Grace Tang
Final Paper

Computational Drug Design and Small Molecule Library Design

Computational Drug Design

In the past decade, significant investments in drug development have not translated into a parallel growth in new drugs [1]. Increasing knowledge and understanding of disease have expanded the number of therapeutic protein targets. However, few drug candidates targeting them reach the market as many fail Phase II clinical trials due to insufficient efficacy [2]. Computational drug design has emerged to harness different sources of information to facilitate the development of new drugs that modulate the behavior of therapeutically interesting protein targets. Ligand-based methods use existing knowledge of active compounds against the target to predict new chemical entities that likely possess similar behavior. In contrast, structure-based methods rely on target structural information to determine whether a new compound is likely to bind and interact. Increasingly popular are integrated methods, which combine a target's ligand information and structural information for the design of new drugs. These different methods are summarized in Figure 1.

Virtual Screening

A specific emphasis within computational drug design is virtual screening. Virtual screening is the process of screening through small molecule libraries for a subset of compounds enriched for interacting with a therapeutic protein target of interest. Its utility predominantly lies in hit and lead-compound identification [3] and is complementary to experimental high throughput screening (HTS). High throughput screening involves robotically assaying many thousands to a few million compounds for binding or function against a protein target. Such experiments are expensive and cannot exhaustively sample chemical space [4], problems virtual screening are

less sensitive too. Ligand-based methods, structure-based methods, and integrated approaches are all largely applicable for virtual screening.

Ligand-based Approaches

Ligand-based approaches are one of the main classes for computer-aided drug design. It is built on the concept of molecular similarity, where compounds with high structural similarity are more likely to have similar activity profiles [5]. Given a protein target of interest, ligand-based approaches are most appropriate when bioactive molecules are available but structural knowledge of the target is unavailable. The number of known active compounds determines the specific approach used to aggregate this information to build a model of suitable ligands for the target. A single known active molecule can be used to screen a small molecule library for similar compounds. Comparison of the active molecule against the library is often performed via fingerprint-based similarity searching where the molecules are represented as bit strings, indicating the presence/absence of predefined structural descriptors [6]. Such methods are popular due to their speed, which is achieved by pre-calculating and storing the fingerprints for the small molecule library.

Multiple known active molecules can be structurally superimposed and overlapping elements extracted to create a two-dimensional or three-dimensional pharmacophore. Pharmacophores contain the spatial constraints between features thought to be important for interacting with the target protein. Common features include hydrogen bond donors and acceptors, positively and negatively charged groups, hydrophobic regions, and aromatic rings. Pharmacophores are subsequently used to screen small molecule libraries for compounds satisfying the constraints

and thus likely to be active against the protein target of interest. Building the pharmacophore can be difficult when large structural differences exist between the bioactive compounds as the correct superpositioning of the molecules is difficult to determine [7]. Superpositioning of the bioactive molecules in three dimensions is also more difficult as there are many degrees of freedom.

A combination of known active and inactive compounds against a protein target permits the building of more complex models using machine learning. A variety of machine learning techniques including decision trees, neural networks, support vector machines, and ensemble methods have been applied to drug design, each with their strengths and weaknesses [8]. These methods are relatively data hungry as many active and inactive compounds must be used to train and subsequently test the models. These models capture the properties discriminating active molecules from inactive molecules in order assess novel small molecules for their likelihood of interacting with the target of interest.

Additionally, one of the most popular ligand-based drug design approaches is quantitative structure-activity relationships (QSAR). The aim of QSAR is to determine the relationship between structural/physicochemical properties of active compounds to their biological activity. Knowledge of the activity levels of the compounds, such as binding affinity (K_D) or inhibitory concentration (IC50), is requisite for QSAR methods. Both pharmacophores as well as machine learning derived models are suitable for QSAR studies [9]. Here, pharmacophores indicate the method of small molecule representation and machine learning correlates the pharmacophoric features with quantitative activity levels.

Structure-based Approaches

Structure-based approaches to drug design require structural knowledge of the protein target. The Protein Data Bank is the most commonly used repository of high-resolution X-ray crystallography structures [10]. In the absence of experimentally derived structures, homology models built from related proteins have also been shown to be successful [11]. Structure-based methods require no *a priori* knowledge of active ligands, permitting their application to novel target proteins about which little is known. As a result, structure-based approaches frequently contribute to the development of new drugs through the discovery and optimization of the initial lead compound [12]. Structure-based methods have already critically contributed to the development of approximately ten drugs [12].

Structure-based drug design approaches are often referred to as docking-based methods. Docking programs serve three main purposes. First, docking programs identify potential ligands from a library of chemical compounds (virtual screening). Next, they predict the binding mode of potential ligands or known ligands. Finally, using the predicted binding pose, these programs calculate a putative binding affinity. Docking programs have shown success in screening large chemical libraries, reducing them to a more manageable subset that is enriched for binders. In cases of true interactions, the predicted ligand pose often correlates well with experimentally solved protein-ligand complexes. While structure-based methods have led to the identification of novel drugs, binding pose prediction is considered its strength [13]. Binding affinity prediction and even the simpler task of rank ordering ligands by affinity have not been successfully with inconsistent results across diverse protein targets and drug classes [13].

As ligand-based drug design approaches are sensitive to the number and types of known active and inactive ligands, structure-based drug design methods are sensitive to the number and diversity of experimental structures of the target protein. A single snapshot of a target protein does not capture its dynamic and flexible nature. Frequently, protein-ligand interactions involve structural rearrangements, both major and minor [14]. Differences between the bound and unbound protein, or active and inactive protein, can result from induced fit or conformational selection or both [15-17]. This issue of conformational flexibility is crucial in ligand binding and readily evident for the important drug target family of G protein-coupled receptors (GPCRs) [18]. Docking into structures reflecting inactive conformations are unlikely to yield agonists and vice versa. Only recently, have increased structural examples of GPCRs permitted enhanced understanding of receptor activation and thereby better informed structure-based drug design.

Integrated Approaches

With the rise bioactive ligand information and protein target structures, there has been a shift towards integrated protein and ligand structural data in computational drug design. At its simplest, building a three-dimensional pharmacophore to find potential ligands and performing a subsequent docking study on the hits constitutes a combined approach. However, more advance techniques integrating the two sources of information into a single method also exist. These integrates approaches fall into two classes: interaction-based and docking similarity-based methods (Figure 1).

Interaction-based methods focus on identifying the key interactions between the protein and ligand using available physicochemical data. These interactions are then used to screen small

molecule libraries for compounds capable of producing such an interaction profile. Representation of the interaction information divides interaction-based methods into pseudoreceptor methods and pharmacophore/fingerprint models (Figure 1). Pseudoreceptor techniques capture the key interactions by modeling the ligand binding site [19]. Different ligands that bind a common protein target must be structurally overlaid in order for conserved protein-ligand interactions to be determined. Overlaying of the small molecule structures can be difficult as ligands can be highly flexible. Correct superpositioning of the ligands is essential for meaningful protein models to be built. Additional difficulty comes in the form of ligand diversity. Only interactions existing in the known ligands can be modeled, meaning diverse ligands with a broad range of interactions are necessary to capture all the key interaction elements of the binding site. This may pose a problem for detecting ligands interacting via a novel binding mode.

In contrast to pseudoreceptor methods, which focus on modeling the ligand-binding site, pharmacophore/fingerprint models capture the key interactions by modeling the ligand using pharmacophores or fingerprints [19]. These models differ from ligand-based pharmacophores and fingerprints in that the features are extracted based on the protein-ligand interactions rather than the ligands alone. A key strength of pharmacophore/fingerprint models is that their simplicity readily lends themselves to similarity-based methods for screening chemical compound libraries. Problems with this method include the number and types of features to include as this determines the types of information captured by the models.

The second major class of integrated approaches is docking similarity-based methods, which merge structure-based docking methods with ligand similarity methods [19]. These approaches subdivide into screening-based and scoring-based (Figure 1). The screening-based methods use ligand similarity to focus the screening on ligands similar to known active compounds for the target. Increased efficiency in screening is important as small molecule libraries can be on the order of millions of compounds. The scoring-based methods integrate ligand information into the scoring function. Traditional scoring functions used in structure-based docking methods evaluate interactions between the protein and ligand. Scoring-based integrated approaches such as Maximum Volume Overlap [20] use how well a compound binding pose overlaps with known ligand binding poses to score. This permits compound poses similar in volume and charge distribution to known ligands to score well. While such methods can improve the ability to find native-like binding orientations, they may perform poorly if the binding mode differs from those of known ligands.

Small Molecule Library Design

Small molecule library design is an important issue in drug design and virtual screening. High quality collections are essential for effective screening studies. The key issue in library design is the balance between library size and library diversity. Two contrasting approaches have therefore emerged for library design: diversity-oriented synthesis (DOS) and fragment-based screening (FBS) [21].

Diversity-oriented Synthesis

Diversity-oriented synthesis aims to fill chemical space and thus promotes production of compounds not found in existing libraries. Such compounds are usual similar in size to those of drug-like compounds and therefore are capable of high affinity and potency. These properties render them easier to detect both in virtual and experimental screens. Additionally, by nature, libraries from diversity-oriented synthesis often cover chemical space previously ignored by scientists, drug developers, and nature [22]. Coverage of such space can be critical when searching for therapeutics against novel targets. Furthermore, a key strength of DOS libraries is its utility in experimental drug discovery against less understood diseases where appropriate protein targets are unknown. Such screening studies may be performed against diseased cells where compounds from DOS libraries have sufficient potency and diversity to yield interesting and effective hits.

However, library design by diversity-oriented synthesis has its weaknesses. Chemical space is extremely large and resulting compound libraries from DOS are therefore large, with millions of compounds. While application to virtual screening may be reasonable, high throughput screening of such libraries is costly, if possible. Huge investments must also be made to synthesize novel compounds from DOS when little knowledge about their activity against relevant targets is known. Additionally, many small molecules from DOS are not considered drug-like based on their physicochemical properties. This is a major concern, as a lack of drug-like properties is a leading cause of poor drug selectivity and attrition [23]. If such molecules are isolated by screening techniques as potential bioactive compounds, further optimization may be necessary to develop analogues with improved behavior such as selectivity and bioavailability.

Fragment-based Screening

In contrast, fragment-based screening aims to represent chemical diversity through fragments. It has been suggested that a few thousand compounds can sufficiently capture the diversity encompassed by tens of millions larger drug-like compounds [24]. Fragment-based libraries are smaller than diversity-oriented synthesis libraries and are therefore efficient and cost effective to screen. Fragments forming high quality interactions with the target protein are pieced together to form larger, more potent and more drug-like lead compounds. Two considerations affect library design in FBS. First, libraries are often created in a context specific fashion by taking into consideration the targeted proteins. Second, fragments are selected to have acceptable physicochemical properties based on the physicochemical properties of successful drugs. The careful inclusion of "good" fragments in the library and subsequent construction of lead compounds from these fragments allow the design of compact lead compounds with high ligand efficiency [25]. Already, drug candidates from fragment-based drug design are in clinical trials [26].

However, drug candidates and lead compounds from fragment-based screening tend to be flat. This is in contrast to the three dimensional nature of natural compounds and compounds from diversity-oriented synthesis. Flat molecules are unlikely to be as specific as three-dimensional molecules. Unlike DOS compounds, fragments are also significantly smaller in size and therefore have lower potency and affinity. This renders them more difficult to detect in virtual and experimental screens. Specifically in experimental screens against cells, fragments are not only too weak to be detected, but also lack specificity. Fragment-based screening is therefore not appropriate when precise protein targets are not known. Even when precise protein targets

are known, in general these proteins must be capable of being solubly expressed and compatible with nuclear magnetic resonance (NMR) spectroscopy or X-ray crystallography. These two techniques elucidate the binding location and binding pose of fragments, information often used in expanding fragments into lead compounds.

Conclusion

In recent years, computational drug design and more specifically virtual screening, has emerged as a powerful tool in drug discovery. There are numerous approaches to drug design depending on the types of information available about the bioactive ligands and the therapeutic protein target. These include ligand-based methods, structure-based methods, and integrated approaches. Regardless of approach, a common goal is to identify new compounds capable of modulating a protein target's activity. As such, screening of small molecule compound libraries both virtually and experimentally is performed. The library quality plays a key role in screening results as poor libraries produce poor hits. The two approaches to library design focus on the two competing aspects of library design: library diversity and library size. Diversity-oriented synthesis focuses on the former and fragment-based screening focuses on the latter. They each have their strengths and weaknesses and have seen success in the pharmaceutical industry. Protein-protein interaction interfaces are an area of drug discovery traditionally dominated by protein-based therapeutics, such as antibodies. Recent breakthroughs using DOS libraries [27] and FBS libraries [28] have led to penetration of these targets.

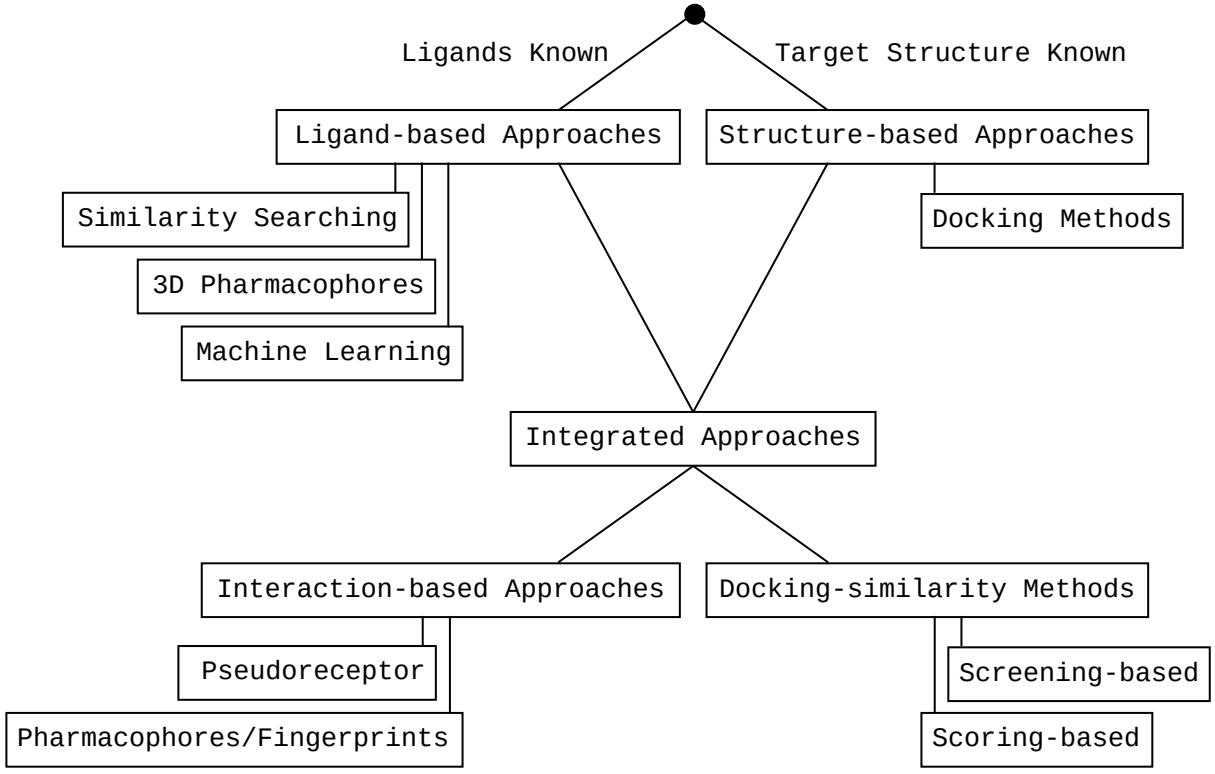

Figure 1. Computational Drug Design. Classification tree of drug design methods

References

1. Pammolli, F., L. Magazzini, and M. Riccaboni, *The productivity crisis in pharmaceutical R&D.* Nat Rev Drug Discov, 2011. **10**(6): p. 428-38.
2. Arrowsmith, J., *Trial watch: Phase II failures: 2008-2010.* Nat Rev Drug Discov, 2011. **10**(5): p. 328-9.
3. Schneider, G., *Virtual screening: an endless staircase?* Nat Rev Drug Discov, 2010. **9**(4): p. 273-6.
4. Blum, L.C. and J.L. Reymond, *970 million druglike small molecules for virtual screening in the chemical universe database GDB-13.* J Am Chem Soc, 2009. **131**(25): p. 8732-3.
5. Martin, Y.C., J.L. Kofron, and L.M. Traphagen, *Do structurally similar molecules have similar biological activity?* J Med Chem, 2002. **45**(19): p. 4350-8.
6. Mishra, V. and C.V. Siva Prasad, *Ligand based virtual screening to find novel inhibitors against plant toxin Ricin by using the ZINC database.* Bioinformation, 2011. **7**(2): p. 46-51.
7. Yamaotsu, N. and S. Hirono, *3D-pharmacophore identification for kappa-opioid agonists using ligand-based drug-design techniques.* Top Curr Chem, 2011. **299**: p. 277-307.
8. Plewczynski, D., S.A. Spieser, and U. Koch, *Performance of machine learning methods for ligand-based virtual screening.* Comb Chem High Throughput Screen, 2009. **12**(4): p. 358-68.
9. Gozalbes, R., et al., *Development and validation of a pharmacophore-based QSAR model for the prediction of CNS activity.* ChemMedChem, 2009. **4**(2): p. 204-9.
10. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

11. Ghosh, A., et al., *Structural insights into human GPCR protein OA1: a computational perspective.* J Mol Model, 2011.
12. Kolb, P., et al., *Docking and chemoinformatic screens for new ligands and targets.* Curr Opin Biotechnol, 2009. **20**(4): p. 429-36.
13. Leach, A.R., B.K. Shoichet, and C.E. Peishoff, *Prediction of protein-ligand interactions. Docking and scoring: successes and gaps.* J Med Chem, 2006. **49**(20): p. 5851-5.
14. Lin, J.H., *Accommodating protein flexibility for structure-based drug design.* Curr Top Med Chem, 2011. **11**(2): p. 171-8.
15. Hansen, G., et al., *Unexpected active-site flexibility in the structure of human neutrophil elastase in complex with a new dihydropyrimidone inhibitor.* J Mol Biol, 2011. **409**(5): p. 681-91.
16. Aleksandrov, A. and T. Simonson, *Molecular dynamics simulations show that conformational selection governs the binding preferences of imatinib for several tyrosine kinases.* J Biol Chem, 2010. **285**(18): p. 13807-15.
17. Silva, D.A., et al., *A role for both conformational selection and induced fit in ligand binding by the LAO protein.* PLoS Comput Biol, 2011. **7**(5): p. e1002054.
18. Salon, J.A., D.T. Lodowski, and K. Palczewski, *The significance of G protein-coupled receptor crystallography for drug discovery.* Pharmacol Rev, 2011. **63**(4): p. 901-37.
19. Wilson, G.L. and M.A. Lill, *Integrating structure-based and ligand-based approaches for computational drug design.* Future Med Chem, 2011. **3**(6): p. 735-50.
20. Fukunishi, Y. and H. Nakamura, *Prediction of protein-ligand complex structure by docking software guided by other complex structures.* J Mol Graph Model, 2008. **26**(6): p. 1030-3.
21. Hajduk, P.J., W.R. Galloway, and D.R. Spring, *Drug discovery: A question of library design.* Nature, 2011. **470**(7332): p. 42-3.
22. Galloway, W.R., A. Isidro-Llobet, and D.R. Spring, *Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules.* Nat Commun, 2010. **1**: p. 80.
23. Leeson, P.D. and B. Springthorpe, *The influence of drug-like concepts on decision-making in medicinal chemistry.* Nat Rev Drug Discov, 2007. **6**(11): p. 881-90.
24. Hajduk, P.J., R.P. Meadows, and S.W. Fesik, *Discovering high-affinity ligands for proteins.* Science, 1997. **278**(5337): p. 497,499.
25. Carr, R.A., et al., *Fragment-based lead discovery: leads by design.* Drug Discov Today, 2005. **10**(14): p. 987-92.
26. Hajduk, P.J. and J. Greer, *A decade of fragment-based drug design: strategic advances and lessons learned.* Nat Rev Drug Discov, 2007. **6**(3): p. 211-9.
27. Di Micco, S., et al., *Identification of lead compounds as antagonists of protein Bcl-xL with a diversity-oriented multidisciplinary approach.* J Med Chem, 2009. **52**(23): p. 7856-67.
28. Murray, C.W. and T.L. Blundell, *Structural biology in fragment-based drug design.* Curr Opin Struct Biol, 2010. **20**(4): p. 497-507.