W96–W99 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh354

ClusPro: a fully automated algorithm for protein–protein docking

Stephen R. Comeau[1], David W. Gatchell[2], Sandor Vajda[1,2] and Carlos J. Camacho[1,2,*]

[1]Bioinformatics Graduate Program and [2]Department of Biomedical Engineering, Boston University, 44 Cummington Street, Boston, MA 02215, USA

Received February 14, 2004; Accepted February 23, 2004

ABSTRACT

ClusPro (http://nrc.bu.edu/cluster) represents the first fully automated, web-based program for the computational docking of protein structures. Users may upload the coordinate files of two protein structures through ClusPro's web interface, or enter the PDB codes of the respective structures, which ClusPro will then download from the PDB server (http://www.rcsb.org/pdb/). The docking algorithms evaluate billions of putative complexes, retaining a preset number with favorable surface complementarities. A filtering method is then applied to this set of structures, selecting those with good electrostatic and desolvation free energies for further clustering. The program output is a short list of putative complexes ranked according to their clustering properties, which is automatically sent back to the user via email.

INTRODUCTION

The prediction of protein–protein interactions is one of the main challenges facing the proteomics community. The ultimate goal is to take the three-dimensional coordinates of two independently crystallized proteins which are known to interact, and to derive a model for the bound structure (1–3). Through the introduction of the Fourier correlation method (4–6), it is now possible to evaluate billions of putative complex structures covering a large set of the translational and rotational space of relative positions between the two molecules. The resulting output of these docking algorithms, typically, is a few near-native structures nested within a multitude of false positive structures that also have favorable surface complementarity. Various discrimination techniques (7–12) are then used to find the needle(s), i.e. near-native structures, in the proverbial haystack.

These discrimination techniques have evolved over the past decade, and they regularly try to incorporate components of the binding energy into the process, as it is assumed that the native structure is at a global free energy minimum. Other discrimination methods have tried to refine the interface of the structures, as the surface side-chains of the independently crystallized proteins are oftentimes in the wrong positions (13). With this refinement of the interface, the van der Waals contact energy is greatly improved, leading to an increase in surface complementarity. Also, many of the correct, i.e. energetically favorable, electrostatic interactions and hydrogen bonds can be established and will contribute to a more successful discrimination.

In our algorithm, we rapidly filter the output from the Fourier correlation algorithm using a combination of desolvation and electrostatic energies (calculated using a Coulombic potential). This approach results in several near-native structures passing through the filter, while eliminating many of the false positives. Our next step takes advantage of the fact that the free energy landscape exhibits its broadest and deepest well near the native structure, inferred to be the global minimum, with various local minima, which are narrower and shallower than the global minimum, scattered throughout the landscape. Therefore, to further discriminate, i.e., eliminate false positives, the putative structures are clustered together, with the center of the most populated cluster being a structure near the native binding site (14). This method is reminiscent of the work of Shortle *et al*. (15) for protein structure prediction, where the cluster containing the largest number of low-energy structures was typically the native fold. The clustering application to protein–protein docking was first introduced in the first Critical Assessment of PRediction of Interactions (CAPRI) (16) (http://capri.ebi.ac.uk/) by Camacho and Gatchell (17).

METHODS

Rigid body docking

Using ClusPro's (18) algorithm (Figure 1), the user has the option of selecting DOT (19,20) or ZDOCK (21,22) to perform rigid body docking, both of which are based on the fast Fourier

*To whom correspondence should be addressed. Tel: +1 617 353 4842; Email: ccamacho@bu.edu

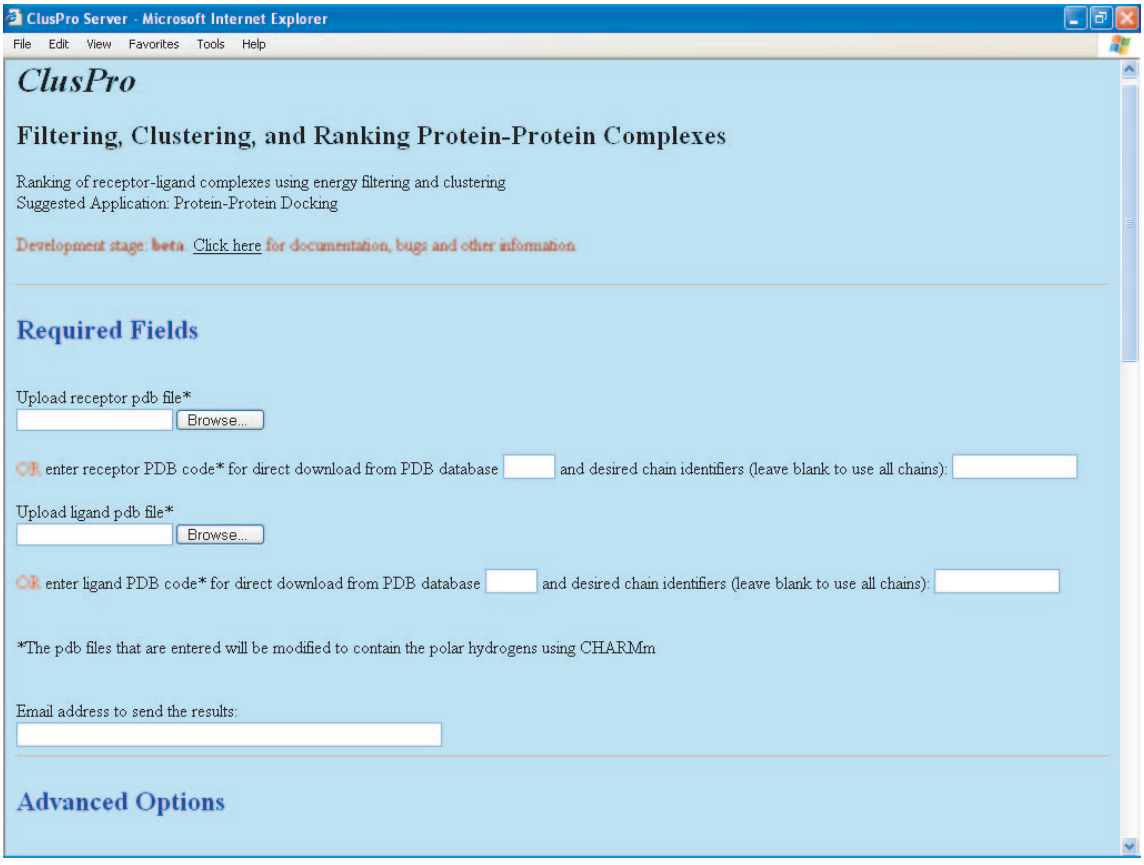

Figure 1. A screenshot of ClusPro's web interface displays the user-friendliness of the server.

transform (FFT) correlation techniques. Although DOT allows for the use of an electrostatic potential in the scoring function, we base the scoring solely on the surface complementarity between the two structures. DOT is run on a 128 Å × 128 Å × 128 Å grid, using a grid spacing of 1 Å. Using a pre-defined list of 13 000 rotations, over 2.7×10^{10} structures are evaluated, retaining 20 000 structures with the best surface complementarity scores, which are then further subjected to the empirical free energy filtering algorithm described below.

ZDOCK's scoring function is different from DOT's in the sense that it combines pairwise shape complementarity (PSC), desolvation, and electrostatics in its calculations. The scoring function of ZDOCK is more intricate, hence ZDOCK typically has a better ratio of true positives to false positives than using the simpler scoring function of DOT. This allows us to keep only 2000 structures, to which the filtering algorithm is then applied.

Filtering using empirical free energy functions

The initial structures from DOT are selected using shape complementarity, an approximation of the van der Waals contact energy, which is useful given that many proteins bury large surface areas upon binding (23). The free energy of the complex structure is typically dominated by van der Waals interactions, which is a very noisy function and difficult to calculate, especially with incorrect side-chain rotamers. Therefore, it is important to use other components of the binding free energy to account for the noise of the van der Waals energies. Therefore, we calculate the desolvation free energy using the atomic contact potential (24), which is a statistical measure of the desolvation free energy, and the electrostatic free energy using a Coulombic model with a distance-dependent dielectric of $4r$ (25). The default values are to retain 1500 structures exhibiting the best electrostatic energies and 500 structures with the best desolvation energies. The number of electrostatic structures is greater due to the relative accuracy of the two energy functions. The atomic contact potential is a smooth potential and, therefore, its calculated energies do not vary much for small distance perturbations, e.g. errors in the coordinates of the crystal structures. However, the electrostatic potential is highly sensitive to these perturbations, especially errors in the coordinates of solvent accessible side-chains, hence we allow an increased number of electrostatic structures, as compared to the number of desolvation structures, to pass through the filter in an attempt to retain more near-native structures. At this stage of ClusPro, regardless of the original docking method selected, there are 2000 candidate complexes retained for further processing.

Pairwise RMSD clustering

The top 2000 energetically favorable structures are then clustered on the basis of a pairwise binding site root mean squared deviation (RMSD) criterion. For each of the 2000 structures, the residues of the moving molecule (designated as the ligand) that have at least one atom within 10 Å of any atom of the still

molecule (designated as the receptor) are recorded into a list. Then, the distance between the Cα of each of those residues and the Cα of the corresponding residues on each of the 2000 ligands is calculated and stored into a matrix. Clusters are then formed by selecting the ligand that has the most neighbors below a previously selected clustering radius. Each member within the cluster is then eliminated from the matrix to avoid overlaps between clusters. This is repeated until at least 30 clusters are formed. The ligand with the most neighbors is the cluster center, and is the representative structure for the cluster. The top cluster centers are then CHARMm (26) minimized in the presence of the receptor, and concatenated into PDB NMR format, compressed in the Unix '.gz' format, and sent back to the user via email.

Computing power and time

The server runs using 16 processors on an IBM pSeries 690, with each running at 1.3 GHz and sharing 32 GB of memory. The average computational time used for a complex is approximately 4 h, but varies greatly depending on the size of the proteins submitted. The most time-consuming step of the algorithm is the filtering method, where the desolvation and electrostatic energies are calculated for each of the 20 000 putative structures. A limitation of the server is the size of the proteins submitted. The receptor protein can be no larger than 11 999 atoms after CHARMm minimization to include the polar hydrogens, and the ligand can be no larger than 4700 atoms after minimization.

DESCRIPTION OF THE WEB INTERFACE

As standard input, the user is asked to supply two PDB (27) coordinate files, one denoted as the receptor and the other as the ligand. The user may upload the files from his computer or he may input the PDB code with or without PDB chain identifiers. The user is also required to supply a valid email address to which the results will be sent.

For parameter selection, the user can choose between DOT and ZDOCK as their docking program, as mentioned earlier. If the user has performed their own docking run of using ZDOCK, DOT or GRAMM (5), they may upload the PDB files associated with the job, as well as the output from the docking algorithm. The list of putative complexes supplied will then be subjected to the filtering and clustering algorithm previously described.

The user may also select the number of desolvation and electrostatic putative structures that pass through the filter, as long as the sum is equal to 2000 structures. For instance, in a system such as barnase/barstar, where electrostatics is the known driving force of interaction, the user may choose to pass the top 2000 electrostatic structures through the filter for clustering. The default, as described, is 1500/500, which is generally successful on a broad range of complexes. The clustering radius may also be varied between 3 and 10 Å. For smaller protein complexes, a smaller clustering radius is suggested because the putative structures are typically closer together, and some specificity is lost when a larger clustering radius is used. For larger complexes, it makes sense to increase the clustering radius as the putative ligand structures are typically more sparse along the surface of the receptor.

A new feature to ClusPro, inspired by Target 10 of the CAPRI experiment, is the docking of homo-multimeric structures with N-fold symmetry (S. R. Comeau and C. J. Camacho, to be published). For this feature, both the receptor and ligand structures must comprise the same sequence, and the number of monomers that constitute the multimer must be defined. Users can select whether the structure will be a dimer, trimer, tetramer or pentamer. One caveat of this feature is that it does not yet produce tetramers with D2 symmetry (e.g. structures that are dimers of dimers).

If the user has a priori information as to where binding should occur, there are also two Perl scripts available on ClusPro for pre-processing of the receptor PDB, 'block.pl' and 'attract.pl'. For example, the user may want to restrict the search on an antibody to just the CDR regions, in which case the user can use block.pl to prohibit any residues that are not part of the CDR from participating in forming putative structures. Conversely, a user may attract more ligand structures to the CDR regions by using attract.pl on the CDR residues. The difference between the two is that block.pl absolutely prohibits binding to the regions blocked, whereas attract.pl does not exclusively attract to the selected residues and putative ligands may still be found elsewhere.

CAPRI

The algorithm implemented in ClusPro has been successfully tested in CAPRI, where Camacho and Gatchell (17) were one of the three teams that predicted the best models on the first assessment (28). Since January 2003, ClusPro is participating in CAPRI as the only automated method to predict complex structures. It should be noted that in both rounds, ClusPro's submission to CAPRI was only hours after the structures were released, while human experts had over a month to perform the docking. No biochemical information was used in the docking of these structures, which is also a very strong advantage that human experts currently have over the web server. Here, we show Target 8 (Figure 2) (29) of Round 3, where ClusPro's third largest cluster center was one of the best predictions, beating many of the human experts. Using the interface RMSD

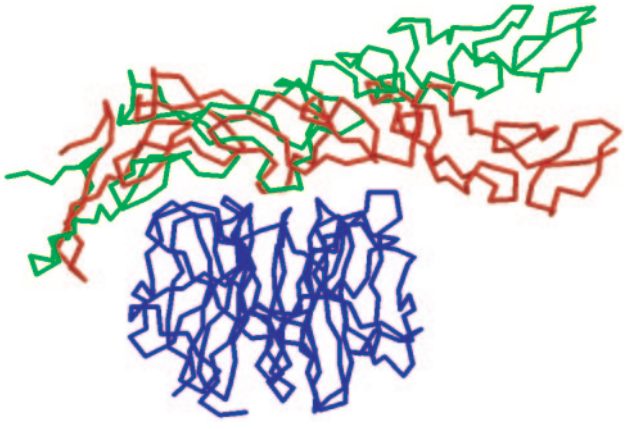

Figure 2. The Nidogen-G3/laminin complex (Target 8) of CAPRI. The laminin structure is colored in blue, the native nidogen molecule is colored in green, and ClusPro's model is colored in red.

metric, ClusPro actually outperformed all other predictions. In the more recent Round 4 of CAPRI, ClusPro's ninth model for Target 12 was also among the top performers, obtaining the best percentage of native contacts, again performing very well compared to human experts.

DISCUSSION

We describe the function of ClusPro, the first fully automated web server for the prediction of protein–protein interactions. The user may allow ClusPro to generate the putative structures, to which the filtering and clustering method is applied. The server may also be used to discriminate putative structures that have been generated by the user, using any one of the server-compatible docking algorithms. ClusPro's user interface is relatively simple; the only inputs needed from the user are two proteins known to interact and a valid email address. Also, the new Symmetry functions of ClusPro can be useful in the prediction of complex structures for homo-multimers.

ClusPro has been rather successful in the blind CAPRI experiment, where it has generated some of the best predictions for the given target structures. The performance of ClusPro may also be increased by adding other stages of discrimination and refinement. Most notably, we plan to parallelize and incorporate the SmoothDock (30) algorithm into the ClusPro algorithm, which will not only re-rank the models on properties other than cluster size, but also lower the RMSD between the near-native structure and the prediction.

ACKNOWLEDGEMENTS

We are grateful to J. C. Prasad for helping set up the server. This research has been supported by grants GM61867 from the National Institute of Health and P42 ES07381 from the National Institute of Environmental Health.

REFERENCES

1. Camacho,C.J. and Vajda,S. (2002) Protein–protein association kinetics and protein docking. *Curr. Opin. Struct. Biol.*, **12**, 36–40.
2. Halperin,I., Ma,B., Wolfson,H. and Nussinov,R. (2002) Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins*, **47**, 409–443.
3. Smith,G.R. and Sternberg,M.J.E. (2002) Prediction of protein–protein interactions by docking methods. *Curr. Opin. Struct. Biol.*, **12**, 28–35.
4. Katchalski-Katzir,E., Shariv,I., Eisenstein,M., Friesem,A., Aflalo,C. and Vakser,I.A. (1992) Molecular surface recognition—determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl Acad. Sci. USA*, **89**, 2195–2199.
5. Vakser,I.A. (1996) Low-resolution docking: prediction of complexes for underdetermined structures. *Biopolymers*, **39**, 455–464.
6. Ritchie,D.W. and Kemp,G.J.L. (2000) Protein docking using spherical polar Fourier correlations. *Proteins*, **39**, 178–194.
7. Weng,Z., Vajda,S. and DeLisi,C. (1996) Prediction of protein complexes using empirical free energy functions. *Protein Sci.*, **5**, 614–626.
8. Gabb,H.A., Jackson,R.M. and Sternberg,M.J.E. (1997) Modelling protein docking using shape complementarity, electrostatics, and biochemical information. *J. Mol. Biol.*, **272**, 106–120.
9. Jackson,R.M., Gabb,H.A. and Sternberg,M.J.E. (1998) Rapid refinement of protein interfaces incorporating solvation: Application to the docking problem. *J. Mol. Biol.*, **276**, 265–285.
10. Moont,G., Gabb,H.A. and Sternberg,M.J.E. (1999) Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, **35**, 364–373.
11. Camacho,C.J., Gatchell,D.W., Kimura,S.R. and Vajda,S. (2000) Scoring docked conformations generated by rigid-body protein–protein docking. *Proteins*, **40**, 525–537.
12. Norel,R., Sheinerman,F., Petrey,D. and Honig,B. (2001) Electrostatic contributions to protein–protein interactions: Fast energetic filters for docking and their physical basis. *Protein Sci.*, **10**, 47–61.
13. Kimura,S.R., Brower,R.C., Vajda,S. and Camacho,C.J. (2001) Dynamical view of the positions of key side chains in protein–protein recognition. *Biophys. J.*, **80**, 635–642.
14. Camacho,C.J., Weng,ZP, Vajda,S. and DeLisi,C. (1999) Free energy landscapes of encounter complexes in protein–protein association. *Biophys. J.*, **76**, 1166–1178.
15. Shortle,D., Simons,K.T. and Baker,D. (1998) Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl Acad. Sci., USA*, **95**, 11158–11162.
16. Janin,J., Henrick,K., Moult,J., Ten Eyck,L., Sternberg,M.J., Vajda,S., Vakser,I. and Wodak,S.J. (2003) CAPRI: a critical assessment of predicted interactions. *Proteins*, **52**, 2–9.
17. Camacho,C.J. and Gatchell,D. (2003) Successful discrimination of protein interactions. *Proteins*, **52**, 92–97.
18. Comeau,S.R., Gatchell,D.W., Vajda,S. and Camacho,C.J. (2004) ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, **20**, 45–50.
19. Ten Eyck,L.F., Mandell,J., Roberts,V.A. and Pique,M.E. (1995) Surveying molecular interactions with DOT. In Hayes,A. and Simmons,M. (ed.), *Proceedings of the 1995 ACM/IEEE Supercomputing Conference*. ACM Press, New York.
20. Mandell,J.G., Roberts,V.A., Pique,M.E., Kotlovyi,V., Mitchell,J.C., Nelson,E., Tsigelny,I. and Ten Eyck,L.F. (2001) Protein docking using continuum electrostatics and geometric fit. *Protein Eng.*, **14**, 105–113.
21. Chen,R. and Weng,Z. (2003) A novel shape complementarity scoring function for protein–protein docking. *Proteins*, **51**, 397–408.
22. Chen,R., Li,L. and Weng,Z. (2003) ZDOCK: an initial-stage protein docking algorithm. *Proteins*, **52**, 82–87.
23. Chakravarti,P. and Janin,J. (2002) Disecting protein–protein recognition sites. *Proteins*, **47**, 334–343.
24. Zhang,C., Cornette,J.L. and DeLisi,C. (1997) Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.*, **267**, 707–726.
25. Camacho,C.J., Kimura,S.R., DeLisi,C. and Vajda,S. (2000) Kinetics of desolvation-mediated protein–protein binding. *Biophys. J.*, **78**, 1094–1105.
26. Brooks,B.R., Bruccoleri,R.E., Olafson,B.D., States,D.J., Swaminathan,S. and Karplus,M. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187–217.
27. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
28. Mendez,R., Leplae,R., De Maria,L. and Wodak,S.J. (2003) Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins*, **52**, 51–67.
29. Takagi,J., Yang,Y., Liu,J.H., Wang,J.H. and Springer,T.A. (2003) Complex between nidogen and laminin fragments reveals a paradigmatic beta-propeller interface. *Nature*, **424**, 969–974.
30. Camacho,C.J. and Vajda,S. (2001) Protein docking along smooth association pathways. *Proc. Natl Acad. Sci., USA*, **98**, 10636–10641.