Published online 05 May 2015 *Nucleic Acids Research, 2015, Vol. 43, Web Server issue* ***W419–W424***
doi: 10.1093/nar/gkv456

CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site

Mateusz Kurcinski†, Michal Jamroz†, Maciej Blaszczyk, Andrzej Kolinski* and Sebastian Kmiecik*

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

Received March 05, 2015; Revised April 07, 2015; Accepted April 24, 2015

ABSTRACT

Protein–peptide interactions play a key role in cell functions. Their structural characterization, though challenging, is important for the discovery of new drugs. The CABS-dock web server provides an interface for modeling protein–peptide interactions using a highly efficient protocol for the flexible docking of peptides to proteins. While other docking algorithms require pre-defined localization of the binding site, CABS-dock does not require such knowledge. Given a protein receptor structure and a peptide sequence (and starting from random conformations and positions of the peptide), CABS-dock performs simulation search for the binding site allowing for full flexibility of the peptide and small fluctuations of the receptor backbone. This protocol was extensively tested over the largest dataset of non-redundant protein–peptide interactions available to date (including bound and unbound docking cases). For over 80% of bound and unbound dataset cases, we obtained models with high or medium accuracy (sufficient for practical applications). Additionally, as optional features, CABS-dock can exclude user-selected binding modes from docking search or to increase the level of flexibility for chosen receptor fragments. CABS-dock is freely available as a web server at http://biocomp.chem.uw.edu.pl/CABSdock.

INTRODUCTION

Although protein–peptide interactions play key roles in cell functions, relatively little is known about the structural details of these complexes. The highly dynamic and transient nature of peptide binding makes experimental investigation difficult. Thus, computer aided support, such as molecular docking, is needed. The major problem in molecular docking is the treatment of protein flexibility (1,2). Namely, current algorithms are not efficient enough to handle the high conformational fluctuations of peptides. An additional difficulty comes from dealing simultaneously with the receptor's flexibility, which, even if small, is extremely costly for most of the computational models.

Computational protein–peptide docking is usually divided into three consecutive steps realized by separate protocols (3): (1) prediction of the binding site(s) on the receptor structure, (2) initial modeling of the peptide backbone in the binding site(s), and finally (3) refinement of the initial protein–peptide complexes to high resolution. The CABS-dock method, presented in this paper, is an attempt to unify all these three steps into single, highly efficient simulation of coupled folding and binding of the peptide to the flexible receptor structure. To our knowledge, such an approach to docking, without prior knowledge of the binding site, has been successful so far only when applied to very short peptides (2–4 amino acids) (4). We tested performance of the CABS-dock protocol on a wide benchmark set of protein–peptide complexes with peptides of 5–15 amino acids. The benchmark contains 103 bound and 68 unbound protein receptor cases (determined experimentally in complex with a peptide and without a peptide, respectively).

MATERIALS AND METHODS

Previous applications and background

The CABS-dock docking protocol was developed and validated during the following simulation studies: mechanism of folding and binding of an intrinsically disordered peptide (5), docking antigen-mimicking peptides to an antibody (6) and docking peptide co-activators to nuclear receptors (7,8). These studies showed that the method is able to predict complex arrangements close to the native structure. Importantly, in all the validation tests mentioned above,

*To whom correspondence should be addressed. Tel: +48 22 822 02 11 (Ext 310); Fax: +48 22 822 02 11 (Ext 320); Email: sekmi@chem.uw.edu.pl
Correspondence may also be addressed to Andrzej Kolinski. Tel: +48 22 822 02 11 (Ext 320); Fax: +48 822 02 11 (Ext 320); Email: kolinski@chem.uw.edu.pl
†These authors contributed equally to the paper as first authors.

peptides were allowed to be fully flexible and no information about the binding site or peptide conformation was used.

The CABS-dock protocol is a multiscale modeling procedure based on the coarse-grained CABS protein model. The CABS model has been designed to provide significant efficiency in the treatment of protein conformational changes, while preserving high local accuracy (enabling seamless reconstruction to all-atom representation). In the CABS model, each amino acid is represented by up to four interaction centers, simulation dynamics is controlled by the Monte Carlo scheme and the force field is based on statistical potentials (force field is summarized in the Supplementary Data, details have been described elsewhere (9)). Additionally to the aforementioned protein docking studies, we have demonstrated that the CABS protein model enables reliable simulations of protein dynamics: long-term folding mechanisms (10,11) and short-term fluctuations close to the native state (12,13). CABS has also been successfully used in protein structure prediction exercises, showing exceptional performance especially in blind predictions of short globular proteins (14) and large protein fragments (15,16). Altogether, these studies demonstrate the validity of the CABS interaction model and sampling scheme in simulations of simultaneous folding and binding, such as performed in the CABS-dock protocol.

Protocol overview

The CABS-dock protocol consists of the following steps (presented also in Figure 1 and Supplementary Figure S1 flow chart):

1. Generating random structures. Random structures of the peptide are generated and randomly placed on the surface of the sphere centered at the receptor's geometrical center (the radius of the sphere is the receptor dimension in the longest direction + 20 Å).
2. Simulation of binding and docking. The CABS-dock procedure utilizes Replica Exchange Monte Carlo dynamics with 10 replicas uniformly spread on the temperature scale. Additionally the temperatures of the replicas constantly decrease as the simulation proceeds to end on the bottom of the energy minima. On output the procedure produces 10 trajectories (one for each replica), each consisting of 1000 time-stamped simulation snapshots for a combined total of 10 000 models. During the simulation, the receptor molecule is kept in near-native conformations by a set of distance restraints binding pairs of C-alpha atoms. The restraints are selected from the distance map calculated on the input structure based on the following conditions: only C-alpha atoms located within a 5–15 Å range from each other are restrained; the minimum sequence gap between restrained residues is set to 5; violation of the restraint by less than 1 Å is not penalized; beyond that the energetic penalty increases linearly. If the user marks some of the residues as semi-flexible or fully flexible, the slope of the penalty is halved or set to 0, respectively, for all restraints assigned to the marked residues.
3. Selection of the final models is a two-step procedure:
   a. Initial filtering. From each of the 10 trajectories, all unbound states are excluded and next 100 lowest binding energy models are selected (or less if a trajectory contains less than 100 bound states, which is rarely the case) for the next step of the procedure.
   b. The *k*-medoids clustering. Selected models (1000 in total) are clustered together in the *k*-medoids procedure. Clustering is performed 100 times with different initial medoids and $k = 10$. Ten consensus medoids are selected as the final models.
4. Reconstruction of the 10 final models. Final models are reconstructed from the C-alpha trace to an all-atom representation and subsequently undergo optimization using Modeller (17) with DOPE statistical potential (18).

PERFORMANCE

Docking without prior knowledge of the binding site

We have validated the CABS-dock docking protocol against the largest dataset of non-redundant peptide–protein interactions (<70% sequence identity with respect to the receptor protein) available to date. The benchmark was originally created to test the FlexPepDock refinement procedure (3) and subsequently extended (with new unbound cases) in a study testing the HADDOCK algorithm (docking driven by knowledge of the binding site) (19).

We assess the quality of docking models using ligand RMSD (root-mean-square deviation) as follows:

- High-quality prediction: RMSD < 3 Å.
- Medium-quality prediction: 3 Å ≤ RMSD ≤ 5.5 Å.
- Low-quality prediction: RMSD > 5.5 Å.

The RMSD is calculated on the peptide only, after superimposition of the receptor structures. We have decided to set up an arbitrary cutoff of 5.5 Å (between low- and medium-accuracy models) on the basis of the work benchmarking the Rosetta FlexPepDock protocol for the refinement of coarse models of protein–peptide complexes (3). In that work, the authors defined an effective 'basin of attraction' of 5.5 Å resolution, from which the FlexPepDock protocol is able to reliably recover near-native protein–peptide models (in 91% of bound docking cases). Importantly, low-quality prediction (as defined by the 5.5 Å cutoff) does not mean that the obtained models are useless for further refinement. In the FlexPepDock benchmark, starting the refinement from structures of 6.5–7.5 Å resolution resulted in near-native models in 48% of cases of bound docking (3).

In our performance test, we used neither information about the bound peptide structure nor about the binding site (blind prediction test). As shown in Figure 2, within the set of resulting models, high- or medium-accuracy models can be found, both in the entire set of predicted models (all) or in the top selected models (top 10) (for the detailed results of bound and unbound cases see Supplementary Tables S1 and S2, respectively). Moreover, the CABS-dock performance for bound cases is on the same level as that for unbound cases (the pairs of bound and unbound counterparts are listed in Supplementary Table S3). This is because, for the majority of benchmark cases, the difference between binding interfaces of bound and unbound protein forms is

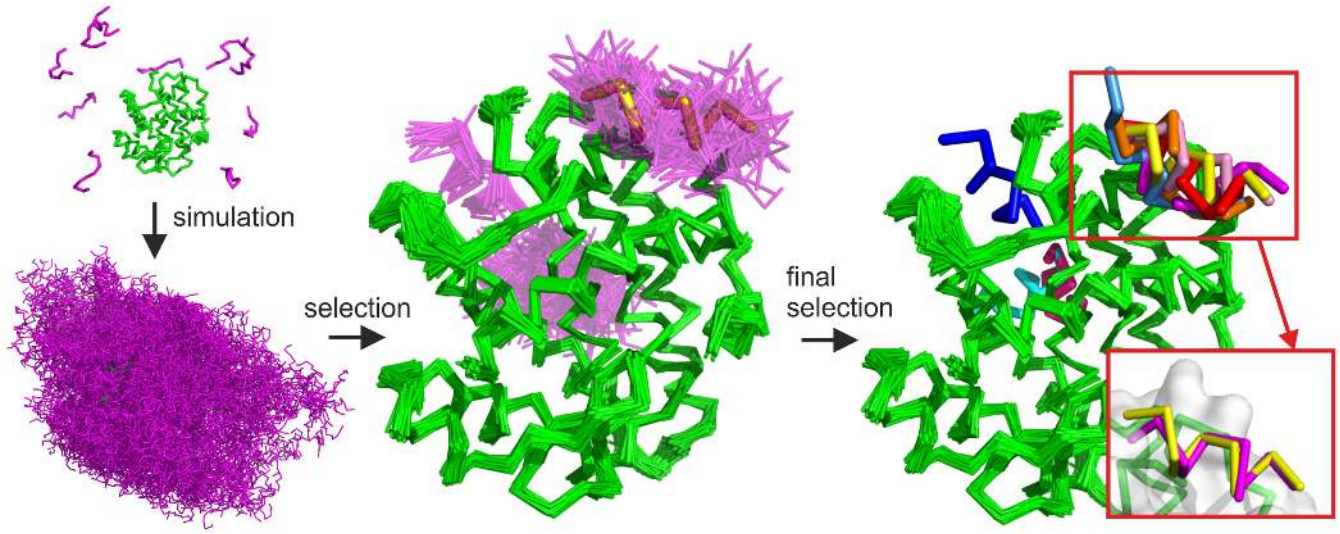

Figure 1. Basic stages of the CABS-dock protocol illustrated on the example benchmark case (PDB ID: 2P1T). The protein receptor is colored in green, modeled peptide conformations in magenta and the reference native peptide structure in yellow. The following CABS-dock stages are visualized: (1) simulation start (from random conformations and positions of the peptide); (2) simulation result (a set of 10 000 models); (3) filtering and clustering result (a set of models grouped in similar binding modes and similar peptide conformations); (4) final models (a set of 10 representative models). In the presented benchmark case, 7 of the 10 final models were docked in the native binding site (marked in red rectangle). Among these, the best accuracy model was within 1.37 Å to the native (shown in the right bottom corner superimposed on the native peptide structure). For a more detailed flow chart of the CABS-dock pipeline, see Supplementary Figure S1.

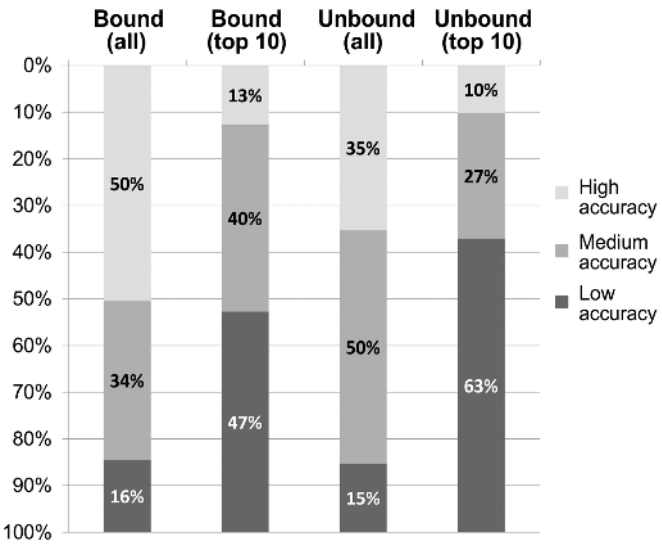

Figure 2. CABS-dock performance summary for 103 bound and 68 unbound benchmark cases. The percentages of high-, medium- or low-accuracy models (quality assessment criteria are given in the text) are reported for the best quality models found in the sets of 10 000 models (all) and in the sets of 10 final models (top 10). Detailed results, for each modeled complex and each prediction run, are available in Supplementary Tables S1 (bound docking cases) and S2 (unbound docking cases).

small (lower than 1 Å (19)). Therefore, such small protein changes are well handled by CABS-dock using the default settings of protein receptor flexibility.

In summary, for over 80% of bound and unbound dataset cases, we obtained models with high or medium accuracy that is sufficient for practical applications (at least for further refinement to higher resolution (3)). The prediction results may qualitatively differ between different prediction runs (due to stochastic nature of the simulation model). The analysis of the accuracy of the predicted models in different prediction runs showed that the modeling cases generally fall into two categories: those with consistent and those with qualitatively distinct predictions (see Supplementary Tables S1 and S2). Therefore in ambiguous prediction cases, we suggest running a few independent runs and analyzing the predictions jointly.

SERVER DESCRIPTION

Input interface and requirements

The required input includes:

- Protein receptor structure in the PDB format or protein PDB code along with the chain identifier(s), for example: 1RJK:A or 1CE1:HL. The following requirements apply to the input PDB files: single or multichain proteins are accepted (chains must be up to 500 amino acids in length); each residue in the provided PDB file should have a complete set of backbone atoms (i.e. N, Cα, C and O); side chain atoms may be missing. Non-standard amino acids are automatically changed to their standard counterparts.
- Peptide sequence (in one letter code, standard amino acids only, maximum 30 amino acids in length)

The optional input includes:

- Peptide secondary structure. If not provided, the PSIPRED method (20) for secondary structure prediction is automatically used. For best results, if the peptide secondary structure was derived experimentally, we sug-

gest providing experimental assignments. The secondary structure should be provided for each residue in a three letter code: H, helix; E, extended state (beta strand); and C, coil (less regular structures). Overpredictions of the regular secondary structure (H or E) are more dangerous for the quality of the results than underpredictions (i.e. for residues with an ambiguous secondary structure prediction assignment, it is better to assign coil (C) than a regular (H or E) structure).
- Project name. Recommended: project names appear in the queue page.
- Email address. Recommended: if provided, the server will send an email notification about job completion.
- Advanced options (described in a separate paragraph below).

Output interface

The output interface is organized under the following tabs: 'Project information', 'Docking prediction results', 'Clustering details' and 'Contact maps'. The content of the latter three is briefly described in the following paragraphs.

Under the 'Docking prediction results' tab (see the screenshot in Figure 3A) the user may view in 3D and download 10 final models (representatives of 10 structural clusters found in the simulation—for details see Protocol overview in Methods). Additionally, the user may download a compressed archive with 10 final models, cluster models (all models that have been classified in structural clustering to particular clusters) and complete trajectories (in the PDB format and C-alpha representation). The archive also contains the input structure of the receptor and a log file with all information to recreate the simulation.

Under the 'Clustering details' tab (see the screenshot in Figure 3B) the user is provided with thorough insight into structural clustering data. An interactive chart shows the composition of clusters of models versus their trajectory affiliation. By clicking on the points representing models in the chart, the user may view in 3D and download particular models in the PDB format. The tab also contains a table summarizing basic information on clusters, such as density, diversity, etc.

The 'Contact maps' tab allows the user to investigate the interaction interface between the peptide and the receptor. An interactive chart shows a contact map between the peptide and the receptor residues. The user may define the value of the contact cutoff distance. The user also has a text list of the contacts shown on the maps.

Advanced options

For more advanced prediction runs, users may perform the following operations:

- Excluding binding modes. The user may select receptor residues that are unlikely to interact with the peptide to exclude some binding modes from the results. The user may provide such a list explicitly (available from the main page by checking the 'Mark unlikely to bind regions' option) or resubmit an already completed job (available in 'Project information' tab) with marked models (binding modes) to be excluded from future results.
- Allowing for the higher flexibility of selected receptor fragments. This advanced option enables removing distance restraints that keep the receptor structure in a near-native conformation (see Protocol overview in Materials and Methods section). For each selected residue, the user may choose from two preset settings: moderate or full flexibility. This option is available from the main page by checking the 'Mark flexible regions' option.
- Increasing simulation length. The user may increase the default (50) number of Monte Carlo macrocycles to be performed (the maximum number is 200). Increasing this number may be beneficial in more difficult modeling cases (e.g. for large receptor structures or for long peptides of more than 20 residues). However, an alternative way for dealing with such more demanding cases is to run independent simulation runs and to analyze the results jointly.

Online documentation

CABS-dock documentation is available online and can be accessed using the links in the menu at the top of every server page. It contains a description of the method and a tutorial explaining how to access and interpret resulting data. The online documentation is updated on a regular basis according to users' needs or method improvements.

Server and output data availability

The CABS-dock server is free and open to all users, and there is no login requirement. After clicking on the submit button, a web link to the results is provided which can be bookmarked and accessed at a later time. Web links to the submitted jobs are also displayed on a queue page (available from the main page), unless the option 'Do not show my job on the results page' is marked. Note that the results will be available for a limited period of time (notification about data storage is displayed at the bottom of the job page).

Server architecture and run-time

The CABS-dock website interface and parsers were developed in the Python scripting language, using Flask framework and Jinja2 template engine. The molecular visualization used in the server is executed using 3Dmol.js (21) and JSmol (22). The CABS-dock website runs on the Apache2 and MySQL database for user queue storage. The CABS-dock queue is checked every 5 min by computational servers and any new jobs are added to the SGE queue. As soon as a job is started on the computational server, job status changes on the CABS-dock website (from 'pending' to 'running').

A Typical CABS-dock run takes about 3 h. However, in case of high server load the running time can take even few times longer. After completion, job results are sent back to the website and job status changes from 'running' to 'done' (or 'error'). Currently, CABS-dock server computations are performed on a Linux supercomputer cluster having about 100 CPU threads.

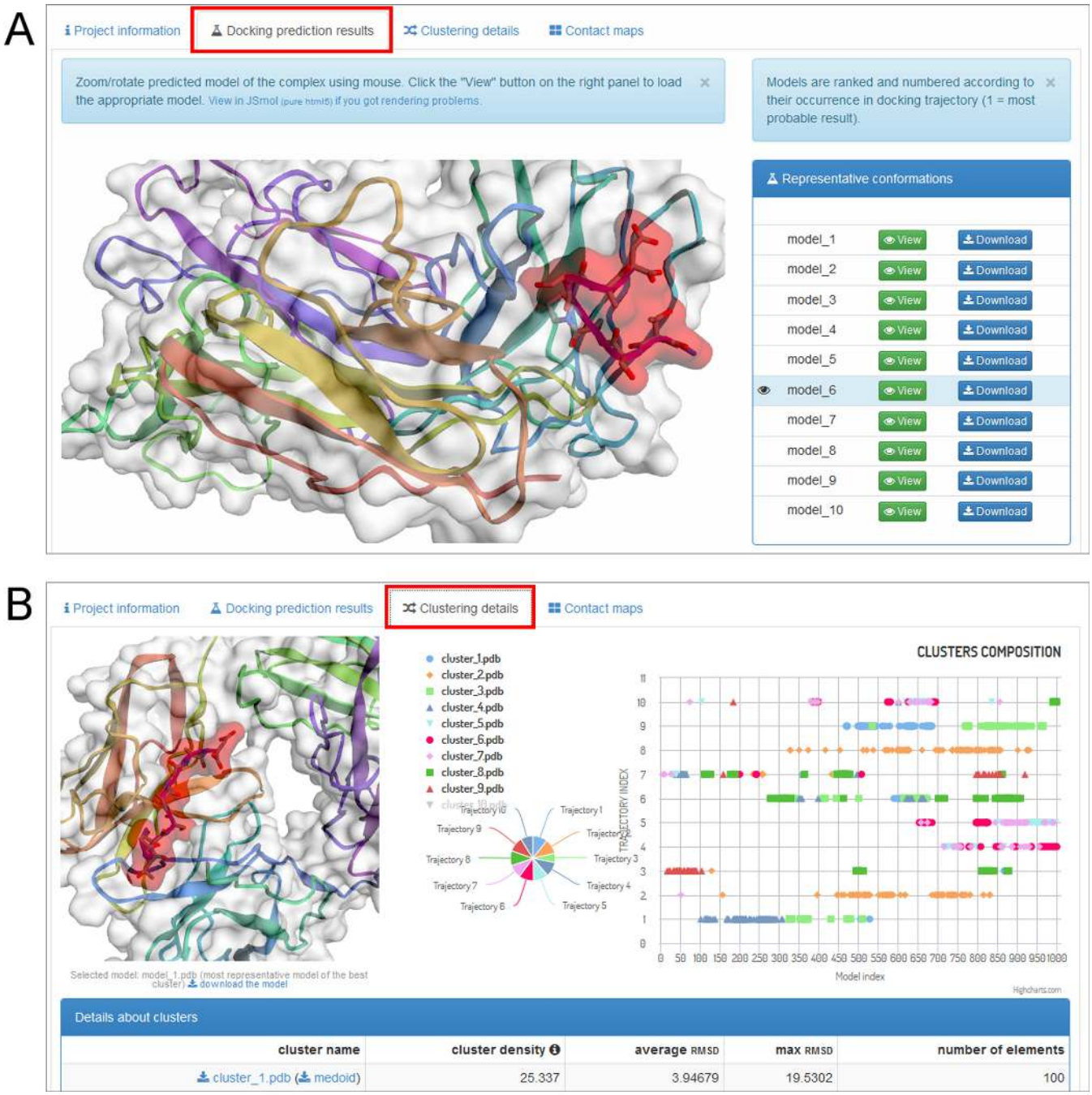

Figure 3. CABS-dock web server screenshots. Example output interface is presented: (A) 'Docking prediction results' tab and (B) 'Clustering details' tab.

SUMMARY

The CABS-dock protocol has been already successfully used in studies of protein–peptide interactions (5–8). Within this work, we developed an easy-to-use web server interface for the CABS-dock protein–peptide docking protocol. We expect that this web server will be widely applied to new systems as well as utilized as an element of new modeling procedures. The promising CABS-dock extensions include: for example, adding a refinement step (19,23); incorporation of experimental data; increasing the flexibility of appropriate receptor fragments (e.g. through predicted restraints (24)).

SUPPLEMENTARY DATA

Supplementary Data is available at NAR Online.

FUNDING

Foundation for Polish Science TEAM project [TEAM/2011–7/6] co-financed by EU European Regional Development Fund operated within the Innovative Economy Operational Program; Polish Ministry of Science and Higher Education [IP2015 016573]. Funding for open access charge: Foundation for Polish Science

TEAM project [TEAM/2011–7/6] co-financed by EU European Regional Development Fund operated within the Innovative Economy Operational Program.
Conflict of interest statement. None declared.

REFERENCES

1. Trellet,M., Melquiond,A.J. and Bonvin,A.J.J. (2015) Information-Driven Modeling of Protein-Peptide Complexes. In: Zhou,P and Huang,J (eds). *Computational Peptidology*. Springer, NY, Vol. **1268**, pp. 221–239.
2. Rubinstein,M. and Niv,M.Y. (2009) Peptidic modulators of protein-protein interactions: progress and challenges in computational design. *Biopolymers*, **91**, 505–513.
3. Raveh,B., London,N. and Schueler-Furman,O. (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins*, **78**, 2029–2040.
4. Hetényi,C. and van der Spoel,D. (2002) Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.*, **11**, 1729–1737.
5. Kurcinski,M., Kolinski,A. and Kmiecik,S. (2014) Mechanism of folding and binding of an intrinsically disordered protein as revealed by ab initio simulations. *Journal of Chemical Theory and Computation*, **10**, 2224–2231.
6. Horwacik,I., Kurcinski,M., Bzowska,M., Kowalczyk,A.K., Czaplicki,D., Kolinski,A. and Rokita,H. (2011) Analysis and optimization of interactions between peptides mimicking the GD2 ganglioside and the monoclonal antibody 14G2a. *Int. J. Mol. Med.*, **28**, 47–57.
7. Kurcinski,M. and Kolinski,A. (2010) Theoretical study of molecular mechanism of binding TRAP220 coactivator to Retinoid X Receptor alpha, activated by 9-cis retinoic acid. *J. Steroid. Biochem. Mol. Biol.*, **121**, 124–129.
8. Kurcinski,M. and Kolinski,A. (2007) Steps towards flexible docking: modeling of three-dimensional structures of the nuclear receptors bound with peptide ligands mimicking co-activators' sequences. *J. Steroid. Biochem. Mol. Biol.*, **103**, 357–360.
9. Kolinski,A. (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.*, **51**, 349–371.
10. Kmiecik,S. and Kolinski,A. (2011) Simulation of chaperonin effect on protein folding: a shift from nucleation-condensation to framework mechanism. *J. Am. Chem. Soc.*, **133**, 10283–10289.
11. Kmiecik,S. and Kolinski,A. (2007) Characterization of protein-folding pathways by reduced-space modeling. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 12330–12335.
12. Jamroz,M., Kolinski,A. and Kmiecik,S. (2014) CABS-flex predictions of protein flexibility compared with NMR ensembles. *Bioinformatics*, **30**, 2150–2154.
13. Jamroz,M., Kolinski,A. and Kmiecik,S. (2013) CABS-flex: server for fast simulation of protein structure fluctuations. *Nucleic Acids Res.*, **41**, W427–W431.
14. Blaszczyk,M., Jamroz,M., Kmiecik,S. and Kolinski,A. (2013) CABS-fold: server for the de novo and consensus-based prediction of protein structure. *Nucleic Acids Res.*, **41**, W406–W411.
15. Kolinski,A. and Bujnicki,J.M. (2005) Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins*, **61**(Suppl. 7), 84–90.
16. Kmiecik,S., Jamroz,M. and Kolinski,M. (2014) Structure prediction of the second extracellular loop in G-protein-coupled receptors. *Biophys. J.*, **106**, 2408–2416.
17. Eswar,N., Webb,B., Marti-Renom,M.A., Madhusudhan,M.S., Eramian,D., Shen,M.Y., Pieper,U. and Sali,A. (2007) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.*, Chapter 2, Unit 2.9.
18. Shen,M.Y. and Sali,A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, **15**, 2507–2524.
19. Trellet,M., Melquiond,A.S. and Bonvin,A.M. (2013) A unified conformational selection and induced fit approach to protein-peptide docking. *PloS One*, **8**, e58769.
20. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
21. Rego,N. and Koes,D. (2014) 3Dmol.js: molecular visualization with WebGL. *Bioinformatics*, **31**, 1322–1324.
22. Hanson,R.M., Prilusky,J., Renjian,Z., Nakane,T. and Sussman,J.L. (2013) JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. *Isr. J. Chem.*, **53**, 207–216.
23. London,N., Raveh,B., Cohen,E., Fathi,G. and Schueler-Furman,O. (2011) Rosetta FlexPepDock web server–high resolution modeling of peptide-protein interactions. *Nucleic Acids Res.*, **39**, W249–W253.
24. Ortiz,A.R., Kolinski,A. and Skolnick,J. (1998) Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 1020–1025.