Towards structure-based protein drug design

Changsheng Zhang* and Luhua Lai*†1

1Beijing National Laboratory for Molecular Science, State Key Laboratory for Structural Chemistry of Unstable and Stable Species, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China; and †Center for Theoretical Biology, Peking University, Beijing 100871, China

Abstract
Structure-based drug design for chemical molecules has been widely used in drug discovery in the last 30 years. Many successful applications have been reported, especially in the field of virtual screening based on molecular docking. Recently, there has been much progress in fragment-based as well as de novo drug discovery. As many protein–protein interactions can be used as key targets for drug design, one of the solutions is to design protein drugs based directly on the protein complexes or the target structure. Compared with protein–ligand interactions, protein–protein interactions are more complicated and present more challenges for design. Over the last decade, both sampling efficiency and scoring accuracy of protein–protein docking have increased significantly. We have developed several strategies for structure-based protein drug design. A grafting strategy for key interaction residues has been developed and successfully applied in designing erythropoietin receptor-binding proteins. Similarly to small-molecule design, we also tested de novo protein-binder design and a virtual screen of protein binders using protein–protein docking calculations. In comparison with the development of structure-based small-molecule drug design, we believe that structure-based protein drug design has come of age.

Protein drugs: from native, engineered to rationally designed

Since the first recombinant protein therapeutic insulin was used as a life-saving treatment for patients with Type 1 diabetes mellitus in 1982, more than 130 therapeutic proteins have been approved for clinical use by the U.S. Food and Drug Administration [1]. Six therapeutic proteins were among the top 20 in the U.S. 2008 drug sales and it has recently been estimated that half of the top 100 selling ethical pharmaceuticals will be therapeutic proteins by 2014 [2]. One reason for the dramatically increased number of protein drugs is that interactions involving proteins are usually specific, and protein drugs cause fewer side effects compared with small-molecule drugs [3]. Another reason is the rapid increment of the knowledge about the disease-related pathogen–host protein interaction and protein–protein interaction network in human body, and interfering or modulating protein–protein interaction has attracted more and more attentions [4,5]. Protein–protein interfaces are usually large and flat, therefore, for most cases, it is difficult to use small molecules to bind directly to a protein–protein interface with high affinity and specificity [5].

Most protein drugs are the recombinant native partners of various target proteins [1–3]. Structure-based methods are often used to design mutations to improve binding affinity or stability, or minimize protein aggregation [3,6]. Monoclonal antibodies are a class of remarkably successful protein drugs against a wide range of target molecules [7,8]. Monoclonal antibodies have the same invariant Y-like scaffold and variant conformational determination regions which recognize various targets. However, antibodies have large sizes and complicated compositions, which results in high production costs: they require manufacturing in eukaryotic expression systems, usually involving stably transfected mammalian cell lines [9,10]. The continuous improvement of in vitro selection technologies, such as phage display, ribosome display and yeast display, promotes the new generation of protein drugs using engineered protein scaffolds [9]. In general, a robust and small monomeric protein is selected as a new scaffold, whose flexible surface sites can be engineered, and the in vitro selection techniques are then used to screen the engineered library. Several scaffolds, such as DARPin [10], Anticalin [11], Monobody and Affibody, are being developed with encouraging success.

Constructing and screening a large library is laborious and costly. This therefore raises the question of whether we can design protein drugs rationally. The number of protein complex structures solved is steadily growing. By exploring these structures, some basic principles about protein–protein interaction have been discovered [12]. Computational design of novel proteins that bind with a certain target based on these principles can not only cut down the cost, but also help to discover more reasonable scaffolds and mutants with high binding affinity, high stability and less aggregation. Despite a few inspiring successful attempts [13–18], computational protein drug design is still in the early stage. In comparison, structure-based small-molecule design has been developed.
for over 30 years, and many successful applications have been reported. In the present paper, we compare computational protein drug design with small-molecule drug design to see what we can learn from the existing strategies and what the differences are between the two fields.

Small-molecule drug design strategies

In the last 30 years, many chemical compounds have been designed based on target structures, including enzymes, receptors or other types of proteins, to interrupt or regulate their activity. Rationally designed small-molecule drugs have been successful in the treatment of almost all types of disease [19].

Target structure-based virtual screening, using protein-ligand docking to identify potential lead candidates, is a fundamental and the most successful approach for structure-based drug design [19–22]. In 1982, Kuntz et al. [23] developed the earliest protein–ligand docking program DOCK. The first trials of drug design by structure-based strategies was also carried out by Kuntz [22] using the DOCK program. In recent years, this field has become highly diversified and is evolving rapidly. Less expensive high-performance computing platforms are available to screen large chemical databases.

By analysing a set of active ligands for a target in a specific binding site, some three-dimensional arrangement of features are often found to be crucial for ligand molecules in order to interact with the target [24]. These features are defined as pharmacophores. Pharmacophore models can be used in compound modification, and also in pharmacophore-based virtual screening, which has been proved to be an effective method for in silico drug discovery [25].

Compared with virtual screening, computational de novo drug design can produce novel molecular entities without structural limitations [26]. Fragments from a fragment library are used to construct a ligand in a target binding pocket by a linking or growing strategy. The prominent problem for the de novo design strategy is that synthesis of the de novo designed molecules is often difficult or laborious.

Protein drug design by protein–protein docking

Protein–protein docking is the procedure of constructing possible complex models from the free structures of two proteins [27,28]. Several protein–protein interface de novo design trials started from protein–protein docking to obtain an initial complex model, then followed by computational sequence design or experimental screening of a designed library [13–16]. Huang et al. [13] designed de novo a heterodimer of the 56-amino-acid β1 domain of streptococcal protein G, which was optimized from a dimeric model of the wild-type protein generated by a fast Fourier transform-based docking program. In the first stage of the computational protocol suggested by Jha et al. [14], for redesigning scaffold proteins to bind to a specified region on a target protein, a low-resolution rigid-body docking was performed. Using this protocol, the human hyperplastic disc protein was redesigned to bind to the kinase domain of p21-activated kinase 1. Guntas et al. [15] designed a variant of the ubiquitin ligase E6AP (E6-associated protein) that bound to a non-natural partner, Ubc12, also using a docking-design computational protocol [15]. A directed library was constructed according to the results of computational design, and the tight binders were obtained by screening the library. Karanicolas et al. [16] designed de novo a pair of proteins: Prb and Pdar. A tightly buried hydrophobic core was designed based on the initial docking model. Then directed evolution was used to increase the binding affinity and create a complex with subnanomolar affinity.

All of these design cases are focused on a particular scaffold protein. Taking the lesson from the target structure-based virtual screen strategy in small-molecule drug discovery, we may also use the protein–protein docking approach to virtually screen a protein library to look for the best scaffolds, with certain binding affinity, as ‘drug leads’. This can be done because: first, thousands of protein structures deposited in the PDB can now be collected to compose the protein library; and, secondly, the protein synthesis and purification technologies have been greatly improved and the cost has been greatly reduced. Of course, in order to do this, the most important thing is that the protein–protein docking programs used should be able to efficiently screen a large protein structure library and enrich the possible binders with a relatively high success rate. Since flexible protein–protein docking is often very time-consuming, for the protein drug-screening step, docking programs that perform global rigid docking should be used. In 1992, the fast Fourier transform-based rigid global protein–protein docking program FTDOCK was developed by Katchalski-Katzir et al. [29]. In this earliest protein–protein docking program, the scoring function just contained geometric complementary. In the following two decades, electrostatic interaction, hydrophobic interaction, statistical contact potentials (for example, in ZDOCK [30] and even force-field potentials (in SDOCK [31]) are introduced to the scoring function of the rigid docking programs, and the reliability and accuracy for predicting protein–protein complex structures have been greatly improved. Recent recent protein–protein rigid docking programs may be ready to be used for protein drug screening.

Enrichment testing for the protein–protein docking program SDOCK

To test whether the positive proteins can be enriched according to the best score of protein–protein docking results, we built test cases in which both the native binding proteins and the decoy proteins are collected. Three target proteins were selected: ubiquitin, CDC42 and Ran. Unbound structures were used as the target protein structures. The
Table 1 | Enrichment testing using SDOCK

<table>
<thead>
<tr>
<th>Target</th>
<th>Native binding proteins (PDB codes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ubiquitin (1UBQ)</td>
<td>1CMX, 1S1Q, 1WRF, 1L6M, 2C7M, 2HDS, 2HTH, 2J7Q, 200B, 2QHO, 2XBB, 3A3A, 3Y4B, 3K90, 3065 and 3PRF</td>
</tr>
<tr>
<td>CDC42 (2GR2)</td>
<td>1D0A, 1GRN, 1GZS, 1K1T, 1NF3, 2WM9, 3EG5 and 3N3V</td>
</tr>
<tr>
<td>Ran (3GJ0)</td>
<td>1A2K, 1L2M, 1IBR and 1K5D</td>
</tr>
</tbody>
</table>

Table 2 | Enrichment factor by SDOCK in the three cases

<table>
<thead>
<tr>
<th>x%</th>
<th>Ubiquitin</th>
<th>CDC42</th>
<th>Ran</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2.0</td>
<td>3.8</td>
<td>5.0</td>
</tr>
<tr>
<td>20</td>
<td>2.3</td>
<td>3.1</td>
<td>3.8</td>
</tr>
</tbody>
</table>

complex structure of a native binding protein and the corresponding target protein need to have been determined by X-ray crystallography at a resolution better than 2.8 Å (1 Å = 0.1 nm). The names of the native binding proteins and the PDB codes of the complexes are listed in Table 1. The bound structure of the native binding protein was used for docking. A total of 116 decoy proteins were selected from the PDB (see the Supplementary Online Data at http://www.biochemsocstrans.org/bst/039/bst0391382add.htm for the list) by the criteria that the protein should have 100–150 amino acids and exist as a monomer in solution. The crystal structure should have a resolution better than 1.8 Å and have no ligand bound. Proteins with 30% or higher sequence identity are removed. The native binding proteins and the decoy proteins are docked to the corresponding target protein using the docking program SDOCK which was a global protein–protein docking program using stepwise force-field potentials. All of the positive and negative proteins for a certain target were ranked according to their best docking score. The enrichment factor has been defined by Bender and Glen [32] for measures of enrichment in virtual screening. If the total number of native binding is \( N_{native}^{all} \) and \( N_{native}^{x\%} \), native binding proteins can be found in the top x% of all docked proteins, then the enrichment factor is:

\[
E_f = \frac{N_{native}^{x\%}}{N_{native}^{all}} \times x\% 
\]

The enrichment factors of the three test cases, when x was set to 10 and 20, are shown in Table 2. The results indicate that the protein–protein docking has the ability to enrich positive proteins.

We also tested the docking-screen strategy experimentally for TNFα (tumour necrosis factor α) [33] inhibitor design project (B. Tang, C. Zhang and L. Lai, unpublished work). A library containing 677 protein structures was constructed and screened using SDOCK. Three proteins among the 22 selected had confirmed binding to TNFα.

Hotspot grafting as a new method for protein drug design

On protein–protein interfaces, typically only several residues called the hotspots dominate binding energy, and other residues make them tightly buried [34]. A promising strategy for designing protein drugs targeting a certain target is searching for appropriate scaffolds from a library to accommodate these key residues and perfectly bury these residues on the interfaces. Hotspots on protein drugs can be compared with pharmacophores in small-molecule drugs. Besides the mutagenesis methods, hotspots on a protein–protein interface can be identified by analysing the complex structures [35,36]. A pharmacophoric pattern matching approach, such as graph theory-based approach or set reduction algorithm, can be used in detecting sites that can accommodate hotspot patterns [37]. An algorithm for hotspots grafting has been developed in our laboratory [37] and an updated one in which the backbone flexibility was considered was also developed recently (C. Zhang and L. Lai, unpublished work). The hotspot pattern was fixed to a scaffold; and then the scaffold was docked to the target protein according to the hotspot superposition; high-scoring complexes were selected for testing. Using this algorithm, a scaffold protein, the pleckstrin homology domain of phospholipase C, was searched for and successfully designed to bind the human EPO (erythropoietin) receptor by grafting the key interaction residues of human EPO on to this scaffold protein [17]. Another strategy to design protein drugs with predetermined hotspot patterns was suggested by Fleishman et al. [18]. In their strategy, a scaffold protein was first docked to the target protein to find favourable configurations with good shape complementary, and then the hotspot patterns, which were mainly from a known antibody, were used to anchor these configurations [18]. Using this computational strategy, two proteins were successfully designed that targeted the conserved stem region of influenza haemagglutinin. These two examples are inspiring for rational protein drug design and more hotspot grafting-based designs are expected.

Protein drug de novo design with completely novel sequences

In small-molecule drug discovery, in order to discover completely novel compounds instead of modifying existing scaffolds, de novo design strategies have been developed. Similarly, can we design novel protein sequences that can bind to the target instead of redesigning certain scaffolds? Obviously, de novo design is the ultimate goal for both small-molecule drug discovery and protein drug discovery, because, with little restriction, excellent drugs meeting all requirements have more possibility of being discovered. An advantage of de novo protein drug design compared with de novo small-molecule drug design is that obtaining proteins with the designed sequences is usually easier than synthesizing a novel compound. However, de novo protein
drug design faces other challenges: the folding problem, the binding problem, and the aggregation problem. To de novo design compounds, fragment growing and linking are the most commonly adopted strategy. Coincidentally, fragment-assemble-based approaches, for example, the algorithms in ROSETTA, have also been developed for protein structure prediction and de novo design of protein structures [38]. Using ROSETTA algorithm, the novel globular protein fold Top7 was designed [39] and the structures of variable regions in homologous proteins can be accurately modelled [40]. In the structurally variable region modelling problem, both the self-folding and interaction with the conserved region should be considered. Thus, we can take examples from this prediction problem to de novo protein drug design. Recently, we have tried to use ROSETTA to design small proteins that can bind to TNFα and found encouraging results (Q. Sheng, C. Zhang and L. Lai, unpublished work). We believe that the de novo protein drug design strategy will become practical in the near future.

**Perspectives and challenges**

For the three structure-based protein drug design strategies proposed above (summarized in Figure 1), although all of them face challenges, they can be applicable in different protein design cases and for different purposes. If the structure of a native binding protein and target protein complex is available and the hotspots on the interface are known, then the hotspot grafting strategy would be a good choice. The hotspots can also be computationally de novo designed rather than mimic the native protein-binding patterns. However, the designed hotspots may not be as dependable as native hotspots. The docking-screen method is used for discovery of native drug leads for further optimization, but the reliability of protein–protein docking should be improved and the scoring strategy for screening should also be studied further to obtain a better enrichment performance. Designing protein drugs with de novo sequences is a challenging task. Because the protein folding problem is far from solved, small proteins (or peptides) with simple folded structures can be used in de novo design for the current stage.

**Conclusions**

We have suggested and tested three strategies for structure-based protein drug design. Along with the rapid development of computational and experimental techniques, most obstacles in both computational and experimental aspects have been overcome and structure-based protein drug design is coming of age. Therapeutic protein drugs developed through structure-based design are expected to come in the near future.

**Funding**

This work was supported in part by the Ministry of Science and Technology of China and the National Natural Science Foundation of China.

**References**


Received 26 June 2011
doi:10.1042/BST0391382
SUPPLEMENTARY ONLINE DATA
Towards structure-based protein drug design

Changsheng Zhang* and Luhua Lai*†

*Beijing National Laboratory for Molecular Science, State Key Laboratory for Structural Chemistry of Unstable and Stable Species, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China, and †Center for Theoretical Biology, Peking University, Beijing 100871, China

PDB codes of decoy proteins in enrichment testing for protein–protein docking screen

1AAJ, 1BFG, 1BKR, 1EKG, 1EWF, 1GYY, 1H03, 1IB1, 1J27, 1JB3, 1KF5, 1KMT, 1L6P, 1LIT, 1LMI, 1LN4, 1LU4, 1LWB, 1NOA, 1O8X, 1SAU, 1SBX, 1SRV, 1STN, 1T3Y, 1TJ6, 1TP6, 1TQG, 1TZV, 1UXZ, 1W4I, 1WKA, 1WW1, 1X6Z, 1ZLB, 2B02, 2CIU, 2CW4, 2CWR, 2CXY, 2D4P, 2EHG, 2END, 2ESK, 2FB6, 2FC3, 2FRG, 2GKG, 2GZV, 2HLQ, 2HUJ, 2IOS, 2JAB, 2JLI, 2LIS, 2NWD, 2PND, 2PPO, 2QJZ, 2QPW, 2R48, 2RB8, 2RE2, 2RH3, 2V4X, 2V75, 2VQ4, 2W0I, 2WJ5, 2WWE, 2X5P, 2YXF, 2YXM, 2YZ1, 2ZG7, 2Z14, 3A7L, 3AG7, 3B79, 3BU9, 3BZT, 3CE7, 3CTG, 3CX2, 3D33, 3D95, 3EAO, 3EOI, 3EUR, 3UY6, 3FH2, 3FKC, 3FSO, 3GAX, 3HAK, 3HD4, 3HNX, 3I2V, 3I7M, 3IHS, 3IU5, 3IV4, 3JU0, 3JZZ, 3KP8, 3LAX, 3LD7, 3LS0, 3M9J, 3M9Q, 3NRW, 3OMD, 3ON1, 3ONH, 3PD7.

*To whom correspondence should be addressed (email lhla@pku.edu.cn).