Bioinformatics 2005 (In press)

FastContact: Rapid Estimate of Contact and Binding Free Energies

Carlos J. Camacho[1*] *and Chao Zhang*[2]
[1]Department of Computational Biology,
University of Pittsburgh, 200 Lothrop St., Pittsburgh, PA 15261, USA.
[2]Plexxikom Inc., Bolivar Drive, Berkeley, CA 94710, USA.

* To whom correspondence should be addressed:

Carlos J. Camacho
Center for Computational Biology and Bioinformatics
University of Pittsburgh
200 Lothrop St.
Pittsburgh, PA 15261, USA
E-mail: ccamacho@pitt.edu

Running heading: *FastContact: Fast Estimate of Binding Free Energy*

Abstract

Summary: Interaction free energies are crucial to analyze binding propensities in proteins. Although the problem of computing binding free energies remains open, approximate estimates have become very useful for filtering potential binding complexes. We report on the implementation of a fast computational estimate of the binding free energy based on a statistically determined desolvation contact potential (Zhang et al., 1997) and Coulomb electrostatics with a distance-dependent dielectric constant (Pickersgill, 1988), and validated in the Critical Assessment of PRotein Interactions experiment. The application also reports residue contact free energies that rapidly highlight the hotspots of the interaction.

Availability: The program was written in Fortran. The executable and full documentation is freely available at http://structure.pitt.edu/software/FastContact

Contact: ccamacho@pitt.edu

Today, nearly all docking methods use some type of scoring function to differentiate between near-native complexes and non-specific encounter complexes. In the First Critical Assessment of PRedicted Interaction meeting, CAPRI, see Janin et al. (2003), computational scoring functions involved free energy-like terms adjusted by free parameters that optimized the discrimination of bound crystal structures (Fernandez-Recio et al., 2003; Gray et al., 2003; Ritchie, 2003; Smith et al., 2003) or more geometrical discriminators, say, buried surface area (Gardiner et al., 2003; Krippahl et al., 2003; Law et al., 2003; Schneidman-Duhovny et al., 2003), or hybrids of these two approaches (Ben-Zeev et al., 2003; Chen et al., 2003).

At the same time, in the literature one finds several more sophisticated, and perhaps more accurate, approaches to estimate different free energy contributions – e.g., free energy perturbation (Kollman, 1993), Poisson-Boltzman (Honig and Nicholls, 1995), atomic continuum electrostatic (Schaefer and Karplus, 1996), and generalized-Born solvation (see, e.g, Qiu et al., 1997). However, since protein docking requires filtering or sampling millions of plausible complex structures and these methods are computationally expensive, they are not used for free energy screening.

Finally, a somewhat different approach to screen protein binding interactions has been developed by Camacho et al. (2000; 2003). These authors use a free energy scoring function developed independently of the bound crystal structures present in the PDB (Berman et al., 2000). Namely, the interaction between two proteins is estimated as ΔG_{bind}, where

$$\Delta G_{bind} = \Delta E_{elec} + \Delta G_{des} \; . \qquad [1]$$

ΔE_{elec} corresponds to the standard intermolecular Coulombic electrostatic potential with a distance-dependent dielectric constant equal to $4r$ (Pickersgill, 1988). ΔG_{des} captures the most essential features of the desolvation free energy in proteins, including hydrophobic interactions, the self-energy change upon desolvating charge on polar atom groups, and side-chain entropy loss. ΔG_{des} is calculated by an empirical contact potential of the form $\Delta G_{des} = g(r) \sum \sum e_{ij}$, where e_{ij} denotes the atomic contact potential (ACP) between atom i of the receptor and j of the ligand. The double sum is taken over all atom pairs and $g(r)$ is 0 for atoms that are more than 7 Å apart, 1 if less than 5 Å apart and in between $g(r)$ is a smooth function varying between these two limits (Zhang et al., 1997). The ACPs have been defined for a total of 18 atom types, and obtained from a diverse set of close to 90 protein structures by converting frequencies of structural factors into atom-atom contacts.

This free energy estimates reasonably well experimental binding affinities from complex crystal structures (Zhang et al., 1997; Zhang et al, 1997a; Kimura et al., 2001). However, filtering decoys with less than optimal side chain packing and structural/charge overlap is not as straightforward. Two problems are the sensitivity of the electrostatic energy to charge overlaps, and the overextended contribution of the desolvation term arising from overlapping contacts that are not at the protein surface. It is worth mentioning that although one could easily remove most overlaps by energy minimization, this is not computationally feasible for a million or so structures. We address these problems by first not allowing two atoms to be any closer than the sum of their van der Waals radii, preventing artificial spikes on the electrostatic term (Vasmatzis et al. 1996; Zhang et al 1999); and second, we provide an option to always require that at least one of the

interacting atoms be exposed to solvent by at least 1 $Å^2$ in the unbound state (Camacho et al., 1999). The latter is done by computing the solvent accessible surface area of each individual protein using Lee and Richards (1971) algorithm. It is worth mentioning that the problem of over counting desolvation contact energies is worst when the receptor-ligand overlap is more than around 300 $Å^3$, for minimum overlap the range of the contact potential is sufficient to constraint the interactions within the surfaces.

This free energy was the main filter of potential binding sites used by Camacho and Gatchell (2003) in the first CAPRI experiment. These authors produced some of the best predictions at CAPRI1-2 (Mendez et al., 2003), appropriately ranking the native-like models. We have also implemented our method as a fully automated public server named *ClusPro* (Comeau et al., 2004). *ClusPro* was the only server validated in the second CAPRI meeting (Gaeta, Italy, 2004), where for 5 (out of 10) targets native-like structures were submitted. Moreover, for two of the targets, the models predicted using Eq. [1] were some of the most accurate among all submissions (Comeau, Vajda and Camacho, Proteins 2005). It is also worth mentioning that, after adding the van der Waals interactions and Eq. [1] to the scoring function, the native-like structures submitted after flexible refinement were also discriminated (Camacho, Proteins, 2005).

In order to share this utility with the research community, we have implemented this fast scoring function in a program called ***FastContact***. The input of *FastContact* is as follow:

FastContact *RTF receptor.pdb ligand.pdb Num_extra_ligands Contacts SASA* [2]

The *RTF_file* defines the united atom composition of each amino acid and it is provided together with the executable. The RTF file includes a list of residue types and their atomic make up, partial charges and van der Waals radii; the data is consistent with CHARMM19 parameters. The user is free to modify this file provided the format of the data remains the same. The *proteins* should be in standard CHARMM (Brooks et al., 1983) or CONGEN (Bruccoleri et al., 1997) format with polar hydrogens only. Since in rigid-body docking one often is interested in scoring several ligand conformations against the same receptor, we provide an option that allows computing the binding free energy for as many extra ligands as needed. The program will read from standard input *Num_extra_ligands* file names of new ligand structures. The main output of the program is directed to the screen and consists on the total electrostatic and desolvation energy. If *Contacts* $\neq 1$, the output details the top 20 residues that have the minimum and maximum contribution to the different free energy components; the residues are renumbered starting with number 1 and the program creates two PDB files, *fort.19* and *fort.20*, with the new numbers. If *Contacts* = 1 no contact energy information is produced. *SASA* $\neq 1$ will check that at least one of the contact residues is at the surface. If this constraint is not deemed necessary then *SASA* = 1.

Computing the solvent accessible surface area (*SASA* $\neq 1$) is the most computational expensive step of the algorithm. Using a single Pentium 4 processor, *FastContact* takes less than 0.1 seconds to compute ΔG_{bind} for two single domain proteins and *SASA* = 1, and 3 seconds if *SASA* $\neq 1$. However, once SASA is computed for one receptor and ligand, extra runs using different orientations of the same ligand structure take less than 0.1 seconds. The maximum number of residues is 1500. If the residue name is not in the

RTF file the program stops; if an atom is not in the RTF then its contribution is made equal to zero and a warning message is spooled to the screen.

The contact information (*Contacts* ≠ 1) is very useful in model refinement of rigid-body docked conformations because in the output list one can read the residues and pair of residues that provide both the most attractive and repulsive free energy. While the former immediately highlights the hot spots of the binding interaction, the latter often suggest side chains that might need to be refined. Also, the residue contact free energies should prove useful in selecting interesting residues for mutagenesis experiments.

FastContact provides a fast estimate of the interaction free energy between two proteins. Because it is based on folding data, the estimate is robust and does not required to be re-parameterized as more complex structures become available. More importantly, as far as we know, it is the only scoring function validated in CAPRI that is made available to the community at large. *FastContact* can now be combined with the user favorite decoys generator and other scoring functions to further refine predictions of complex structures.

Acknowledgments

We are grateful to Sandor Vajda, Charles DeLisi and Zhiping Weng for their help and support while the authors were at Boston University. CJC is grateful for the support of the University of Pittsburgh. We are also thankful to Christoph Champ for setting up the link to download. ACP and ACP-based binding energy function was developed by C. Zhang in collaboration with Drs. J. Cornette and G. Vasmatzis while working at Professor Charles DeLisi's laboratory.

References

Ben-Zeev,E., Berchanski,A., Heifetz,A., Shapira,B. and Eisenstein,M. (2003) Prediction of the unknown: Inspiring experience with the CAPRI experiment. *Proteins*, **52**, 41-46.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., P.E. Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235-242.

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 dynamic calculations. *J. Comput. Chem.,* **4**, 187-217.

Bruccoleri,R.E., Novotny,J., Davis,M. and Sharp,K.A. (1997) Finite difference Poisson-Boltzmann electrostatic calculations: increased accuracy achieved by harmonic dielectric smoothing and charge antialiasing. *J. Comp. Chem.,* **18**, 268-276.

Camacho,C.J. and Gatchell,D. (2003) Successful discrimination of protein interactions. *Proteins,* **52**, 92-97.

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.

Camacho,C.J., Weng,Z., Vajda,S. and DeLisi,C. (1999) Free energy landscapes of encounter complexes in protein-protein association. *Biophys. J.,* **76**, 1166-1178.

Chen,R., Li,L. and Weng,Z. (2003) ZDOCK: An initial-stage protein-docking algorithm. *Proteins*, **52**, 80-87.

Comeau,S.R., Gatchell,D., Vajda,S. and Camacho,C.J. (2004) ClusPro: An automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, **20**, 45-50.

Dunbrack, R., and Cohen, F. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci*. **6**, 1661-1681.

Fernández-Recio,J., Totrov,M. and Abagyan,R. (2003) ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins*, **52**, 113-117.

Gardiner,E.J., Willett,P. and Artymiuk,P.J. (2003) GAPDOCK: A genetic algorithm approach to protein docking in CAPRI round 1. *Proteins*, **52**, 10-14.

Gray,J.J., Moughon,S.E., Kortemme,T., Schueler-Furman,O., Misura,K.M.S., Morozov,A.V. and Baker,D. (2003) Protein-protein docking predictions for the CAPRI experiment. *Proteins*, **52**, 118-122.

Honig,B. and Nicholls,A. (1995) Classical electrostatics in biology and chemistry. *Science.* **268**, 1144-1149.

Janin,J., Henrick,K., Moult,J., Ten Eyck,L., Sternberg,M.J.E., Vajda,S., Vakser,I. and Wodak,S.J. (2003) CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins*, **52**, 2-9.

Kimura,S.R., Brower,R., 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.

Kollman,P.A. (1993) Free energy calculations: applications to chemical and biochemical phenomena. *Chem. Rev.*, **93**, 2395-2417.

Krippahl,L., Moura,J.J, Palma,P.N. (2003) Modeling protein complexes with BiGGER. *Proteins*, **52**, 19-23.

Law,D.S., Ten Eyck,L.F., Katzenelson,O., Tsigelny,I., Roberts,V.A., Pique,M.E. and Mitchell,J.C. (2003) Finding needles in haystacks: Reranking DOT results by using shape complementarity, cluster analysis, and biological information. *Proteins*, **52**, 33-40.

Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.,* **55**, 379-400

Méndez,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.

Pickersgill,R.W. (1988) A rapid method of calculating charge-charge interaction energies in proteins. *Protein Eng*. **2**, 247-248.

Qiu,D., Shenkin,P.S., Hollinger,F.P. and Still,W.C. (1997) The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A.,* **101**, 3005-3014.

Ritchie,D.W. (2003) Evaluation of protein docking predictions using *Hex* 3.1 in CAPRI rounds 1 and 2. *Proteins*, **52**, 98-106.

Schneidman-Duhovny,D., Inbar,Y., Polak,V., Shatsky,M., Halperin,I., Benyamini,H., Barzilai,A., Dror,O., Haspel,N., Nussinov,R. and Wolfson,H.J. (2003) Taking geometry to its edge: Fast unbound rigid (and hinge-bent) docking. *Proteins*, **52**, 107-112.

Smith,G.R. and Sternberg,M.J.E. (2003) Evaluation of the 3D-Dock protein docking suite in rounds 1 and 2 of the CAPRI blind trial. *Proteins*, **52**, 74-79.

Vasmatzis,G., Zhang,C., Cornette,J.L. and DeLisi C. (1996) Computational determination of side chain specificity for pockets in class I MHC molecules. *Mol. Immunol*., **33**, 1231-9.

Zhang,C., Chen,J. and DeLisi,C. (1999) Protein-protein recognition: exploring the energy funnels near the binding sites. *Proteins*, **34**, 255-67.

Zhang,C., Cornette,J.L. and DeLisi C. (1997a) Consistency in structural energetics of protein folding and peptide recognition**.** *Protein Sci*., **6**, 1057-1064.

Zhang,C., Vasmatzis,G., Cornette,J.L. and DeLisi,C. (1997) Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.,* **267**, 707-726.