Methods xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Modeling of protein–peptide interactions using the CABS-dock web server for binding site search and flexible docking

Maciej Blaszczyk [1], Mateusz Kurcinski [1], Maksim Kouza [1], Lukasz Wieteska, Aleksander Debinski, Andrzej Kolinski, Sebastian Kmiecik *

Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

ARTICLE INFO

Article history:
Received 4 May 2015
Received in revised form 6 July 2015
Accepted 8 July 2015
Available online xxxx

Keywords:
Protein–peptide docking
Flexible docking
Molecular docking
Peptide folding
Peptide binding
CABS-dock

ABSTRACT

Protein–peptide interactions play essential functional roles in living organisms and their structural characterization is a hot subject of current experimental and theoretical research. Computational modeling of the structure of protein–peptide interactions is usually divided into two stages: prediction of the binding site at a protein receptor surface, and then docking (and modeling) the peptide structure into the known binding site. This paper presents a comprehensive CABS-dock method for the simultaneous search of binding sites and flexible protein–peptide docking, available as a user's friendly web server. We present example CABS-dock results obtained in the default CABS-dock mode and using its advanced options that enable the user to increase the range of flexibility for chosen receptor fragments or to exclude user-selected binding modes from docking search. Furthermore, we demonstrate a strategy to improve CABS-dock performance by assessing the quality of models with classical molecular dynamics. Finally, we discuss the promising extensions and applications of the CABS-dock method and provide a tutorial appendix for the convenient analysis and visualization of CABS-dock results. The CABS-dock web server is freely available at http://biocomp.chem.uw.edu.pl/CABSdock/.

1. Introduction

Despite the significant progress in experimental and theoretical studies of protein–peptide interactions, the understanding of their role in the cellular machinery remains quite limited. Over the years it has become clear that understanding a particular protein function as a combination of its functional domains is not complete [1]. It has been shown that in higher eukaryotes up to 40% of protein–protein interactions (PPIs) are mediated by peptides [2]. Peptides responsible for PPIs are not necessarily independent molecules, but more often appear as disordered regions within proteins (at termini, between domains or flexible loops) that can act as a separate peptide molecule. The view that proteins can be understood through their discrete segments has already provided important insights into protein function [1]. Especially, protein–peptide interactions can be found in intracellular signaling pathways, cell localization, immune response, and protein degradation. Their new functional roles are constantly being discovered [3]. Importantly, many of these interactions are implicated in human diseases such as cancer or autoimmune disease [4–7]. Therefore, structure-based studies directed toward the design of completely new or modified receptor-interacting peptides have become a hot spot of current biomedical research.

In comparison to PPIs, protein–peptide interactions are more transient and interaction affinity is significantly weaker. Together with the high conformational flexibility of peptides, these factors make structural characterization of protein–peptide complexes really challenging. Therefore, there is an urgent need for the development of complementary computational approaches, such as effective molecular docking [8]. Assuming that the structure of a protein receptor has been solved experimentally or modeled with good accuracy, the modeling protocol for searching new protein–peptide interactions usually has two or three major steps:

(1) The first step involves identification of the binding site on the protein surface. This goal can be accomplished by bioinformatics methods using data from already known protein and protein–peptide structures or simply protein sequences [9–12]. They mostly aim at creating a library of sequence, structure or surface landscape motifs that could be universally detected in unknown proteins [13–15]. It should be noted that due to its simplicity,

* Corresponding author.
E-mail address: sekmi@chem.uw.edu.pl (S. Kmiecik).
[1] These authors contributed equally to this paper.

http://dx.doi.org/10.1016/j.ymeth.2015.07.004
1046-2023/

information only about sequence patterns that occur within binding sites is not sufficient for accurate binding site prediction and could result in a high ratio of false-positives [1].

(2) Second, the peptide is docked to a known binding site using local docking techniques, such as adapted molecular dynamics [16], Rosetta FlexPepDock [17,18], HADDOCK [19,20] or PepCrawler [21] methods (see reviews [2,8]).

(3) Third, those methods for local peptide docking may also serve in the final modeling step: high resolution refinement of initially generated peptide poses.

The first two steps of modeling protein–peptide interactions can also be achieved using techniques for the combined search of binding sites and peptide poses [13,22,23]. Usually, these methods allow the identification of a binding site, although the quality of resulting peptide models is often unsatisfactory [2].

Recently, we have developed a CABS-dock method and a web server for the simultaneous prediction of binding sites and protein–peptide docking [24]. The CABS-dock simulation engine, based on the coarse-grained CABS model, enables efficient docking search of fully flexible peptides over the entire surface of flexible proteins in a reasonable time, typically 1–8 CPU hours (which is thousandfold shorter than analogical simulations using rapid molecular dynamics adapted to peptide docking [22]). CABS-dock has been extensively tested over the largest benchmark set of non-redundant protein–peptide interactions available to date (including docking to bound and unbound receptor forms). For over 80% of bound and unbound dataset cases, we obtained high or medium accuracy models (expected to be of sufficient accuracy for high resolution refinement) [24].

In comparison to other protein–peptide docking tools (listed above), the CABS-dock offers the following major advantages: (1) the method does not require knowledge of the binding site nor any information about the peptide conformation, (2) during docking peptide conformation is allowed to be fully flexible, and (3) it is possible to simulate significant conformational changes of the protein receptor structure (see Section 3.2.1). These advantages become even more apparent in comparison to general purpose protein–ligand docking tools, which are usually less efficient in sampling conformational changes than methods dedicated to protein–peptide docking. The possible CABS-dock disadvantages include: (1) lack of option to guide the docking with the knowledge of the binding site (this will be available in the next CABS-dock update planned in 2015), however, it is possible to exclude some receptor areas from the docking search, and thereby to enforce more effective search in a closer neighborhood of the potential binding site (see Section 3.2.2); (2) a small set of 10 best scored models may not show the high accuracy models (that may be present in the large set of CABS-dock predicted models), however, this is also the case for the other docking methods (scoring problem is discussed in Section 3.4).

In this work, we evaluate CABS-dock performance and focus on particular examples of protein–peptide docking. The examples discussed illustrate CABS-dock performance using the default server mode as well as its advanced options. We also address the possibility of improving CABS-dock performance using an external scoring method over a large set of CABS-dock generated models. This can be achieved by a two-step procedure involving: (1) reconstruction and local optimization of CABS-dock models, followed by their (2) scoring using short simulations by all-atom molecular dynamics with explicit solvent. An Appendix is also provided with this paper where we provide a tutorial for the display and analysis of CABS-dock models and trajectories using VMD [25], a molecular graphics program.

2. Method

The CABS-dock method [24] is based on the CABS model (described in detail in Ref. [26]) that was originally designed for the structure prediction of globular proteins and simulation of protein dynamics. CABS comes from the letters of pseudo-atoms used to represent a single protein residue: carbon alpha (CA), carbon beta (B) and side chain (S). An additional pseudo-atom, defined in the geometrical center of the virtual CA–CA bond, is used to define the main chain hydrogen bonds. To speed up calculations, the coordinates of the CA atoms are restricted to the beads of a dense cubic lattice with lattice spacing arbitrary set to 0.61 Å. The remaining pseudo-atoms are located off the lattice and follow the movement of the main chain. The force field is based on knowledge-based statistical potentials derived from structural regularities seen in known protein structures. Sampling is controlled by the asymmetric Metropolis criterion. Additionally, CABS uses the Replica Exchange protocol for better sampling coverage of the energy landscape.

The CABS model was initially used for protein structure prediction and it performed exceptionally well in CASP6 (Critical Assessment of protein Structure Prediction, a community-wide blind test of structure prediction approaches). Using the CABS-based approach, the Kolinski–Bujnicki group scored best or second best, depending on the evaluation method [27,28]. CABS-based protocols for the *de novo* and consensus prediction of protein structure were made freely available to the academic community on an automated web server [29]. The CABS model has also been successfully used to simulate the dynamics of denatured protein states [30], protein folding mechanisms [31–36], the flexibility of globular proteins [37–39] and its influence on protein aggregation [40]. Finally, the CABS model has been optimized for the investigation of protein interactions and prediction of structures of protein complexes. It has been used to build a model of human telomerase [41], protein–peptide and protein–protein docking [42–45] and to investigate the mechanism of simultaneous folding and binding of an intrinsically disordered peptide [46].

The pipeline of the CABS-dock server is a multistage protocol that consists of multiple programs and associating scripts, with the CABS model (version dedicated to handle multimeric protein chains) at its center. As shown in Fig. 1, the whole procedure consists of four main stages (1) flexible docking by the CABS algorithm; (2) initial filtering of probable solutions from all generated models; (3) further selection of representative models by the clustering protocol and (4) reconstruction to all-atom representation and local optimization of final models. The method is fully automated. As an input it requires only the 3D structure of the receptor and the peptide sequence. On output the server returns 10 top scored models of the protein–peptide complex. Each step of the procedure is briefly described in the following paragraphs.

2.1. Flexible docking with the CABS model

The CABS-dock method requires two inputs: (1) amino acid sequence of the peptide, and (2) 3D structure of the protein receptor (the obligatory and non-obligatory input recommendations are listed in [24]). In the first CABS-dock modeling step, 10 copies of the protein–peptide system are generated as starting models for the Replica Exchange Monte Carlo sampling method. Each starting copy contains a random peptide structure that is placed in a random position within 20 Å from the input receptor structure (see Fig. 2, left). During simulation of coupled peptide binding and folding, the CABS-dock protocol allows full flexibility of the peptide