select optimal poses for any type of ligand, within a margin of error which can be sampled by a course lattice method. This knowledge-based function outperforms more than 20 other published ones in several docking decoy tests, by analyzing interatomic distance distributions from the repeating units of high resolution small molecule crystallography structures (Bernard & Samudrala, 2009). In part the strength of this method is the quality of the intermolecular contacts: the crystals of small molecules are much more regular than those of proteins, and so more accurate structures are modeled from the electron density maps. However, exhaustive consideration of the statistical derivation makes this function even better. We considered radial versus normalized frequency distributions, mean versus cumulative reference state, reduced versus complete composition, and the maximum interatomic distance to be considered (cutoff). Across a diverse set of protein interactions with small molecules, other proteins, and DNA, the radial mean reduced derivation performed with the most accuracy (Bernard & Samudrala, 2009). The result is a highly accurate discriminatory function with perhaps fine enough resolution to make a continuous function that could act as a force field.

Future work to improve scoring functions includes efforts to bolster the accuracy of knowledge-based or empirical functions to address the goals of molecular dynamics approaches. If forces are to be divided into physical contributions, proper handling of entropic and solvation contributions are needed (section 6.e). Further improvements include representing three dimensionality to model the physical intricacies of electron sharing through hydribidized orbitals (e.g. sp3), and multi-body potentials that can account for resonance patterns (Ngan et al., 2005); there are enough high resolution structures in the Cambridge structure database (Allen, 2002) to approach these goals (Bernard and Samudrala, 2009).

4.d. Relative affinity ranking

Ultimately there are two roles for the ligand pose selected by a scoring function: to be the representative for ranking amongst the best scoring poses of other ligands, and to identify the pattern of contacts that might be retained or improved during optimization. Ideally, protein-ligand scoring functions should be able to identify the native or near-native ligand pose from a set of incorrect conformations (i.e. discrimination), and to distinguish between small molecules that do and do not bind a target protein (i.e., relative affinity ranking). This is unfortunately not the case with current methods, as discriminatory functions perform poorly at correlating scores with experimental binding energies. An ideal ranking function would accurately calculate the free energy of binding. Relation to the affinity estimation for another ligand (another drug or physiologic substrate) would be sufficient to estimate biological activity: this thermodynamic understanding would indicate which ligand would outcompete the others by binding strength. The kinetic considerations (e.g. target tissue concentration) could be designed around this understanding. Clearly ranking functions could be extremely useful to computational drug discovery, but currently no function has been shown to consistently reach these goals. An accurate relative affinity ranking function is needed in the field of structure-guided drug screening and design if these predictive methods are to serve as a useful an complementary tool to prospective experimental investigation.

Knowledge-based functions perform quite well at discrimination (Bernard and Samudrala, 2009) but inaccurately provide scores proportional to the size of the ligand, due to their simple additive nature, and therefore may be of limited utility for relative affinity ranking. Empirical scoring functions fitted to experimental binding energies perform rather poorly, especially for classes of molecules not included in the training set, and significantly lack in discriminatory ability. Often experimental complexes are used to correlate scores with experimental binding affinities; in practice this is not useful, as

Table 1. Summary of recent prospective drug discovery screens.

				bench experiments					
disease	**protocol**	**in silico**	**targets**	**in vitro →**	**hit**	**best(μM)**	**type**	**Group**	**Year**
DiabetesII	throughput:		1	400,000 →	85 <100μM	4.2	IC50	Schoichet/Doman	2002
Malaria	throughput:		organism	2,687 →	19 <1μM	0.003	IC50	Sullivan	2006
Malaria	throughput:		organism	2,160 →	36 <1μM	0.010	IC50	DeRisi	2006
β-lactamase	throughput:		1	70,563 →	0 <30μM	none†	IC50	Schoichet/Roth	2008
Malaria	throughput:		organism	1,700,000 →	5973 <1.25μM	nr	IC50	Winzeler/Shultz	2008
					648 <100nM	nr	IC50		
Malaria	throughput:		organism	1,986,056 →	13533 <2μM	nr	IC80	Garcia-Bustos	2010
DiabetesII	docking:	150,000	1	7 →	5 <100μM	21	Ki	Zhang	2000
Cancer	docking:	100,000	1	2 →	2 <100pM	0.00003	Kd	Shakhnovich	2001
DiabetesII	docking:	235,000	1	365 →	127 <100μM	1.7	IC50	Schoichet/Doman	2002
Malaria	QSAR:	12	1‡	12 →	3 <100μM	5.7	IC50	Freire	2002
Malaria	QSAR:	9	4	9 →	2 <1μM	0.3	IC50	Freire	2003
Malaria	docking:	241,000	1	100 →	5 <10μM	1.0	IC50	Avery	2004
HumanGPCRs	docking:	150,000	5	309 →	50 <5μM	0.021	EC50	Becker	2004
Malaria	docking:	355,000	2	84 →	7 <10μM	9.5	IC50	Avery	2006
Depression	dock,QSAR:	40,000	1	78 →	9 <1μM(Ki)	0.3	IC50	Becker	2006
β-lactamase	docking:	70,563	1	18 →	4 <200μM	70†	IC50	Schoichet/Austin	2008
Malaria	docking:	2,344	14	16 →	7 <1μM	<1	IC50	Samudrala/vanVoorhis	2008
Tropical	QSAR:	3,665	11,714	2 →	1 NMR	nr	NMR	Sali/Marti-Renom	2009
HumanMany	QSAR:*	3,665	1,133	30 →	20 <1μM	0.001	Ki	Schoichet/Roth	2009
Dengue	peptides:	14	1	14 →	2 <1μM	7.0	IC50	Samudrala/Michaels	2010

* search for target, in vivo activity already known.
† extensive reversibility studies.
‡ the human homolog was also considered as an antitarget.
nr not reported.

the objective is to find new compounds that bind to a protein target. The most relevant experiment is to test known inhibitors against alternative protein structures that are not bound by the small molecule of interest, and then evaluate the correlation coefficient (which is invariably lower; Fan et al., 2009).

The ability to accurately discriminate the correct binding mode of individual ligands and that to rank the relative binding affinity between different ligands can be treated as distinct computational modeling problems. All protein ligand scoring functions can be applied as ranking functions, but dissecting apart docking and ranking allows for considerations more important to each problem. For example, counting hydrogen bonds and calculating loss of torsional entropy is essential to ranking ligands, but many conformations of the same ligand can be equivalent for these factors (Kitchen et al., 2004). Therefore, the methods used for discrimination and relative affinity ranking should be separated into distinct functions and developed independently, which has not previously been the case.

4.e. Comparison of docking methods

Many methods have been created to dock ligands to proteins (Kitchen et al., 2004). But bias and overtraining have impeded attempts of evaluation in the field of computational biology, as demonstrated for protein structure prediction with the solution of the CASP experiments (Moult, 2005). Blinded or independent examinations are proper means for unbiasing assessments of predictive methods. Minimizing bias optimizes the estimation of the accuracy in prospective experimentation, which is the purpose of these methods. A recent experiment performed such an independent test between seven docking programs (Surflex, LigandFit, Glide, GOLD, FlexX, eHiTS, and AutoDock) on 1300 holo structures from the PDBbind database. Ligand conformations were converted through SMILE strings using two different tools (Corina, Omega2). Two commercial products (GOLD, eHITS) outperformed the other methods, with mean accuracy <3.0 Å RMSD and >55% of cases <2 Å RMSD. The use of holo rather than apo structures is a caveat to the relevance of these findings to prospective drug discovery. As well, it is likely that the examined methods were trained on some of the same structures as those used to test them, which gives unfair advantage. While prospective experimentation is the only true test of a computational method, this study describes the most independent comparison of methods for drug discovery known to us (Plewczynski et al., 2010).

4.f. Ligand comparison

Small molecule structure activity relationships are applied to find active substances similar to initial hits found through bench or computational techniques (Hansch, 1969). The underlying concept follows that the activity of the substrate transition state can be analogized by chemical similarity to any other compound (Abdel-Rahman et al., 2004). It follows that the activity of a hit ligand can be analogized by chemical similarity to any other compound. The ligand comparison is calculated by comparing the geometric distribution of electronegativity and hyrophobicity for the hit ligand against a database of existing small molecules. Improvement accurate predictions of the similar active substance are found when limiting the database to known bioactive molecules. This approach is powerful in part because of the small requirement for computational resources compared to docking.

While the structure activity relationship of small molecule organics has been applied to ligand optimization traditionally, the concept of similar chemical structures having similar bioactivity has recently been applied to discover initial hits (Nezami et al., 2002; Nezami et al., 2003; Orti et al., 2009; Keiser et al., 2009; Table 1). The rationale here is to use known substrates or predicted ligands in place of the initial hit. It is logical that the physiologic substrate would be a productive starting place for detection of similar active substances. This brand of applications of structure activity relationship will make a large impact in making drug discovery more efficient, and expanding our understanding of the co-evolution of proteins by their similar physiologic substrates.

Other computational methods compare structural and chemical properties among protein-ligand binding sites directly without considering ligands (Gold et al., 2006; Das et al., 2009; Weill et al., 2009; Chen and Honig, 2010). For example, the method of Das and colleagues dissects a binding site into a profile of probabilities that a surface patch with a particular physicochemical property will present at a specific distance to another on the binding site surface (Das et al., 2009). When the binding site and tertiary structure is known or predicted, this analysis enables rapid detection of target identification, understanding of multitarget effects, and suggests compounds to screen for pharmacologic inhibition. Binding sites can be predicted by sequence analysis (Wang et al., 2008; Horst and Samudrala, 2010) or mapping by structural similarity (Orti et al., 2009).

5. Recent technical Improvements

5.a. Automated binding site identification

A variety of sequence and structure based approaches are used to predict protein-ligand binding sites. For many globular soluble enzymes the binding pocket is easily identified by its characteristic narrowness and depth, which allows harbor of small molecules. This analysis can be automated by geometric measurements, for example surface concavities can be found by comparing the accessibility of different sized spheres to the solvent exposed surface (Greaves and Warwicker, 2005). Meanwhile, many protein active sites are not as obvious from the protein structure; these harder problems demand sophisticated bioinformatic tools (Gutteridge and Thornton, 2005).

Often a ligand can be mapped to the query structure from a holo template protein identified by sequence or structural similarity (Lopez et al., 2009; Orti et al., 2009; Roy et al., 2010). Where ligand mapping is not available and when results are not consistent, conservation analysis is useful. Particularly, proteins from poorly characterized families cannot always be understood by direct similarity analysis. Sequence analysis can evaluate multiple aspects of evolutionary conservation and residue identity to predict binding sites with comparable accuracy to structure-based methods (Berezin et al., 2004; Fischer et al., 2008; Wang et al., 2008). Structural analysis or structure prediction can be combined with conservation calculations to improve interpretation (Landau et al., 2005; Roy et al., 2010; Horst and Samudrala, 2010; Horst et al., 2010). Our group has found across many protein active sites that hidden Markov model estimates of relative conservation entropy is the most accurate single predictor of residue functional importance (Wang and Samudrala, 2006; Horst et al., 2010).

Differences in residue identity within otherwise similar binding sites control metabolite specificity and variation in enzymatic reactions (Ashworth et al., 2006; Jiang et al., 2008). Thus the residues that specify ligands are often not conserved. More advanced analysis is indicated to find these residues; function prediction methods may be useful to select atomic contacts to targets during computational drug discovery. Our group has demonstrated that machine learning can be used to transfer dissections of structure and function from many proteins to predict the active sites of highly different query proteins (Wang et al., 2008; Horst and Samudrala, 2010; Horst et al., 2010). Our methods predict protein-ligand binding sites de novo using an algorithm that generates meta-functional signatures (MFS) by combining multiple sources of information reflecting functional importance. MFS can be applied to a protein sequence or structure and has been shown to be more effective in identifying functional sites than other popular methods (Wang et al., 2008; Horst and Samudrala, 2010; Horst et al., 2010).

5.b. Docking with protein target dynamics

Biologically active proteins are in continuous motion, yet the majority of protein structure information is limited to the most stable form of a protein when crystallised in artificial conditions. Induced fit is a widely recognized challenge in computational drug screening, wherein the protein undergoes significant conformational changes upon ligand binding (Koshland, 1958). As a consequence, traditional rigid protein-ligand docking is insufficient for structure guided drug screening, and is often misleading. The active cleft surface is treated as rigid, though a conformational shift occurs upon binding a physiologic substrate, inhibitor, or interacting protein. This conformational shift brings together the mediator functional groups of the catalytic reaction. The energetic force to bind the reagent metabolite is generally enthalpic, so the bound holo conformational state of the protein is closest to the optimal pharmacologic target. Dynamics simulations increase the possibility of surveying a physiologically relevant conformation beyond using the static crystal structure alone. For example, our group showed that for a group of HIV-1 protease inhibitors, using molecular dynamics to model changes in the target

protein improves the correlation coefficient of predicted score versus measured affinity from 0.35 to 0.88 (Jenwitheesuk and Samudrala, 2003a).

Modeling target and ligand flexibility aids the multitarget approach. Multiple stable conformations or highly flexible portions of a ligand increase the range of possible target clefts in which the ligand might fit. The benefit of ligand flexibility for action on multiple proteins is exemplified in the difference between first and second generation HIV protease inhibitors (Freire, 2002).

Along with our work demonstrating the importance of target protein dynamics in computational docking, many other groups have incorporate target flexibility into their software. But as for the rotatable bonds of ligands discussed in section 4.b, each additional bond considered for rotation dramatically increases the sample space, and so slows down the search. Therefore our approach of using short spans of molecular dynamics (200 steps) appears to be most computationally reasonable, and is used widely (Jenwitheesuk et al., 2008).

5.c. Structure modeling for target docking

The concept of template based modeling can be understood and applied in different ways. One approach that has been shown to work is to model the query protein based on a template, and then dock to this model. However, this is not the only way to make use of a template. It is not always necessary to build a structural model. If there is a known drug or ligand interaction for a template protein, this information may be transferred directly based on the similarity between the proteins (Orti et al., 2009). If docking is indicated, it may be more relevant to dock to the template itself rather than a model built using the template - the accuracy of the template is known, while the model built with the template is guaranteed to be less accurate (Fan et al., 2009). A good template will have a highly similar binding site, sufficiently similar that the differences in residue identity can be modeled after docking.

The structures of all human GPCRs have been modeled with I-TASSER (Zhang, 2008), the best existing protein structure prediction method (albeit an older version), and are freely available (Zhang et al., 2006). Various publicly available methods are capable of modeling structure and ligand docking for GPCRs. For example, our group combined I-TASSER with our consensus refinement method (Liu et al., 2009) to perform amongst the very best groups in a prospective prediction experiment to predict structure and ligand conformation for the second human GPCR Xray structure (Michino et al., 2009). Meanwhile, the proof of concept for all modeling drug discovery for GPCRs was accomplished in 2004 by Becker and colleagues, as discussed above in section 3c. Briefly: the authors modeled five GPCRs based on the bovine rhodopsin structure (PDB id 1f88; the only GPCR structure known at the time), used the anchor and grow approach in DOCK4.0 (Ewing et al., 2001) for ~150,000 compounds selected from ~1,600,000 based on physical properties, and ranked the resulting protein-compound pair conformations using in house software. The outcome of this study includes 50 substances with EC50 <5μM activity, a novel EC50 <100nM compound for four of the five target GPCRs, and an agonist lead compound (Becker et al., 2004; Table 1). However, there was no comparison performed to check for enrichment versus docking to the template rhodopsin structure.

A recent study explored the opportunity of template based modeling and docking for 38 proteins, 2950 ligands of known bioactivity, and 95,316 decoy ligands (Fan et al., 2009). The exploration was relatively thorough for protein structure modeling, using templates across a broad range of sequence identity (20-99%). In this study the consensus result of docking against multiple template based models was better than docking to the single best model or even the apo structure of the protein (in most cases), and in many cases the consensus model accuracy approached that of docking against the target holo structure. Meanwhile, this study also compared bioactive ligand selection enrichment for docking to the homology model templates versus the models. There was a slight trend for holo templates of sequence identity below 40% to more accurately select the bioactive ligands than models derived from the holo template (R=0.22 across sequence identity range). There was no clear range for which it would be better to use homology models. When using apo templates or models derived from them, the correlation for sequence identity dropped (R=0.07): sequence identity is not predictive of whether it is better to use the apo template itself or a model derived from it (Fan et al., 2009). On average, docking to templates produced insignificantly higher enrichment for bioactive ligand selection than docking to models of the target protein (Student's paired 1 tailed t-test p=0.29). So, based on this study using the latest versions of MODELLER and DOCK, it appears that for the purposes of docking, there is no great benefit to spending the computational resources to build all atom models of target proteins. Meanwhile, the success of the consensus of models suggests that clustering may be useful for finding the best template on which to dock, and that improvement in structure prediction methods may breach the accuracy of docking to homolog holo structures. Nevertheless, the high resolution of the template is for now a better data source of analysis, whereupon our ability to detect the evolutionary connection between homologous proteins is the most powerful tool.

5.d. Ligand-target networks

Metabolic systems bring an environmental substrate through a series of reactions that add or remove chemical moieties. The majority of the substrate is often maintained through the process, such that each protein controlling the metabolic network will recognize similar features of the substrate. Therefore if a drug is selected or designed to inhibit a particular protein target, it is highly likely that the drug will inhibit multiple proteins of the metabolic network (Csermely et al., 2005; Hopkins, 2008). Thus many drugs achieve higher efficacy by unintentional pathway multitargeting (Kohanski et al., 2010), with benefits described throughout this text.

Network targeting involves activity of a compound across multiple pathways. Multiple routes of attack may be necessary to effectively stop neoplasms or pathogens that have multiple compensatory pathways to allow survival and proliferation. More and more we are learning that simple linear or cyclic pathways are the exception rather than the rule, so even to inhibit a single pathway it seems that multiple indirectly connected proteins must be inhibited (Hopkins, 2008). If one adopts a multitarget philosophy, the principle difference is a need to monitor the interconnectivity of the targets, maximizing relevance to the clinical question.

6. Emerging concepts

6.a. Starting with nature

The current drug discovery process itself both mimics and expedites the natural evolution of bioactive products. Living organisms have influenced the creation and relative abundance of chemicals on Earth. For example the production of oxygen by conifers which enabled aerobic metabolism: the cyclic feedback between life and that which is traditionally considered nonlife (small molecule organic compounds) describes a co-evolutionary pattern which can be exploited in drug discovery.

The current diversity of natural chemicals emerged within the same evolutionary soup. This shared evolutionary chemical context sets the stage for various organisms to use the same compounds to control different processes, making one molecule relevant to diverse physiological activity. The observation that structural folds are largely conserved, even when sequence and function are not, provides logical evidence that one compound can be an excellent initial candidate for many different protein targets. The topological forms of proteins (folds), present much more consistency than those of small molecules. For example, the proteins of various metabolic pathways appear to have evolved from the same template protein, with mutations conferring the ability to perform different chemical alterations. Meanwhile the binding site within a particular pathway is relatively conserved, and a ligand which gets in the way of a reaction in one protein will be promiscuous to the pathway. The result of these patterns of evolutionary divergence is that natural chemicals are highly multitargeting (Jenwitheesuk et al., 2008; Dancik et al., 2010).

The network of targets for existing drugs reveals physiologic relationships between the proteins within or between proteomes (Keiser et al., 2009). Particularly, not all human disease targets are predicted to be bound by natural small molecules, and it may be that the respective interaction networks are distinct (Dancík et al., 2010). The relatively unique human drug target network may be explained as bearing those more unique protein functions for which there are minimal compensatory self-righting mechanisms. The uniqueness of the target proteins seems to coincide with constrictions in the protein

interaction network, rather than network hubs which tend to be targetable by natural compounds. Presumably these network constriction human targets are not canonical enzymes, receptors, or channels - i.e. natural compounds are not their substrates. Thus for these targets, natural products and perhaps their derivatives may be insufficient.

Nonetheless, it is clear that there is some piece missing from the immediately preceding argument and referenced data, as 614 of the 974 new chemical entities discovered from 1981 to 2006 were natural products or derivatives thereof, many of which do target host proteins (Newman and Cragg, 2007). Leaders in bench drug discovery look to exotic organisms for drug leads continually (e.g. scorpion venom). Natural compounds can be very difficult to make outside of the source organism, and most exotic organisms are not cultivatable on a large scale. These compounds are the products of intricate protein mediated metabolic pathways not usually understood well enough to be genetically engineered into *E. coli* or yeast. Computational aid to retrosynthetic analysis enabled mass production of natural active products via total synthesis (Corey et al., 1985).

Thus natural products may not be able to inhibit or activate all host targets, but for any protein that acts upon a natural substrate, they likely will be useful. Thousands of years ago we recognized the pharmacologic capacity of many natural materials, and over the past few decades nature has still been the greatest source for new drugs. Natural compounds may not comprise ideal decoys for complex substrates such as DNA or other proteins, but we can keep looking to them as one principle source for bioactive compounds. The evolutionary pressure of competition clearly selected for organisms ready to fight other organisms - the resulting arsenal of molecular weapons are a robust starting point for rational drug discovery.

6.b. Peptides and their derivatives

Peptides represent a natural modular scaffold that can be easily designed to mimic natural substrates and binding partners for drug discovery. Knowledge-based protein structure prediction methods can be applied by reverse engineering the amino acid sequence of a natural binding partner to optimize binding. For example, our group created peptide inhibitors by redesigning the sequence of the dengue viral entry protein substrate, which prevents infectivity of dengue virus at the micromolar level (Costin et al., 2010; Table 1).

Peptides present some benefits for computational drug discovery relative to standard organic small molecules. One benefit is the modularity, which enables design, massive replication, and low production cost. Another aspect is that the chemical nature of side chain and main chain moieties are evolved to stabilize proteins, and therefore in some cases bind active sites more tightly than organic small molecules. The rapid degradation by endopeptidases is generally seen as a disadvantage because of inactivation and clearance, but: protease recognition is designable to some extent, peptide degradation minimizes immunogenicity, and some clinical indications call for rapid clearance.

Disadvantages of peptides also include susceptibility to nonspecific endoproteases (which are nearly everywhere in the body) and low oral bioavailability. Even with these disadvantages, peptide inhibitor design can be useful as part of an in vitro model to find or verify targets, and to identify specific binding site contacts to be targeted by small molecules. However, modifications to overcome disadvantages are chemically straightforward: multimerisation (e.g. polyethylene glycol), lipidisation, and adding peptidomimetic moieties (e.g. alternate atoms to substitute the amide bonds). Expressible peptides can be modified chemically to produce vast functional diversity suitable for many pharmacologic applications (Bellmann-Sickert and Beck-Sickinger, 2010).

6.c. Off-label drug use

FDA approved drugs present similar advantages to natural compounds due to their known bioactivity. Added benefits of screening existing drugs include the known safety and ADME profile, demonstration that the compound will get through first pass metabolism and get to at least some sites of action, and a hint of certainty that they will have the promiscuity of ligand-protein interactions discussed for natural compounds. Perhaps most importantly, since they are already approved for use in humans, the only barrier to clinical trials is demonstration of efficacy (Jenwitheesuk et al., 2008).

While it appears that use of existing drugs enriches screens for hit compounds, no one has done the proper side by side background control of testing a random sample of compounds. Current Pharma compound databases are designed to optimize bioactivity and ADME profiles in the case of presenting a hit inhibitor, e.g. following Lipinski's rule of 5 (Lipinski et al., 1997). However, four bench screens searching for inhibitors of *Plasmodium falciparum* demonstrate a trend towards enrichment for existing drugs (Table 1). Massive screens of ~2 million compounds from the chemical libraries of Novartis (Plouffe et al., 2008) and GlaxoSmithKline (Gamo et al., 2010) resulted in 0.35% and 0.68% micromolar hit rates, respectively. One thousand-fold smaller bench screens of ~2 thousand existing drugs for *Plasmodium falciparum* resulted in 0.71% (Chong et al., 2006) and 1.7% (Weisman et al., 2006), suggesting slight enrichment. Meanwhile, our computational screen of the same drug database selected 16 compounds, of which 44% are micromolar inhibitors (Jenwitheesuk et al., 2008; see Table 1 for further details of these studies). Although these giant Pharma companies have put decades of data and analysis into the design of their chemical libraries, similar if not better success rates can be achieved on a 1000x smaller scale if these screens are simply run with existing drugs. Moreover, our group has shown that publicly available computational methods can vastly enrich this search, and thus suggest existing drugs to be the starting set for any computational drug discovery project.

Understanding the biologic activity of known drugs of course makes it easier to repurpose them for desired physiologic effects. It is important to note here within this chapter on automated tools for drug discovery, that deep understanding of existing drugs and the disease of interest enable enrichment far beyond that currently available with contemporary computational methods. Accordingly, off-label uses are continuously being discovered. Carbamazepine, a widely used anticonvulsant and mood stabilizer, seems to combat hepatic fibrosis (Hidvegi et al., 2010). A lead for polycystic kidney disease was recently discovered by intuiting the target, for which an inhibitor was already developed in effort to treat diabetes (Natoli et al., 2010).

The trend for drugs approved for treatment of one disease to effectively treat another underscores the importance of epidemiologic studies to track disease patterns in medicated patients. Clinical informatics is an emerging field meant to handle questions like this. Meanwhile, the reward for repurposing an existing drug is highly similar to discovering its first use. In the US, intellectual property and patents are defined by the purpose; if you can figure out a new use for a hula hoop, you can patent it. A new use for an existing chemical entity is unique intellectual property. The only successful generalization of profit for a drug has been through manufacture of the physical drug itself. Thus, opportunity awaits in repurposing old drugs to new tricks.

6.d. Off-target effects

Virtual drug screening methods have been employed to help identify sources of off-target drug effects and investigate their potential to cause adverse or desirable side effects (Jenwitheesuk and Samudrala, 2007; Keiser et al., 2009). Desirable off-target effects include unintended multitargeting of other proteins in the target pathogen (Jenwitheesuk et al., 2008), fighting other infectious agents (Jenwitheesuk and Samudrala, 2003b; Jenwitheesuk and Samudrala, 2005a; Jenwitheesuk and Samudrala, 2007), and balancing untoward effects of other drugs being used in a polypharmacologic regimen. Through proper screening of relevant host and pathogen proteins and metabolites, current methods can enrich the design of off-target pharmacology.

Off-target effects can be predicted by ligand docking methods (Jenwitheesuk and Samudrala, 2003b; Jenwitheesuk and Samudrala, 2005a; Jenwitheesuk and Samudrala, 2007; Xie et al., 2007; Jenwitheesuk et al., 2008), ligand structure activity relationships (Keiser et al., 2007; Eckert and Bajorath, 2007; Keiser et al., 2009), and comparison of protein binding sites (Gold et al., 2006; Xie et al., 2007; Weill et al., 2009; Das et al., 2009). After decades of development (Hansch, 1969), SAR methods are emerging as clinically useful (Keiser et al., 2009; Table 1). Meanwhile, methods to compare protein binding sites and affinity ranking methods are still in their infancy, yet the latter has already demonstrated clinically significant utility (Jenwitheesuk and Samudrala, 2007).

Although virtual screening methods have been useful to inform drug design, many current methods are not able to account for off-target drug effects because they require structural information which is not available for most of the human proteome (Xie and Bourne, 2005). Further, due to the difficulty of crystallizing membrane proteins, structures for these proteins are highly underrepresented, making up less than 1% of the structures in the PDB (Walian et al., 2004). Nonetheless, nearly half of available drugs act on G protein-coupled receptors, a major class of membrane signal receptors (Gudermann et al., 1995). Therefore it is important to consider membrane proteins in the identification of off-target drug interactions. Although structural data is lacking, protein sequence data covers nearly the entire human proteome (UniProt Consortium 2007). Therefore it may be useful to develop computational protein sequence analysis methods to identify the similarity of protein ligand binding sites through their meta-functional signatures (Wang et al., 2008), which could model drug toxicity explicitly across human and pathogen proteomes. The most useful off-target screening methods will combine comparative analysis of ligand structure, protein structure, protein sequence, and the types of interactions between protein and ligand.

6.e. Affinity, entropy, enthalpy, optimization

Dissecting the contributions of entropy and enthalpy to changes in the free energy of a system through bench calorimetry has enabled a much more rationalizable approach to computational drug discovery. This work, led by the Freire group, stems from the universal approach of balancing losses in entropy with gains in enthalpy. The novelty is both the focus on enthalpic improvements, and using isothermal titration calorimetry as a tool by which to separately measure the enthalpic and entropic contributions to affinity (Luque and Freire, 2002; Freire, 2009; Ladbury et al., 2010).

Affinity is improved with larger losses in free energy, such that either gains in entropy or loss in enthalpy could drive a reaction. Improvements in one (entropy or enthalpy) can overcome deleterious effects on the other. Meanwhile, scientists traditionally measure only affinity (K_d) or inhibition (IC50, EC50, K_i). These are one dimensional measures of binding strength, which are highly useful, but it can be difficult to interpret the correspondence of ligand structural changes to whole affinity differences. By dissecting the contributions of enthalpy and entropy to the gains or loss in affinity, one can see how the changes are made; ligand changes that effect conformational freedom represent entropic changes, while improved interactions and fit are enthalpic (Freire, 2009). By separating the measurement of these effects in bench studies, the optimization process gets direct logical feedback. A change to the ligand designed to improve enthalpic contributions might have much more severe entropic consequences than anticipated. Attempts to gain affinity driven by entropy might not make a significant change because of constraining the protein for an entropic loss. Without the separation of analyses afforded by calorimetry, the lack of improved affinity might be misinterpreted as enthalpic losses, which would misdirect further attempts at optimization. Thus the relatively simple concept of separating affinity measures into enthalpic and entropic contributions through isothermal titration calorimetry enables feedback for straight forward rational design (Freire, 2009; Ladbury et al., 2010).

Decrease of conformational restrictions in the protein or ligand correspond to favorable entropic changes. Entropy estimations are useful to interrelate affinities between different ligands (affinity ranking). However, it has been argued that optimization efforts are better spent on improving the enthalpy of binding (Freire, 2008). Considerations for design include that every added hydrogen bond has both enthalpy of desolvation and of binding, and that each 1.4 kcal/mol of enthalpy change drives the reaction thermodynamically by an order of magnitude. These considerations are so important that Freire has suggested that binding enthalpy should be measured by isothermal titration calorimetry every time a new hydrogen bond donor or acceptor is considered (Freire, 2008).

Separate measures of enthalpy and entropy can enable better estimates of both contributions (Luque and Freire, 2002; Freire, 2009; Kawasaki et al., 2010; Ladbury et al., 2010), but what should go into the enthalpy calculation? Many types of enthalpic contributions are understood and well approximated. Details such as the contribution of hydrogen bonds are modeled by comparing the docked donor-acceptor distance to the ideal distance for proton sharing, in the context of the similar interactions available in the solvent. Binding enthalpy was estimated for 25 ligands in 7 proteins within a standard error of 0.4 kcal/mol, by supplementing estimates of conformational enthalpy change, with estimations of changes in solvent accessibility for solvent molecules in shells up to 5-7Å away from the ligand, and a correction for protonation (Luque and Freire, 2002). Modeling changes in enthalpy across different ligands may therefore be possible, and useful for estimating affinity rank.

The contribution of space filling to enthalpy had not advanced substantially since the shape complementarity analysis of Lennard-Jones (1924). The Freire group recently presented a study on how filling an empty protein cavity affects enthalpy (Kawasaki et al., 2010). For the example of filling clefts in the binding pocket of HIV-1 protease, a pattern of effects emerged across a limited spectrum of moiety size. When the cavity was not completely filled by the ligand moiety van der Waals forces gave benefit for enthalpy but at the cost of entropy. When the moiety was enlarged, the protein accommodated more optimal filling of the cavity space, adjusting around the ligand to reach a more enthalpically favored conformation. Entropy increased, driving the reaction. The interactions enabled by optimal space filling may have allowed the protein to stably go through pivot motions around this region, such that stabilizing interactions at the ligand interface allow other areas of the protein to be more flexible and thus the reaction becomes entropically favored. There is an apparent overstretching point at which the ligand pushes the protein into a more strained set of conformations, which penalize by both entropy and enthalpy. Thus proper filling of the space can add entropic and enthalpic driving force to binding (Kawasaki et al., 2010).

Through the analysis provided by the Freire group in the past decade, we have gained the ability to dissect very basic contributions of designed ligand moieties. Bench isothermal titration calorimetry analysis enables specific feedback to improve our estimates of entropy and enthalpy, and inform changes for computational design. This combination of a relatively simple but highly accurate bench technique with computational modeling is an emerging tool which can carry us forward to the next generation of drug discovery.

6.f. False hits

The concept of false hits was demonstrated elegantly by the recent work of the Shoichet group, in showing that hit compounds can inhibit protein activity by pathological mechanisms (Babaoglu et al., 2008). The "false hit" inhibitory mechanisms of beta-lactamase inhibitors discovered by high throughput techniques include many aggregators, covalent bonders, and promiscuous inhibitors. Poignantly, none of the 1,274 initial hits were found to be specific reversible inhibitors, which are pharmacologically desirable. Meanwhile, two of sixteen computationally derived hits were specific reversible micromolar inhibitors (Table 1). Thus, the approach of computational screens are bolstered by the fact that they model bioactivity in an explicitly physiological manner, whereas wet lab systems model the physical interaction and therefore can get side tracked by irrelevant behavior - a behavior which could be highly dangerous to the host (Babaoglu et al., 2008)!

6.g. Finding targets of known inhibition

Many drugs have no known mechanism. For many more drugs, the mechanistic basis of side effects are not understood. Mechanisms are the deep understanding of an interaction which enable improved design and analogy to less understood cases. They let us understand the exceptions, such as variable response.

Target elucidation allows us to understand clinical paired disease patterns. Based on observations that the opportunistic pathogen CMV is cleared from AIDS patients undergoing antiretroviral therapy, one might anticipate the nonspecific mechanism of HIV-1 inhibition allowing return of immunity and nonspecific clearing of CMV (Deayton et al., 1999). However, CD4 T-lymphocyte counts do not correlate with clearance (Reed and Morse, 1998). Our docking study predicts that amprenavir and indinavir target the CMV protease specifically (Jenwitheesuk & Samudrala, 2005a). Our group presented a similar descriptive prediction for HIV-1 inhibition by the common antibiotic minocycline being through HIV-1 integrase (Jenwitheesuk & Samudrala, 2007).

As well, we can understand the interrelation of bioactive compounds (metabolites and drugs) and the relevant proteome through the network of overlapping target-ligand interactions. A recent tour de force was applied to predict the interactions of all drugs

to the human proteome. The resulting network is a road map for polypharmacologic effects - leads and suggestions for caution (Keiser et al., 2009).

In the case for which the activity profile of a vast drug library is known for a particular pathogen, which is becoming more common, analysis of similarly active compounds can facilitate understanding of the targetable aspects of the pathogen. Targets can be selected based on the profile of activity across the library (Plouffe et al., 2008; Table 1). Depending upon the clinical indication, a target may be selected for uniqueness of library activity relative to the host and commensal organisms, or perhaps for similarity to targets of other diseases to maximize the chances of discovery of an existing drug multitargeting the disease of interest.

6.h. Personalized pharmacology

As the accuracy for models of protein-ligand interactions improves, along with it comes the ability to personalize these predictions. In models of individual susceptibility versus resistance, or predictions of disease progression, differences in genotype have already been modeled with high accuracy. The most common difference relevant to this problem is the nonsynonymous single nucleotide mutation or polymorphism. The change of one or more residues by mutation alters specific contacts to increase or decrease affinity, thereby making the mutant organism susceptible or resistant, respectively.

Our group designed a sequence analysis tool to predict the significance for this type of mutation (Horst et al., 2010), but much work remains. Our group also created a group of tools to take a patient's HIV-1 protease and reverse transcriptase sequence mutations and predict the profile of resistance versus susceptibility to the commonly used antiretroviral medications (Jenwitheesuk and Samudrala, 2003; Wang et al., 2004; Jenwitheesuk et al., 2004; Jenwitheesuk and Samudrala, 2005b), and integrated them into a freely available web server that uses the consensus of the structural and logistic regression techniques to select the optimal drug for HIV-1 patients (this web server has handled over 1,000 separate queries; <http://protinfo.compbio.washington.edu/pirspred>; Jenwitheesuk et al., 2005).

Other personalization includes screening for untoward side effects, such as inhibition of CYP450 proteins or monamine oxidases. As well, we differ not only in our human genotype, but that of our symbiotic bacteria. Personalized pharmacology may one day include identifying *E. coli* strain by genotyping stool samples, so an antibiotic regimen can be selected that will not cause imbalance to one's enteral flora.

Finally, dosage can be prescribed based on models of enteral uptake using the genes that code for microvilli intercellular junctions, and models of metabolism based on the CYP450 genes, and immunogenicity by the antibodies of memory T-cells and mast cells. As well, dosage can be prescribed by gene copy number variant, and relative susceptibility.

6.i. Open source drug discovery.

Through the development of robust, free, and publicly available computational methods for drug discovery we can increase efficiency and decrease costs for researchers and institutions involved in drug discovery worldwide. Computational methods have demonstrated the ability to greatly reduce the cost of hit and lead compound discovery (Becker et al., 2006; Jenwitheesuk et al., 2008; Babaoglu et al., 2008; Orti et al., 2009; Costin et al., 2010; Table 1). Therefore they have the potential to enable the development and distribution of drugs to combat diseases that disproportionately affect impoverished nations (also known as tropical or third world diseases), such as malaria and dengue fever. Since tropical diseases mostly affect the poor, the historical perspective has been that there is little to no incentive for pharmaceutical companies to invest in the development of these drugs. Nonetheless it should be noted that some of the largest Pharma companies have recently devoted massive resources to join the fight against Malaria, including Novartis (Plouffe et al., 2008) and GlaxoSmithKline (Gamo et al., 2010; Table 1). As well, from the standpoint of computational methodology, in head-to-head comparisons the best performing computational methods for drug discovery are not freely available nor publicly funded software (Michino et al., 2009; Plewczynski et al., 2010). Reasons for a partial shift to open publication and application of resources to minimally profitable diseases are intriguing, but beyond the scope of this text; for now these are the exceptions rather than the rule. The importance of reducing drug development costs through computation is unwavering.

Although many existing tools used in drug discovery are freely available, the skills necessary to use them and interpret the output typically requires a large amount of knowledge, which comprises an obstacle to wide spread use. It is rare even for medical scientists capable of performing animal studies and clinicians capable of performing clinical trials to possess the necessary knowledge to use computational predictive methods. In response to these barriers, a trend to release the identity of predicted compound-target interactions has emerged amongst publicly funded computational research groups (Jenwitheesuk and Samudrala, 2003b; Desai et al., 2004; Jenwitheesuk and Samudrala, 2005a; Desai et al., 2006; Jenwitheesuk and Samudrala, 2007; Xie et al., 2007; Jenwitheesuk et al., 2008; Keiser et al., 2009; Orti et al., 2009; Costin et al., 2010; Table 1). Moreover, the trend has been to share the outcome for initial experiments amongst these leads publicly. For example over the past decade our group has been committed to making all of our software, ideas, and data freely available to advance the science, and to release our predicted hit compounds in a way that maximizes impact and availability.

In addition to making all data publicly available, it would be useful to develop an easily accessible public web sever usable by non-scientists and scientists alike to expedite communication of knowledge to advance the discovery of novel drugs. Using a web server could be as simple as uploading the structure or sequence of a single target protein or set of related target proteins. A comprehensive analysis of the target(s) would predict inhibitors and substrates of the target(s). Antitargets with the potential to interact with each of the lead compounds could also be identified and presented to the user. Potential compounds tested for activity against the target(s) would come from a library of existing bioactive small molecules. When available experimental data such as ADME, bioavailability, or binding affinity could be stored for each compound and presented to the user in a standardized way. A few open source drug discovery projects have begun to address these goals to promote the discovery and development of novel therapies to neglected diseases (Jenwitheesuk et al., 2008; Orti et al., 2009).

6.j. Multitarget design

While we have argued that the search for a compound with a desired activity can be expedited by evaluating multitargeting compounds, we have not yet elaborated the principle of a single compound multitargeting a single disease. This relates to the off-target properties discussed above, by the concept that natural compounds and known drugs are more likely to be multitargeting. Here we extend the assertion that compounds can be selected to target other proteins in the same disease. This concept was perhaps first formalized by Erlich, who described a magic bullet that would inhibit cancer by multiple mechanisms (Ehrlich, 1911). One such example is Gleevec (a.k.a. imatinib, STI-571), which serendipitously targets both BCR-Abl and c-Abl, inhibiting the two principle known causes of cell proliferation in chronic myelogenous leukemia (CML; Kaelin, 2004). Gleevec has been the most widely used treatment for CML for 8 years.

The most effective drugs in humans (e.g. aspirin, Gleevec) inevitably interact with and bind to multiple proteins, a feature that traditional models based on single target drugs fail to take into account. Yet there is substantial evidence that these multitarget compounds have a higher incidence of untoward side effects than single target compounds (Peters et al., 2009). The multitarget approach is necessary because every drug has to be effective at its site of action (for example, HIV-1 protease inhibitors have to bind and inhibit the protease molecule) and has to be readily metabolized by the body (for example the cytochrome P450 enzymes, which are responsible for metabolizing the majority of drugs). Computational screening for multitarget binding and inhibition is effective because it exploits the evolutionary fact that protein structure is conserved much more in nature than is function or sequence.

It is ironic and surprising that reduced affinity sometimes corresponds to higher efficacy. This appears to be due to "weak linkage" of multiple target proteins within a particular physiologic network (Rogawski, 2000). Low affinity multitarget drugs may perturb networks more efficiently than high-affinity, single-hit drugs (Csermely et al., 2005). Simultaneous effects on multiple targets can

decrease the therapeutic dose, so that untoward side effects can be handled by lower doses; simply, a compound with three targets of similar affinity will be effective at one third the tissue concentration. The effects of salicylates on multiple proinflammatory signals exemplify that multiple mechanisms causing homeostatic imbalance can be targeted by a single drug; the low effective dose facilitated by multitargeting has made aspirin one of the most popular drugs in the world (Huang, 2002).

Pathogens and cancers develop resistance to single drug therapy. Inhibitor resistance is largely overcome in the multitargeting approach by the exponentially decreased probability of resistant mutations simultaneously arising in genes encoding proteins corresponding to all targets. The multitarget approach can be extended to incorporate the variability of target proteins across a disease pathogen population (Freire, 2002).

Computational predictions are obviously not perfect. The accuracy of recent docking with dynamics and structure activity relationship predictions contain less than 50% true positive hits at best (table 1). However, if one compound is predicted to hit multiple targets, the odds increase for actually inhibiting at least one target. Thus we have taken the approach to target as many essential proteins as there are crystal structures for a pathogen or disease (Jenwitheesuk et al., 2008). The complexity of possible multitarget effects indicate that occasionally it may be relevant to test in whole disease organism screens or even animal models of disease before evaluating which of the predicted multitarget interactions actually occur physically.

6.k. Multidisease screens and reversing the disease-drug search

Old Western movies keep alive the iconography of "cure alls" popularized into nostalgia by traveling salesman of the mid 19th century. These tinctures were meant to solve any medical problem, or at least a group of quite unrelated problems. In this chapter we share some examples of single drugs that combat multiple diseases. We also preach the repurposing of existing drugs, exploration of natural compounds, and the use of chemical derivatives of each; i.e. we continue with the concept of exploiting existing bioavailable, nontoxic, nonimmunogenic, multitargeting compounds. So it would be logical to test the ability of all these compounds to target any and all disease targets.

Given the limited set of compounds we propose to be used, the chance of finding a target for one particular disease might not be great, but with contemporary methods the chance of finding a disease for a particular drug is extremely probable. Multidisease screens can find the "opportunities" that do exist; the screening process can drive the drug-disease selection, rather than the disease (tradition). This concept represents a reversal of the conceptual framework underlying drug discovery, wherein we "play to our strengths." At each point of the modeling process we rely on the best scoring instances from the scoring functions. While somewhat ambiguous instances arise for all methods, scoring functions make it easy to know when the models are of little or great utility. Thus if we scan for instances for which the accuracy estimates indicate useful models, rather than searching for the best model for one's pet project, we may truly access those diseases, targets, and compounds which are most realistically modeled with existing computational methods.

Obviously computational drug discovery methods work in some cases. Obviously computational drug discovery methods do not work in all cases. One approach to solving this problem is to improve the methods; while that process goes on, should we not also work to find the cases for which the methods work? One captivating feature of this paradigm shift is that it minimizes the need for improved ranking functions, which as discussed above is the part of drug modeling in which the field has made the least progress.

The Sali group recently presented a project in which they let the available pharmacopeia (FDA approved compounds in DrugBank) be the driving force to choosing the organism and protein to target. Specifically, they started with ten disease associated genomes, modeled as many of the proteins as feasible with template based modeling, predicted protein-ligand matches by ligand mapping from template proteins, analogized protein-ligand matches to protein-drug matches by QSAR analysis, and finally ran four protein-drug pairs that appeared promising and relevant; three of the four demonstrate specific reversible binding (Orti et al., 2009; Table 1). In abstraction, the project used only the best scoring predictions of full modeling on a widely cast net. While the analysis already done in this work may hold other therapeutically relevant hits or leads, it already represents evidence that bolstering computational predictions over many possible targets can be expected to be productive if the decisions are made by the scoring functions.

7. Summary

Incurable or untreatable diseases comprise a salient group of applications for computational drug discovery. Etiologies for incurable diseases include pathogens (e.g. acquired immunodeficiency syndrome, ebola, polio, human papilloma virus), neoplasms (i.e. cancers), genetic abnormalities (e.g. Down, Creutzfeldt-Jakob, and Proteus syndromes), autoimmunity (e.g. lupus erythematosus, asthma, multiple sclerosis), and inappropriate response to environment (e.g. prions, type II diabetes mellitus). Of those for which treatment exists, therapy manages symptoms but does not remove recurrence of disease upon ceasing treatment (e.g. treatment of AIDS). Many life threatening diseases have no treatment whatsoever. The motivation for computational approaches to drug discovery is to spur the bench and clinical studies to find cures for all diseases and alleviate human suffering. Amidst these great successes in pharmacologic discovery, it is important to consider that cures exist for many chronic and opportunistic diseases in the form of proper preventive behaviors (e.g. diet, exercise, hygiene), for which psychology is perhaps a more relevant solution than pharmacology.

The opportunity addressed by computational techniques is to abstract the knowledge from the many instances of physiologic interactions chronicled over the past century, to the clinical situations that plague humanity. The links that allow these abstractions are the genetic code, which helps us to find the most relevant instances, and the structural models which help us predict how the interactions will occur.

Our research group, the groups of Shoichet, Freire, Becker, Avery, Sali, and others have demonstrated the early maturity of computational modeling of protein-ligand interactions by predicting compounds for desired pharmacologic activity and testing them in prospective experiments. These methods not only save time and resources but are beginning to be more accurate than in vitro screening methods (Doman et al., 2002; Jenwitheesuk et al., 2008; Babaoglu et al., 2008; Table 1).

POST SCRIPT

Note 1: It should be noted that authors including ourselves often discuss the goal of drug discovery only in the context of inhibitors. However, pharmacologic activators are desired, particularly for nonpathogenic ailments such as depression and pain, so all discussions of pharmacologic inhibitors here and elsewhere should be understood to be generalized to all pharmacologically active substances. Meanwhile, depending upon the target it may be more difficult to design an activator (agonist) or inhibitor (antagonist); for example the types of contacts and similarity to the physiologic substrate may be exploited differently by each.

References

Abdi MH, Beswick PJ, Billinton A, Chambers LJ, Charlton A, Collins SD, Collis KL, Dean DK, Fonfria E, Gleave RJ, Lejeune CL, Livermore DG, Medhurst SJ, Michel AD, Moses AP, Page L, Patel S, Roman SA, Senger S, Slingsby B, Steadman JG, Stevens AJ, Walter DS (2010). Discovery and structure-activity relationships of a series of pyroglutamic acid amide antagonists of the P2X7 receptor. *Bioorg Med Chem Lett*. 20:5080-5084.

Abdel-Rahman HM, Kimura T, Hidaka K, Kiso A, Nezami A, Freire E, Hayashi Y, Kiso Y (2004). Design of inhibitors against HIV, HTLV-I, and Plasmodium falciparum aspartic proteases. *Biol Chem*. 2004 Nov;385(11):1035-9.

Adams C, Brantner V (2006). "Estimating the cost of new drug development: is it really 802 million dollars?". *Health Aff (Millwood)* **25** (2): 420–8.

Allen FH (2002). The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr B*. 58(3):380–388.

Babaoglu K, Simeonov A, Irwin JJ, Nelson ME, Feng B, Thomas CJ, Cancian L, Costi MP, Maltby DA, Jadhav A, Inglese J, Austin CP, Shoichet BK (2008). Comprehensive mechanistic analysis of hits from high-throughput and docking screens against beta-lactamase. *J Med Chem*. 51(8):2502-11.

Becker OM, Marantz Y, Shacham S, Inbal B, Heifetz A, Kalid O, Bar-Haim S, Warshaviak D, Fichman M, Noiman S. (2004) G protein-coupled receptors: in silico drug discovery in 3D. *Proc Natl Acad Sci USA*. **101** (31):11304-9.

Bellmann-Sickert K, Beck-Sickinger AG (2010). Peptide drugs to target G protein-coupled receptors. *Trends Pharmacol Sci*. ePub 22Jul2010.

Berezin C, Glaser F, Rosenberg Y, Paz I, Pupko T, Fariselli P, Casadio R, and Ben-Tal N. 2004. ConSeq: The Identification of Functionally and Structurally Important Residues in Protein Sequences. *Bioinformatics*. 20:1322-1324.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Research 28: 235-242.

Bernard B, Samudrala R (2009). A generalized knowledge-based discriminatory function for biomolecular interactions. *Proteins*. 76:115-128.

Borisy AA, Elliott PJ, Hurst NW, Lee MS, Lehar J, Price ER, Serbedzija G, Zimmermann GR, Foley MA, Stockwell BR, Keith CT (2003). Systematic discovery of multicomponent therapeutics. Proc Natl Acad Sci USA. 100:7977–7982.

Cammarata A, Menon GK (1976). Pattern recognition. Classification of therapeutic agents according to pharmacophores. *J Med Chem*. 19(6):739-48.

Campbell J, Campbell J, David R (2005). An insight into the practice of pharmacy in ancient Egypt. *Pharm Hist*. 35(4):62-68.

Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ (2005). The Amber biomolecular simulation programs. *J Comput Chem*. 26(16):1668-88.

Chen BY, Honig B (2010). VASP: A Volumetric Analysis of Surface Properties Yields Insights into Protein-Ligand Binding Specificity. *PLoS Comput Biol*. 6. e1000881.

Chong CR, Chen X, Shi L, Liu JO, Sullivan DJ Jr. (2006) A clinical drug library screen identifies astemizole as an antimalarial agent. *Nat. Chem. Biol*. 2, 415–416.

Csermely P, Agoston V, Pongor S (2005). The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci*. 26:187-182.

Costin JM, Jenwitheesuk E, Lok S-M, Hunsperger E, Conrads KA, Fontaine KA, Rees CR, Rossmann MG, Isern S, Samudrala R, Michael SF (2010). Structural optimization and de novo design of dengue virus entry inhibitory peptides. *PLoS Negl Trop Dis*. 4:e721.

de Paulis T, Reinhard JF Jr, Oshana S, Kauffman M, Donahue S (2007). Tolerability, pharmacokinetics, and neuroendocrine effects of PRX-00023, a novel 5-HT1A agonist, in healthy subjects. *J Clin Pharmacol*. 47: 817–824.

de Vos P (2010). European materia medica in historical texts: Longevity of a tradition and implications for future use. *J Ethnopharmacol*. 2010 Jun 16.

Dancík V, Seiler KP, Young DW, Schreiber SL, Clemons PA (2010). Distinct biological network properties between the targets of natural products and disease genes. J Am Chem Soc. 132(27):9259-61.

Das S, Kokardekar A, Breneman CM. 2009. Rapid Comparison of Protein Binding Site Surfaces with Property Encoded Shape Distributions. *J Chem Inf Model*. 49:2863-72.

Deayton J, Mocroft A, Wilson P, Emery VC, Johnson MA, Griffiths PD (1999). Loss of cytomegalovirus (CMV) viraemia following highly active antiretroviral therapy in the absence of specific anti-CMV therapy. *AIDS*. 13(10):1203-1206.

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

Ekins S (2005). Systems-ADME/Tox: resources and network approaches. *J Pharmacol Toxicol Methods*. 53(1):38-66.

Ehrlich P (1910). Die Behandlung der Syphilis mit dem Ehrlichschen Präparat 606. *Deutsche medizinische Wochenschrift*. 1893–1896.

Ehrlich P (1911). Aus Theorie und Praxis der Chemotherapie. *Folia Serologica*. 7,697–714.

Ewing TJ, Makino S, Skillman AG, Kuntz ID (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput-Aided Mol Des*. 15:411-428.

Fan H, Irwin JJ, Webb BM, Klebe G, Shoichet BK, Sali A (2009). Molecular docking screens using comparative models of proteins. *J Chem Inf Model*. 49(11):2512-2527.

Fischer JD, Mayer CE, Söding J (2008). Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*. 24(5):613-620.

Fleming A (1929). On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. *Br J Exp Pathol*. 10:226-236.

Fleming A, Voureka A, Kramer IR, Hughes WH (1950). The morphology and motility of Proteus vulgaris and other organisms cultured in the presence of penicillin. *J Gen Microbiol*. 4, 257–269.

Freire E (2002). Designing drugs against heterogeneous targets. *Nat Biotechnol*. 20(1):15-6.

Freire E (2008). Do enthalpy and entropy distinguish first in class from best in class? *Drug Discov Today*. 13(19-20):869-74.

Freire E (2009). A thermodynamic approach to the affinity optimization of drug candidates. *Chem Biol Drug Des*. 74(5):468-72.

Gamo FJ, Sanz LM, Vidal J, de Cozar C, Alvarez E, Lavandera JL, Vanderwall DE, Green DV, Kumar V, Hasan S, Brown JR, Peishoff CE, Cardon LR, Garcia-Bustos JF (2010). Thousands of chemical starting points for antimalarial lead identification. *Nature*. 465(7296):305-310.

Gold ND, Jackson RM (2006). SitesBase: Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J Mol Biol*. 355:1112-1124.

Greaves R, Warwicker J (2005). Active site identification through geometry-based and sequence profile-based calculations: burial of catalytic clefts. *J Mol Biol*. 349(3):547-557.

Gutteridge A, Thornton JM (2005). Understanding nature's catalytic toolkit. *Trends Biochem Sci*. 30:622–629.

Grzybowski BA, Ishchenko AV, Kim CY, Topalov G, Chapman R, Christianson DW, Whitesides GM, Shakhnovich EI (2002). Combinatorial computational method gives new picomolar ligands for a known enzyme. *Proc Natl Acad Sci USA*. 99(3):1270-3

Hansch C (1969). Quantitative approach to biochemical structure-activity relationships. *Acc Chem Res*. 2(8):232–239.

Hidvegi T, Ewing M, Hale P, Dippold C, Beckett C, Kemp C, Maurice N, Mukherjee A, Goldbach C, Watkins S, Michalopoulos G, Perlmutter DH (2010). An autophagy-enhancing drug promotes degradation of mutant alpha1-antitrypsin Z and reduces hepatic fibrosis. *Science*. 329(5988):229-32.

Hirschel B, Francioli P (1998). Progress and problems in the fight against AIDS. *N Engl J Med*. 338(13):906-8.

Hopkins AL (2008). Network pharmacology: the next paradigm in drug discovery. *Nature Chem Bio*. 4:682-690.

Horst J, Samudrala R (2009). Diversity of protein structures and difficulties in fold recognition: the curious case of protein G. *F1000 Biol Rep*. 1:69.

Horst JA, Samudrala R (2010). A protein sequence meta-functional signature for calcium binding residue prediction. *Pattern Recognit Lett*. 31(14):2103-2112.

Horst JA, Wang K, Horst OV, Cunningham ML, Samudrala R (2010). Disease risk of missense mutations using structural inference from predicted function. *Curr Protein Pept Sci*. 11(7):500-515.

Hróbjartsson A, Gøtzsche PC (2001). Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *N Engl J Med*. 344(21):1594–1602.

Hróbjartsson A, Gøtzsche PC (2010). Placebo interventions for all clinical conditions. *Cochrane Database Syst Rev*. 2010(1):CD003974.

Huang S (2002). Rational drug discovery: what can we learn from regulatory networks? *Drug Discov Today*. 7:S163-S169.

Huff T (2003). The Rise of Early Modern Science: Islam, China, and the West, Cambridge University Press.

Jensen RA (1976). Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*. 30:409-25.

Jenwitheesuk E, Horst JA, Rivas KL, Van Voorhis WC, Samudrala R (2008). Novel paradigms for drug discovery: computational multitarget screening. *Trends Pharmacol Sci*. 29(2):62-71.

Jenwitheesuk E, Samudrala R (2007). Identification of potential HIV-1 targets of minocycline. *Bioinformatics*. 23(20):2797-2799

Jenwitheesuk E, Samudrala R (2005). Identification of potential multitarget antimalarial drugs. *JAMA*. 294(12):1490-1491.

Jenwitheesuk E, Wang K, Mittler JE, Samudrala R (2005). PIRSpred: a web server for reliable HIV-1 protein-inhibitor resistance/susceptibility prediction. *Trends Microbiol*. 13(4):150-151.

Jenwitheesuk E, Samudrala R (2005). Virtual screening of HIV-1 protease inhibitors against human cytomegalovirus protease using docking and molecular dynamics. *AIDS*. 19(5):529-531.

Jenwitheesuk E, Wang K, Mittler JE, Samudrala R (2004). Improved accuracy of HIV-1 genotypic susceptibility interpretation using a consensus approach. *AIDS*. 18(13):1858-1859.

Wang K, Jenwitheesuk E, Samudrala R, Mittler JE (2004). Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance. *Antivir Ther*. 9(3):343-352.

Jenwitheesuk E, Samudrala R (2003). Identifying inhibitors of the SARS coronavirus proteinase. *Bioorg Med Chem Lett*. 13(22):3989-3992.

Jenwitheesuk E, Samudrala R (2003). Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations. *BMC Struct Biol*. 3:2.

Kaelin WG Jr (2004) Gleevec: prototype or outlier? *Sci STKE*. 225:pe12.

Kawasaki Y, Chufan EE, Lafont V, Hidaka K, Kiso Y, Mario Amzel L, Freire E (2010). How much binding affinity can be gained by filling a cavity? *Chem Biol Drug Des*. 75(2):143-51.

Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KL, Edwards DD, Shoichet BK, Roth BL (2009). Predicting new molecular targets for known drugs. *Nature*. 462:175-181.

Kellera TH, Pichotaa A, Yina Z. 2006. A practical view of 'druggability'. *Curr Opin Chem Biol*. 10(4): 357-361.

Khersonsky O, Roodveldt C, Tawfik DS (2006). Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr Opin Chem Biol*. 10(5):498-508.

Kitchen DB, Decornez H, Furr JR, Bajorath J (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov*. 3:935-949.

Koshland DE (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci USA*. 44(2):98–104.

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. (1982). A geometric approach to macromolecular-ligand interactions. *J Mol Biol*. 161:269-288.

Ladbury JE, Klebe G, Freire E (2010). Adding calorimetric data to decision making in lead discovery: a hot tip. *Nat Rev Drug Discov*. 9(1):23-7.

Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res*. 33:W299-W302.

Lennard-Jones JE (1924). On the Determination of Molecular Fields. *Proc R Soc Lond A*. 106(738):463–477.

Liu T, Horst JA, Samudrala R (2009). A novel method for predicting and using distance constraints of high accuracy for refining protein structure prediction. *Proteins*. 77:220-234.

Lloyd NC, Morgan HW, Nicholson BK, Ronimus RS (2005). The composition of Ehrlich's salvarsan: resolution of a century-old debate. *Angew Chem Int Ed Engl*. 44(6):941-4.

Lopez G, Ezkurdia I, Tress ML (2009). Assessment of ligand binding residue predictions in CASP8. *Proteins*. 77(supp9):138-146.

Luque I, Freire E (2002). Structural parameterization of the binding enthalpy of small ligands. *Proteins*. 49(2):181-90.

Manniche L (1989). An Ancient Egyption Herbal. British Museum Publications Ltd., London.

Michino M, Abola E; GPCR Dock 2008 participants, Brooks CL 3rd, Dixon JS, Moult J, Stevens RC (2009). Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. Nat Rev Drug Discov. 8(6):455-463.

Moult J (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol*. 15(3):285-289.

Natoli TA, Smith LA, Rogers KA, Wang B, Komarnitsky S, Budman Y, Belenky A, Bukanov NO, Dackowski WR, Husson H, Russo RJ, Shayman JA, Ledbetter SR, Leonard JP, Ibraghimov-Beskrovnaya O (2010). Inhibition of glucosylceramide accumulation results in effective blockade of polycystic kidney disease in mouse models. *Nat Med*. 16(7):788-792.

Newman DJ, Cragg GM (2007). Natural products as sources of new drugs over the last 25 years. *J Nat Prod*. 70(3):461–477.

Nezami A, Laque I, Kimura T, Kiso Y, Freire E. (2002). Identification and characterization of allophenylnorstatine-based inhibitors of plasmepsin II, an antimalarial target. *Biochemistry*. 41:2273–2280.

Nezami A, Kimura T, Hidaka K, Kiso A, Liu J, Kiso Y, Goldberg DE, Freire E (2003). High-affinity inhibition of a family of Plasmodium falciparum proteases by a designed adaptive inhibitor. *Biochemistry*. 42(28):8459-64.

Ngan S-C, Inouye M, Samudrala R (2006). A knowledge-based scoring function based on residue triplets for protein structure prediction. *Protein Eng Des Sel*. 19:187-193.

Ortí L, Carbajo RJ, Pieper U, Eswar N, Maurer SM, Rai AK, Taylor G, Todd MH, Pineda-Lucena A, Sali A, Marti-Renom MA (2009). A kernel for open source drug discovery in tropical diseases. *PLoS Negl Trop Dis*. 3(4):e418.

Peters JU, Schnider P, Mattei P, Kansy M (2009). Pharmacological promiscuity: dependence on compound properties and target specificity in a set of recent Roche compounds. *ChemMedChem*. 4(4):680-686.

Plewczynski D, Laźniewski M, Augustyniak R, Ginalski K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J Comput Chem*. 2010 Sep 1.

Plouffe D, Brinker A, McNamara C, Henson K, Kato N, Kuhen K, Nagle A, Adrián F, Matzen JT, Anderson P, Nam TG, Gray NS, Chatterjee A, Janes J, Yan SF, Trager R, Caldwell JS, Schultz PG, Zhou Y, Winzeler EA (2008). In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc Natl Acad Sci USA*. 105:9059-64.

Ponder JW, Case DA (2003). Force fields for protein simulations. *Adv Protein Chem*. 66:27-85.

Reed JB, Morse LS (1998). CMV retinitis and the controversies associated with highly active antiretroviral therapy and the immune recovery hypothesis. *AIDS Patient Care STDS*. 12(3):181-185.

Rickels K, Mathew S, Banov MD, Zimbroff DL, Oshana S, Parsons EC Jr, Donahue SR, Kauffman M, Iyer GR, Reinhard JF Jr. (2008). Effects of PRX-00023, a novel, selective serotonin 1A receptor agonist on measures of anxiety and depression in generalized anxiety disorder: results of a double-blind, placebo-controlled trial. *J Clin Psychopharmacol*. 28: 235–239.

Rogawski MA (2000). Low affinity channel blocking (uncompetitive) NMDA receptor antagonists as therapeutic agents – towards an understanding of their favorable tolerability. *Amino Acids*. 19, 133–149.

Roy A, Kucukural A, Zhang Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*. 5:725-738.

* Sadreyev RI, Grishin NV (2006). Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds. *BMC Struct Biol*. 6:6.

Schneider G, Fechner U (2005). Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov*. 4: 649–63.

Strebhardt K, Ullrich A (2008). Paul Ehrlich's magic bullet concept: 100 years of progress. *Nat Rev Cancer*. 8(6):473-80.

Wang K, Samudrala R (2006). Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*. 7: 385.

Weill N, Rognan D (2009). Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. J Chem Inf Model 49(4):1049-62.

Weisman JL, Liou AP, Shelat AA, Cohen FE, Guy RK, DeRisi JL. (2006) Searching for new antimalarial therapeutics amongst known drugs. *Chem. Biol. Drug Des*. 67, 409–416.

Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S Jr, Weiner PJ (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc*. 106:765-784.

Xie L, Bourne PE (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comp Biol*. 1: e31.

Xie L, Wang J, Bourne, PE (2007). In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comp Biol*. 3:e217.

Zhang Y, Devries ME, Skolnick J. (2006) Structure modeling of all identified G protein-coupled receptors in the human genome. PLoS Comput Biol. 2(2):e13.

Zhang Y (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*. 9:40.