ELSEVIER

*Teaser This review discusses current search strategies and evaluation methods for investigating protein–protein docking, two important issues that are quite different from those of protein–ligand docking.*



# Search strategies and evaluation in protein–protein docking: principles, advances and challenges

**Sheng-You Huang**[1,2,3]

[1] Research Support Computing, University of Missouri, Columbia, MO 65211, USA
[2] University of Missouri Bioinformatics Consortium, University of Missouri, Columbia, MO 65211, USA
[3] Department of Computer Science, University of Missouri, Columbia, MO 65211, USA

**Protein–protein docking is attracting increasing attention in drug discovery research targeting protein–protein interactions, owing to its potential in predicting protein–protein interactions and identifying 'hot spot' residues at the protein–protein interface. Given the relative lack of information about binding sites and the fact that proteins are generally larger than ligand, the search algorithms and evaluation methods for protein–protein docking differ somewhat from those for protein–ligand docking and, hence, require different research strategies. Here, we review the basic concepts, principles and advances of current search strategies and evaluation methods for protein–protein docking. We also discuss the current challenges and limitations, as well as future directions, of established approaches.**

**Sheng-You Huang** has been working in the molecular docking field since 2003, and on protein–protein docking in particular since 2007. He is currently a system administrator in the Research Support Computing and the University of Missouri Bioinformatics Consortium at the University of Missouri. He is also a research assistant professor in the Department of Computer Science at the University. His research interests include molecular modeling, bioinformatics, and computational biophysics and/or biology and their application to drug discovery. He actively develops novel docking algorithms and energy-scoring functions for protein–ligand interactions, protein–protein interactions, protein–RNA interactions, and modeling of quantitative structure-function relationships of therapeutically important proteins. He obtained his PhD in 2003 from the Wuhan University where he studied computer simulations in aspects of soft matter and biological physics. His research has resulted in the publication of scientific software, book chapters and more than 60 peer-reviewed journal articles.

Given that protein–protein interactions have an important role in many biological functions in living organisms [1], determination of the structure of the protein–protein complexes involved in these interactions is vital for revealing biological process pathways, to investigate the mechanisms interacting between proteins and to identify the crucial 'hot spot' residues in interactions that are important for drug discovery [2–9]. With the rapid development of structural proteomics projects, the 3D structures of many protein–protein complexes have been determined using various techniques, such as X-ray crystallography and nuclear magnetic resonance spectroscopy, and have been deposited in the Protein Data Bank (PDB) [10]. However, compared with the progress achieved for individual proteins, development in experimentally determining the complex structures between proteins is still limited because of the technical difficulties and high cost of the experimental methods involved [11]. Therefore, computational tools, such as protein–protein docking, which predicts the binding mode and free energy between individual protein structures, are needed to complement the experimental methods for the identification of protein–protein interactions and determination of their complex structures. Since pioneering work by Janin and Wodak [12], the protein–protein docking field has advanced considerably and many protein–protein docking algorithms have been developed over the past two decades [13–22].

*E-mail address:* huangshe@missouri.edu.

1359-6446/06/$ - see front matter  http://dx.doi.org/10.1016/j.drudis.2014.02.005



Please cite this article in press as: S.Y. Huang, Search strategies and evaluation in protein–protein docking: principles, advances and challenges, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.02.005

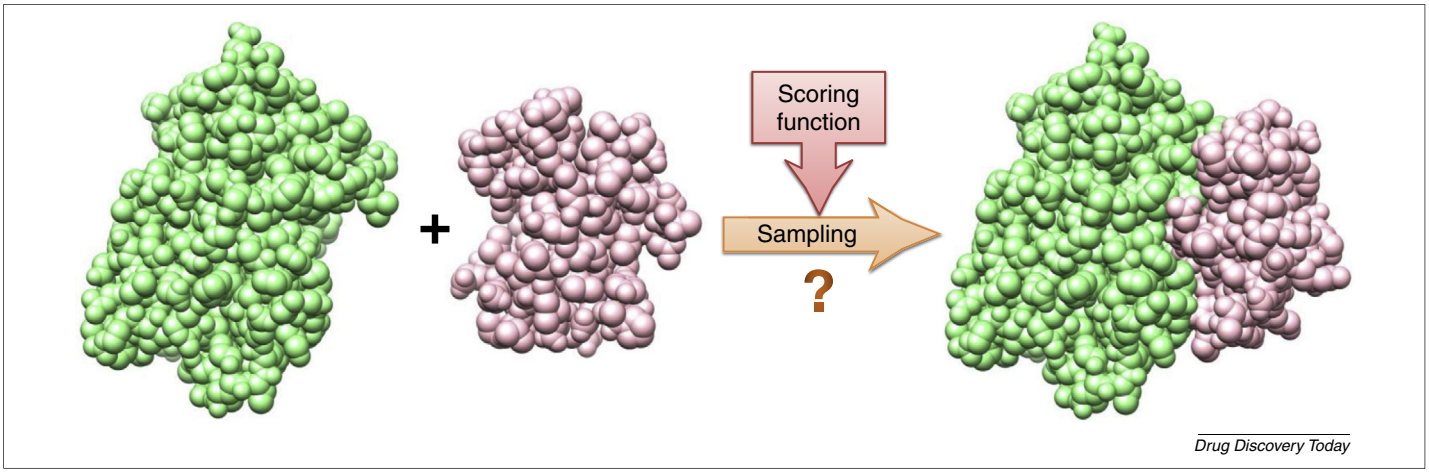


**FIGURE 1**

An illustration of protein–protein docking where the binding complex of two individual proteins (PDB code 1UDI) is constructed by sampling putative binding conformations that are evaluated and ranked by a scoring function.

Similar to protein–ligand (small molecule) docking [23], protein–protein docking also comprises two important components: sampling and scoring (Figure 1) [18]. These two processes can be coupled together during the docking process or can occur in different stages as adopted in post-docking approaches. Sampling is a search process that generates possible binding orientations and/or conformations (i.e. modes) between two molecules. It can be further divided into (i) rigid-body sampling of binding orientations; and (ii) conformational sampling of molecules, whereby rigid-body sampling is performed by the orientational search algorithm and conformational sampling is achieved by explicit protein flexibility consideration. Scoring is the measurement using a scoring function of the binding tightness and/or score between two molecules in a binding mode. The evaluated binding modes are then ranked according to their binding scores so that a set number of top binding modes can be selected as the final docking solutions. Therefore, in molecular docking, up to three aspects (orientational search, protein flexibility, and scoring functions) can be involved in a docking process.

Although sampling and scoring are the two main components for both protein–protein and protein–ligand docking, they do not necessarily use the same algorithm. For scoring, both protein–protein docking and protein–ligand docking use similar types of scoring method, which can normally be grouped into three basic categories: (i) force-field based; (ii) knowledge based; and (iii) empirical, as well as a combination of two or all of them [21,24]. In addition, because of the same type of targets (i.e., proteins) in both protein–protein and protein–ligand docking, algorithms that consider protein flexibility are also similar in both docking types; these normally include side chain and/or backbone flexibility, loop rearrangements, domain movements, and so on [14–18]. However, because less information is available relating to binding sites and the large size of proteins in protein–protein docking [18,23], the orientational search algorithm often requires strategies for protein–protein docking that are different from those for protein–ligand docking. Thus, many global and/or local search strategies have been developed for various protein–protein docking algorithms. In addition, given the larger size of proteins and larger binding interface, the evaluation method for protein–protein docking is also different from that for protein–ligand docking. Here, we give a detailed overview of the basic concepts, principles and specific features of current search strategies and evaluation methods in protein–protein docking. We also discuss challenges and limitations in existing algorithms and make suggestions for potential future research directions.

## Protein–protein docking: an overview of search strategies

The search strategies in currently available protein–protein docking algorithms can be grouped into three basic categories [(i) exhaustive global search; (ii) local shape feature matching; and (iii) randomized search] and one broad category of post-docking approaches (Table 1).

### *Exhaustive global search*

As mentioned above, because of a lack of information about binding sites, the investigation of protein–protein docking requires a global search for the binding orientations over six degrees of freedom (3D translational plus 3D rotational). Therefore, theoretically, the computational cost for an exhaustive global search has an order of $O(N^6)$, where $O(N^3)$ is from a 3D translational search and $O(N^3)$ is for a 3D rotational search. In an actual docking, one protein is normally fixed (the so-called 'static molecule') and the other protein is moved around the static protein (the so-called 'moving molecule'). Search over three rotational degrees of freedom is often separated from that over three translational degrees of freedom, in that the moving molecule is first rotated by an Euler angle in 3D rotational space. Then, for the rotation, an exhaustive search is carried out for the moving protein relative to the static protein in the complete 3D translational space. The above process is repeated until the entire 3D rotational space is sampled completely.

Given the typical size of approximately 60 Å for a protein, the search for the relative translations of two proteins will need to cover approximately $120 \times 120 \times 120$ Å$^3$ in the 3D translational space for a single rotation. If a grid spacing of 1.2 Å is used for discretizing the translational space during the search, there will be


Please cite this article in press as: S.Y. Huang, Search strategies and evaluation in protein–protein docking: principles, advances and challenges, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.02.005