BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 21 2010, pages 2784–2785
doi:10.1093/bioinformatics/btq504

Structural bioinformatics

Advance Access publication September 24, 2010

SimiCon: a web tool for protein–ligand model comparison through calculation of equivalent atomic contacts

Manuel Rueda[1], Vsevolod Katritch[1], Eugene Raush[2] and Ruben Abagyan[1,2,*]

[1]UCSD Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093 and [2]Molsoft, LLC, 3366 North Torrey Pines Court, Suite 300, La Jolla, CA 92037, USA

Associate Editor: Anna Tramontano

ABSTRACT

Summary: SimiCon is a web server designed for an automated identification of equivalent protein–ligand atomic contacts in different conformational models of a complex. The contacts are computed with internal coordinate mechanics (ICM) software with respect to molecular symmetry and the results are shown in the browser as text, tables and interactive 3D graphics. The web server can be executed remotely without a browser to allow users to automate multiple calculations.

Availability: SimiCon is freely available at http://abagyan.ucsd.edu/SimiCon

Contact: rabagyan@ucsd.edu

Received on June 18, 2010; revised on July 26, 2010; accepted on August 29, 2010

1 INTRODUCTION

Docking and virtual ligand screening with crystal structures or homology models play important roles in the understanding of receptor–ligand interactions, with numerous applications in drug discovery. A critical requirement for any useful application of docking algorithms is their ability to find accurate poses of the ligand and its contacts with the receptor atoms.

The validation of the geometry of a complex or 'pose-prediction' is based on the ability to reproduce the binding mode of a ligand observed in the cognate reference X-ray or NMR structure. The precision of the models is usually checked with the root mean squared deviation (RMSD), a popular measure that accounts for the distances of the ligand atoms of the model from the ligand atoms in the reference structure. To provide an example, in small molecule docking context, a heavy-atom RMSD $\leq$2 Å with respect to the reference pose is widely accepted as good (Cole *et al*., 2005).

RMSD, however, has shortcomings that can lead to misclassification of both correct and incorrect poses (Cole *et al*., 2005; Kroemer *et al*., 2004). For instance, although a low RMSD value is indicative of a strong similarity, it does not show how well the critical interactions are conserved. In this regard, literature contains numerous examples where hydrogen bond interaction patterns differ between reference and model, despite having low RMSD values (Kroemer *et al*., 2004). Conversely, a high RMSD might come from differences in a flexible region of the ligand (not important for the overall binding mode), or from a nearly correct pose of a symmetric molecule (Abagyan and Marsden 2003; Kroemer *et al*., 2004). Another major flaw of the RMSD is that it requires receptor superposition, which is rather ambiguous when models represent distinct experimental or modeled conformations of the receptor.

As an alternative to the RMSD, interaction-based measures reflect more adequately important aspects of the molecular recognition by measuring conserved key protein–ligand contacts. In small molecule context, attempts of standardization of contact-based measures have been made before (Abagyan and Totrov 1997; Cole *et al*., 2005; Deng *et al*., 2004; Hawkins *et al*., 2008; Kroemer *et al*., 2004; Marcou and Rognan, 2007); however, some technical caveats hampered their wide acceptance. For instance, one problematic issue is the assignation of the atomic equivalence between atoms of the two complexes. Thus, while Cα/Cβ contact-based metrics are widely employed for evaluation of protein–protein docking model quality in CAPRI assessment (Janin *et al*., 2003), establishing atom equivalence for chemical compounds, on which the equivalence does not follow the amino acid sequence, is non-trivial. For small molecules, equivalence can be established using a specific set of rules for unique enumeration of atoms in the chemical structure, such as in IUPAC, unique SMILES and chirality rules of Cahn–Ingold–Prelog (Sidney *et al*., 1966). Such rules are implemented in some chemical packages, but often they are not in 3D molecular modeling software. Most importantly, the presence of any symmetry elements in a compound also adds complexity to the equivalence problem, as the same exact contact can be made by any of the symmetry-related atoms.

To our knowledge, there is no universal solution to overcome the contact-based measure caveats. Some researchers try looking for atomic equivalence manually; others end up transforming the formats to their favorite package-specific topologies, aiming at generating consensus naming. According to our experience, even the 'package-specific-naming' solution may lead to different naming schemes if the atom names were dissimilar before the conversion, or if the ligands come from SMILES representations. To avoid these difficulties in assigning equivalent interactions, docking assessments (Michino *et al*., 2009) required modelers to submit data in a rigid protein data bank (PDB) template format with defined names and sequential positions of atoms. Unfortunately, analysis of models delivered in even such rigid formats has not been fully automated and requires manual processing to take into account symmetry both in the receptor side chains and in many cases in the small molecule ligands.

Here, we present a solution to this issue based on a web server where the atomic equivalence and symmetry of the receptor and the

*To whom correspondence should be addressed.

ligand are taken into account. The method can be applied not only for protein–ligand docking pose ranking, but also for the assessment of any protein–chemical complex such as, but not limited to, co-factor or drug positioning in multi-chain proteins and community-wide structure assessments of modeled complexes.

2 METHODS AND IMPLEMENTATION

SimiCon web interface was written in Perl using the common gateway interface (CGI) module. The core calculation of the intersection of interatomic heavy-atom contacts is performed with internal coordinate mechanics (ICM) 3.7 molecular modeling and docking software (Abagyan and Totrov, 1994), which has unique enumeration rules accurately implemented and tested in previous applications. An automated ICM script uses the following steps to calculate equivalent contacts (*EC*) in *reference* and *target* models: (i) all receptor–ligand atomic contacts are identified for the reference model using selected cutoff distance, typically 4 Å; (ii) for the set of contact atoms in the reference model, equivalent atoms in the target model are identified. If any atom of ligand has a symmetric atom (e.g. in phenyl group), all possible equivalence schemes are calculated. For symmetric side chains in receptor (i.e. Arg, Phe, Tyr, Val, Leu, Asp and Glu) equivalent atoms are enumerated only within a side chain to avoid large combinatorics; (iii) atomic contacts of the target model, equivalent with contacts in the reference are calculated; and (iv) lists of *EC*, as well as all contacts in reference model (*RC*) and target contacts (*TC*) model are calculated. These lists can be used to calculate some general metrics, such as the coverage = *EC/RC* and the accuracy = *EC/TC* (Janin *et al*., 2003). Results can be obtained through a web browser, or remotely by the execution of a script using Perl LWP library (see Help page at web site).

2.1 Input data

The user can upload PDB coordinate files or retrieve the structure using PDB code (Berman *et al*., 2000). The server is optimized for parsing proteins as receptors and chemicals as ligands (labeled as HETATMs in PDB). The recommended standard cutoff distance for the atomic contacts is 4 Å, but the user can choose a range from 2 to 12 Å.

2.2 Results and visualization

For most PDB complexes the calculation takes 2 s and is presented in plain HTML text, tables and 3D interactive molecular objects. The 3D interactive objects can be visualized online by using the activeICM/active X plugin (Raush *et al*., 2009) or be downloaded as a single file to be browsed with all its attached objects locally with the ICM browser. Both activeICM and ICM browser are freely available to the public. The interatomic contacts for the reference, the target and the intersection can be downloaded as a comma separated value (.csv) file, compatible and supported by almost all spreadsheets and database management systems.

Apart from a standalone calculation, we envision that some users may wish to use the *command line* to execute the CGI with multiple targets. For that purpose, we provide an example Perl script implemented using the LWP library that will avoid the necessity for 'screen scraping' of HTML. Full instructions are provided on the Help page at web site.

The server also contains four pre-computed examples, showing different scenarios where SimiCom can be applied: (i) Heme contacts with chains A and B of human cytocrom P450 2D6 (PDB:2f9q); (ii) best overall *in silico* model of the Adenosine A_{2A} receptor submitted to a recent community-wide G protein-coupled receptor (GPCR) assessment (PDB:3eml); iii) ATP contacts with chains A and B of Human Pyridoxal Kinase (PDB:2yxu); and (iv) staurosporine contacts with epidermal growth factor receptor (EGFR) Kinase domain in complex with AFNN941 protein kinase in wild-type and mutated forms (PDBs:2itw, 2itu).

ACKNOWLEDGEMENTS

We thank Karie Wright for help with manuscript preparation.

Funding: Marie Curie OIF Fellowship of the European Commission (M.R.); National Institutes of Health (grants R01GM071872 and R01GM074832 to V.K. and R.A.).

Conflict of Interest: none declared.

REFERENCES

Abagyan,R. and Marsden,B.D. (2003) Identifying errors in three dimensional protein models. In Chasman,D. (ed.) *Protein Structure: Determination, Analysis, and Applications for Drug Discovery*. CRC Press, Cambridge, MA, pp. 277–314.

Abagyan,R. and Totrov,M. (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, **235**, 983–1002.

Abagyan,R.A. and Totrov,M.M. (1997) Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J. Mol. Biol.*, **268**, 678–685.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Hawkins,P.C.D. *et al*. (2008) How to do an evaluation: pitfalls and traps. *J. Comput. Aided Mol. Des.*, **22**, 179–190.

Cole,J. *et al*. (2005) Comparing protein-ligand docking programs is difficult. *Prot. Struct. Funct. Bioinform.*, **60**, 325–332.

Deng,Z. *et al*. (2004) Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.*, **47**, 337–344.

Janin,J. *et al*. (2003) CAPRI: a critical assessment of PRedicted Interactions. *Proteins*, **52**, 2–9.

Kroemer,R.T. *et al*. (2004) Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.*, **44**, 871–881.

Marcou,G. and Rognan,D. (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.*, **47**, 195–207.

Michino,M. *et al*. (2009) Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nat. Rev. Drug Discov.*, **8**, 455–463.

Raush,E. *et al*. (2009) A new method for publishing three-dimensional content. *PLoS One*, **4**, e7394.

Sidney,R. *et al*. (1966) Specification of molecular chirality. *Angew. Chemie Int. Ed.*, **5**, 385–415.