Molecular docking as a popular tool in drug design, an in silico travel

Jerome de Ruyck
Guillaume Brysbaert
Ralf Blossey
Marc F Lensink
University Lille, CNRS UMR8576
UGSF, Lille, France

Abstract: New molecular modeling approaches, driven by rapidly improving computational platforms, have allowed many success stories for the use of computer-assisted drug design in the discovery of new mechanism- or structure-based drugs. In this overview, we highlight three aspects of the use of molecular docking. First, we discuss the combination of molecular and quantum mechanics to investigate an unusual enzymatic mechanism of a flavoprotein. Second, we present recent advances in anti-infectious agents’ synthesis driven by structural insights. At the end, we focus on larger biological complexes made by protein–protein interactions and discuss their relevance in drug design. This review provides information on how these large systems, even in the presence of the solvent, can be investigated with the outlook of drug discovery.

Keywords: structure-based drug design, protein–protein docking, quaternary structure prediction, residue interaction networks, RINs, water position

Introduction

Sequencing of the human genome has led to an increase in the number of new therapeutic targets for pharmaceutical research. In addition, high-throughput crystallography and nuclear magnetic resonance methods have been further developed and contributed to the acquisition of the atomic structures of proteins and protein–ligand complexes of an increasing level of detail.\(^1\) When the three-dimensional structure of the target, even from experiments or computing, exists, a frequently used technique to design inhibitor molecules is structure-based drug design (SBDD), which is depicted in Figure 1.

The most popular method in SBDD is molecular docking. Initially, docking – a term which was coined in the late 1970s – meant the refinement of a model of a complex structure by optimization of the separation between the partners, but with fixed relative orientations. Later, this relative orientation was allowed to vary, but the internal geometry of each of the partners was held fixed. This type of modeling is often being referred to as rigid docking.\(^2,3\) Currently, thanks to further increases in computational resources, it has become possible to model changes in internal geometry of the interacting partners that may occur when a complex is formed. This type of modeling is also known as flexible docking.

Moreover, computational modeling of the quaternary structure of complexes, formed by two or more molecular interaction partners, is nowadays also feasible. Examples are protein–protein complexes and complexes between proteins and nucleic acids.\(^4\)
In this review, we focus on modern usages of molecular docking. The first sections are dedicated to the design of new drug candidates starting from known crystal structures of crucial proteins. We then turn to protein-protein docking, including a discussion of the importance of water molecules in the docking procedure – how they are managed and, in the end, how they can influence binding probes. A list of the modeling programs discussed in this review is presented in Table 1.

Molecular docking and quantum mechanics, a combined mechanistic tool

A detailed understanding of the mechanisms of enzymes at atomic and electronic detail is of crucial importance in biomedical research. This would require solving the quantum mechanics (QM) of molecules, and thus, the computational costs of ab initio QM methods have limited their application. It is, for example, tenuous to elucidate a complete enzymatic mechanism, and therefore, methods have been devised to approximate the treatment. Several groups used combined approaches where calculations typically use a molecular mechanics (MM) force field for the system as a whole and apply an ab initio (QM) treatment to the site of interest. Using this QM/MM method, they were able to tackle different aspects of the biological systems studied such as electronic properties, interaction sites, or even conformational changes occurring in the protein active sites. Nowadays, a more advanced application of this approach is the ONIOM method. ONIOM stands for “our own N-layered integrated molecular orbital and molecular mechanics”. Originally developed by Morokuma in Dapprich et al and Svensson et al this computational technique models large molecules by defining more than two layers within the structure that are treated at different accuracy levels. By this way, the ONIOM method can treat relatively large molecules and can be applied in many areas of research and specifically organic and enzymatic reaction mechanisms. The modeling process involves two major steps: building the model and then mapping the enzymatic chemical process.

In order to highlight the importance of the very first step, we now describe the use of molecular docking to obtain a good starting point needed to elucidate the mechanistic pathway of an unusual flavoenzyme, isopentenyl diphosphate isomerase (IDI). Usually, flavoproteins are listed as redox catalysts, but in this specific case, there is no redox exchange observed. This enzyme catalyzes the isomerization of isopentenyl pyrophosphate into dimethylallyl pyrophosphate, which is the primary building block for all isoprenoid compounds. These are vital for every (micro-)organism, as these molecules are involved in every single cellular mechanism. Two types of IDIs have been reported. On the one hand, type I IDI (IDI-1) was discovered some decades ago. It has been extensively studied, and the mechanism was well established. On the other hand, type 2 IDI (IDI-2) was discovered more recently and is not so well characterized. This is a flavoprotein that requires reduced flavin mononucleotide cofactor and a divalent cation, but the mechanism is still vague and difficult to approach by experimental methods. As a presumption, it could be
a protonation from N$_2$ of reduced flavin mononucleotide leading to a carbocation, subsequently followed by a deprotonation step (Figure 2A).53,36 Thus, it is a critical enzyme for several classes of pathogenic microorganisms, and it is totally absent from human. Considering that novel chemotherapeutic strategies are urgently needed,57 new mechanism-based inhibitors of IDI-2 were sought. Recently, a structure-based approach was initiated for inhibitor development, since a high-resolution structure had been published.38 The goal was to investigate the putative mechanism by using QM/MM techniques. In this context, molecular docking provided the starting and ending points of the reaction path. As the apo-protein had been crystallized, the strategy was to dock both the substrate, isopentenyl pyrophosphate, and the product, dimethylallyl pyrophosphate, to the structure (Figure 2B). Then, protonation states were carefully inspected, and the energy of both the structures was minimized. Several other research groups are now using these methods to address the enzymatic mechanisms of a wide range of potential drug targets and to develop new mechanism-based inhibitors.39–44

### Modeling ligand–protein interactions in drug design

Ligand binding is the key step in enzymatic reactions and, thus, for their inhibition. Therefore, a detailed understanding of interactions between small molecules and proteins may form the basis for a rational drug design strategy.35–44 This approach was widely considered in order to design molecules addressing a broad range of major pathologies such as cancers49,50 or cardiovascular diseases.51–53

Another example, which is emphasized here, is the successful use of docking to design lead compounds as new anti-infectious agents against *Mycoplasma tuberculosis* or *Plasmodium falciparum*. These two pathogens are the key actors in the development of tuberculosis and malaria, respectively, which are the two major causes of mortality in developing countries.54 In order to target this scourge, several research teams have studied, for a long time now, the nonmevalonate isoprenoid biosynthesis pathway (2-methyl-D-erythritol-4-phosphate [MEP] pathway). Indeed, these parasites rely on this cascade to produce their own isoprenoid compounds, critical for their survival.55–57 The second step of the pathway is the reduction of 1-deoxy-d-xylulo-5-phosphate to MEP catalyzed by 1-deoxy-d-xylulose-5-phosphate reductoisomerase (DXR).58 In addition, humans and animals do not rely on the MEP pathway, making DXR an attractive target in the search for novel families of drugs. Currently, several inhibitors of DXR have been synthesized and evaluated.39

The purpose of this subsection is to present the usage of structural data in order to improve the efficiency of a new family of drugs. In the absence of crystallographic structures of DXR from *P. falciparum* (pf-DXR) or *M. tuberculosis*, molecular modeling, based on the structure of DXR from *Escherichia coli*,60 allowed several research groups to further elucidate the structure and function of the enzyme and also facilitated structure-based inhibitor design. Consequently, models of DXR from the pathogens were built61 and used to develop efficient screening methods in order to identify potential lead compounds.62 Later, these models were validated by X-ray crystallography.63,64

<table>
<thead>
<tr>
<th>Sections</th>
<th>Type of modeling</th>
<th>License</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD</td>
<td>Molecular docking and quantum mechanics, a combined mechanistic tool, Modeling ligand–protein interactions in drug design, Solvent effect as an important parameter</td>
<td>Protein–ligand</td>
<td>Commercial</td>
</tr>
<tr>
<td>Gaussian (ONIOM)</td>
<td>Molecular docking and quantum mechanics, a combined mechanistic tool</td>
<td>QM/MM</td>
<td>Commercial</td>
</tr>
<tr>
<td>AUTODOCK</td>
<td>Modeling ligand–protein interactions in drug design, Solvent effect as an important parameter</td>
<td>Protein–ligand</td>
<td>Open</td>
</tr>
<tr>
<td>GLIDE</td>
<td>Modeling ligand–protein interactions in drug design, Solvent effect as an important parameter</td>
<td>Protein–ligand</td>
<td>Commercial</td>
</tr>
<tr>
<td>RosettaDock</td>
<td>Biological complexes and quaternary structures, the protein–protein docking approach</td>
<td>Protein–protein</td>
<td>Open</td>
</tr>
<tr>
<td>pyDOCK</td>
<td>Biological complexes and quaternary structures, the protein–protein docking approach</td>
<td>Protein–protein</td>
<td>Open</td>
</tr>
<tr>
<td>AquaSol</td>
<td>Solvent effect as an important parameter</td>
<td>Solvent effect</td>
<td>Open</td>
</tr>
</tbody>
</table>

Abbreviations: ONIOM, our own N-layered integrated molecular orbital and molecular mechanics; QM, quantum mechanics; MM, molecular mechanics.
designed and synthesized. They were found to be highly potent inhibitors of pf-DXR, having $K_i$ values in the nanomolar range. Thus, these molecules were more active than fosmidomycin, the reference in the field of DXR inhibition (Figure 3). Recently, structure-guided design and virtual screening were successfully applied in order to identify and evaluate new molecules with a potent inhibitory effect on *P. falciparum*.

In light of these promising results, one can conclude that considerable progress has been made in the past, and the goal of obtaining clinically effective antimalarial drugs seems reachable. 70–77

**Biological complexes and quaternary structures, the protein–protein docking approach**

Protein–protein interactions (PPIs) play a central role in all biological processes. These processes result from the physical interaction of several protein molecules, thus forming the macromolecular assemblies that effectuate cellular function. Many large-scale studies focusing on PPI have emerged in recent years, using graphs with nodes and edges to represent the protein components interacting with each other. 78,79 Such binary representations capture a wealth of information but are inherently abstract and incomplete, since they contain no information as to time, place, or specificity. Such detailed information is indispensable for the guidance of mutagenesis studies or the design of inhibitor molecules.

Protein–protein docking actually predates protein–ligand (small molecule) docking, as the concept of protein docking introduced by Wodak and Janin 80 was later extended to the interaction between macromolecules and small ligands. 81 The treatment of flexibility in the binding process is considerably easier with small molecules, even though a considerable computational cost is involved, and small molecule docking has become one of the most active research areas in computational drug discovery. Most if not all of present-day protein–protein docking algorithms have been developed in light of the critical assessment of prediction of interactions (CAPRI) experiment, which is a community-wide collaboration that has accelerated the development of computational protein docking methods. 82 CAPRI, which is modeled after critical assessment of protein structure prediction, 83 organizes blind prediction trials; participants model their complex, and the models are assessed through comparison with an unknown crystal structure, 82 made available to the assessors on a confidential basis and prior to publication.

The early years have been essential for the development of docking algorithms, 84,85 with the incorporation of more elaborate scoring functions owing to efficient implementations of fast Fourier transform algorithms in docking 86–88 as one of the key advancements. This also spawned the CAPRI
scoring experiment, which was designed to help developers test scoring functions independently from docking calculations. In the scoring experiment, participants are given access to an enriched ensemble of docking models, contributed on a voluntary basis by participants in the docking experiment. The scorers select models from this ensemble and make a docking submission that is assessed using the standard CAPRI evaluation criteria. Although scorers are generally apt at discriminating near-native solutions from a selection of incorrect decoys, a correct ranking of these remains problematic. Benchmark data sets dedicated to scoring protein complexes have been developed recently that should facilitate the development of scoring functions.

During the last years, development was shifted to more realistic docking scenarios. CAPRI no longer offers so-called “bound” targets, where the structure of one or both of the partners is supplied in their bound conformation. Dockers nowadays routinely use unbound structures, that is, the structures of the binding partners as they occur in solution, and some degree of conformational change needs to be taken into account. Moreover, often only the sequence of the interacting partners is provided, and a step of homology modeling is required prior to docking. Although these matters significantly complicate the docking procedure, they represent the realistic scenarios that computational biologists nowadays are presented with.

The PPI interfaces have received increased attention during the past years. A prediction of the residues involved in the interaction interface may be used to guide the docking itself. Subsequent optimization of the interaction interface requires modification of side chain orientation; protein docking algorithms increasingly include flexibility treatments in their docking procedures, and more recent implementations favor the simultaneous docking of ensembles of unbound conformers. Together, the reliable prediction of interface residues and the incorporation of global and local flexibility in the docking algorithms provide invaluable information to inform mutagenesis studies and to steer drug design applications.

The scoring functions of docking programs have been successfully improved with additional descriptors based on residue interaction networks (RINs). RINs consist of networks generated from three-dimensional structures, where nodes correspond to residues and edges to detected interactions. RINs are small-world networks, and their topological analyses have been used in particular to study protein–protein interfaces and protein–ligand binding and to optimize scoring functions for the evaluation of docking poses. Using different approaches, it has been demonstrated that combining the network measures such as closeness centrality, betweenness centrality, degree, or clustering coefficient with energy terms can improve the ranking in scoring functions. Chang et al used two different types of RINs, a hydrophobic one and a hydrophilic one, for each complex, and then calculated a network-based score considering the average degrees and clustering coefficients. They developed a scoring method that enhanced discrimination of the scoring method of RosettaDock by more than 10% on a subset of protein–protein docking benchmark 2.0. Pons et al generated the RIN of each protein individually before docking and calculated four measures of the network, including closeness centrality. By integrating a score based on the closeness values into the pyDock scoring function, they improved by as much as 36% the top ten success rate on the protein–protein docking benchmark 4.0. Furthermore, residue centrality analyses as performed by del Sol et al, which are based on the average shortest path length, can also be used on docked poses to evaluate the central residues located in the interface. These residues could subsequently be targeted for mutagenesis experiments or drug design.

The docking predictions can be used in combination with homology-based methodologies and integrated into PPI networks to enhance these with structural information.
The Interactome3D\textsuperscript{120} web service incorporates structural data into PPI networks to improve them with interface information. These structural data either come from experiments or are modeled through a comparative modeling pipeline. Mosca et al\textsuperscript{120} illustrate the value of this tool by showing it allowed them to suggest a potential mechanism of action common to several disease-causing mutations. Indeed, they observed the mutations on structures of the complement cascade pathway involving, in particular, the complement component 3 (C3) and the component factor H (CFH) interaction. Several disease-causing mutations were located at the interface of proteins, and these key elements could be targeted by drugs in order to stabilize the C3–CFH interface. Thus, with this type of network, it is possible to contextualize mutations related to different diseases involved in a pathway and draw potential links between them. It can help to better define the target to aim for and, hence, improve the drug design.\textsuperscript{120} Docking predictions could then be additionally integrated to these networks. Furthermore, it was predicted that on average a drug binds to six different targets, including both the primary target and additional “off-targets”.\textsuperscript{121} Following this idea, reverse docking can be performed, which consists in the screening of one single molecule against multiple receptors instead of screening multiple small molecules against several receptors.\textsuperscript{122,123} Homology modeling may be useful to enrich the screening, when experimental structures are not available. The building of structural PPI networks may then be used in drug design to predict the targets the drug may bind to, with their related potential adverse drug reactions.\textsuperscript{122} They can help to identify which proteins would be affected by a drug designed to disrupt a particular interface because they highlight the domains that are involved in PPI. These structural PPI networks can also be exploited for drug repositioning, considering the use of known approved drugs or the reconsideration of late-stage failures.

**Solvent effect as an important parameter**

Proteins in solution are surrounded by water molecules. Water molecules around proteins organize in hydration shells that show correlated fluctuations.\textsuperscript{124} They are responsible for electrostatic screening\textsuperscript{25} and make important contributions to enzyme substrate recognition and catalysis and to molecular recognition in general.\textsuperscript{126,127} Considerable effort has been devoted to the modeling of water molecules in protein–ligand docking procedures, where the importance of water-mediated contacts has long been recognized. Well-known docking packages such as GOLD, AUTODOCK, or GLIDE can incorporate water molecules explicitly to predict protein–ligand docking poses.\textsuperscript{128–130} But very few methods exist that allow the prediction of hydration water positions at protein–protein interfaces.

The important contribution of water in the binding\textsuperscript{105} between proteins is readily realized when considering the high-affinity barnase–barstar complex.\textsuperscript{131} The extracellular ribonuclease barnase is always expressed with its inhibitor barstar in order to prevent the bacterium from degrading its own RNA. The complex is noted for its extremely tight binding, with a $k_{\text{on}}$ rate of $10^8$ M$^{-1}$ s$^{-1}$ and an affinity $k_{\text{d}}$ $=$ $10^{-14}$ M. The complex has been extensively studied both experimentally and computationally, explaining in detail its binding energetics.\textsuperscript{132} The binding interface, which is mainly composed of polar and charged residues, contains as many as 51 associated water molecules, of which no less than 18 are fully buried.\textsuperscript{133} Water plays a key role in the binding process; it was shown that interfacial layers of water molecules exhibit anisotropic behavior and form a collaborative network that facilitates the binding of the interfaces.\textsuperscript{134}

Interfacial water molecules play a critical role also in both the stability and the specificity of colicin DNase–immunity protein complexes.\textsuperscript{135} The complex between endonuclease colicin E2 and its Im2 immunity protein (E2/Im2) was presented as CAPRI target T47 with an addition to the docking experiment: groups submitting standard docking predictions were invited to also predict the positions of water molecules in the interface of the complex, using the method of their choice.\textsuperscript{136} These were then compared to the water positions in the crystal structure, a high-resolution (1.72 Å) structure determined at cryogenic temperatures (100 K).\textsuperscript{137} The docking itself presented little challenge, as both cognate (PDB 1emv; E9/Im9, PDB 7cei; E7/Im7) and noncognate (PDB 2wpt; E9/Im2; CAPRI T41) templates were available, but the prediction of interfacial water molecules proved to be much more difficult: only four of the 88 high-quality (root mean square with target $< 1.0$ Å) models submitted, that is, $< 5\%$, were found to have a water-mediated contact recall fraction $> 50\%$. A water-mediated contact is defined as a receptor–ligand contact where either ligand and receptor molecules have one or more heavy atoms within a 3.5 Å distance of the same water molecule. These results attest the relative immaturity of protein interface water prediction and show that further work is needed to attain a performance that is of practical use for drug design applications. Nevertheless, some promising observations could be made, namely, that three highly conserved water molecules, which are believed to be part of the protein–protein interface hotspot, were...
among the best predicted water positions and that another water molecule, involved in the specificity for the family of complexes, was also relatively well predicted. 136

Hydrophilic association characterizes most nonobligate protein complexes. Also in transient protein–protein interactions, which lie at the basis of most cellular processes, water plays an essential, mediating role. 138 Although larger in size, protein–protein interfaces constitute weaker binding sites with respect to small molecules. Successful well-known drugs such as aspirin and ibuprofen transiently bind such protein–protein interfaces and do not shut down, but rather modulate overstimulated signal transduction pathways. PPIs are increasingly targeted in drug design, which is now entering the systems biology era. 139 The successful development of drugs targeting such protein–protein interfaces indubitably benefits from a reliable prediction of interfacial water molecules. 140

For the association of large assemblies, continuum approaches may prove useful for the prediction of water molecule positions at interfaces and, in particular, for the energetic characterization of (large) complexes. Recently, Smouei et al. 141 modeled the formation of amyloid fibrils, protein aggregates that cause brain tissue damage, and compared the findings with experiment. They employed molecular dynamics simulations and, for the calculation of solvation-free energies, a continuum description using an extension of the standard Poisson–Boltzmann equation. This extension, the Poisson–Boltzmann–Langevin equation, considers the water molecules as point dipoles. 142 A solver for the Poisson–Boltzmann–Langevin equation had previously been developed by Koehl and Delarue. 11

**Conclusion**

High-throughput X-ray crystallography of a target alone or in complex with small molecules has significantly grown these last years. With the development of increasingly more sophisticated computational tools, SBDD is becoming a key step in the development of target-based therapies. These integrative approaches, which are primarily driven by increasingly powerful computational platforms, have allowed many success stories of the use of computer-assisted drug design in the discovery of new drugs. In addition, molecular docking approaches are being used to reach other goals such as the elucidation of noncanonical enzymatic mechanisms or the depiction of the quaternary structure of biological protein complexes. Such analyses, which are used in close coupling with traditional medicinal chemistry techniques, are increasingly relevant with drug design entering the systems biology era.

**Acknowledgments**

We acknowledge the support from the research federation FRABio (CNRS FR3688), “Structural and Functional Biochemistry of Biomolecular Assemblies”. JdR acknowledges funding from the Nord-Pas-de-Calais Regional Council and M de Wergifosse for fruitful discussion about ONIOM method. RB and MFL acknowledge financial support from the French Agence Nationale de Recherche, project “Fluctuations in Structured Coulomb Fluids”, grant number ANR-12-BSV5-0009-01. MFL is grateful to the French Agence Nationale de Recherche for his grant number ANR-13-BSV8-0002-0.

**Disclosure**

The authors report no conflicts of interest in this work.

**References**


Molecular docking and structure-based drug design


