Computational Biology and Chemistry 64 (2016) 403–413

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Assessing the similarity of ligand binding conformations with the Contact Mode Score

Yun Ding[a], Ye Fang[b,d], Juana Moreno[a,d], J. Ramanujam[b,d], Mark Jarrell[a,d], Michal Brylinski[c,d,*]

[a] Department of Physics and Astronomy, Louisiana State University, Baton Rouge, LA 70803, USA
[b] School of Electrical Engineering and Computer Science, Louisiana State University, Baton Rouge, LA 70803, USA
[c] Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA
[d] Center for Computation & Technology, Louisiana State University, Baton Rouge, LA 70803, USA

ARTICLE INFO

Article history:
Received 20 May 2016
Received in revised form 17 August 2016
Accepted 25 August 2016
Available online 6 September 2016

Keywords:
Contact Mode Score
CMS
eXtended Contact Mode Score
XCMS
Root-mean-square deviation
RMSD
Molecular docking
Ligand binding conformations

ABSTRACT

Structural and computational biologists often need to measure the similarity of ligand binding conformations. The commonly used root-mean-square deviation (RMSD) is not only ligand-size dependent, but also may fail to capture biologically meaningful binding features. To address these issues, we developed the Contact Mode Score (CMS), a new metric to assess the conformational similarity based on intermolecular protein-ligand contacts. The CMS is less dependent on the ligand size and has the ability to include flexible receptors. In order to effectively compare binding poses of non-identical ligands bound to different proteins, we further developed the eXtended Contact Mode Score (XCMS). We believe that CMS and XCMS provide a meaningful assessment of the similarity of ligand binding conformations. CMS and XCMS are freely available at http://brylinski.cct.lsu.edu/content/contact-mode-score and http://geaux-computational-bio.github.io/contact-mode-score/.

1. Introduction

Molecular docking is a computational technique routinely used in protein function analysis and drug discovery (Cheng et al., 2012; Yuriev et al., 2015). Docking calculations usually consist of two successive stages, the prediction of the favorable orientation of a small molecule when bound to its target protein followed by the estimation of binding affinity and/or free energy of binding. Scoring functions widely used in molecular docking evaluate protein-ligand conformations in terms of the shape and electrostatic complementarity, as well as the presence of stabilizing interactions such as hydrogen bonds, salt bridges, and hydrophobic contacts (Yusuf et al., 2008). Since these factors hinge on the ligand binding mode, accurately predicted protein-ligand conformations are required for meaningful scoring.

A common practice in benchmarking docking programs is to evaluate predicted conformations against experimentally solved complex structures using the root-mean-square deviation (RMSD) (Kabsch, 1978). Typically, predictions within an RMSD of 2 Å are considered successful, whereas values higher than 3 Å indicate docking failures. A standard RMSD function quantifying the difference between two poses of the same molecule is computed as follows:

$$RMSD(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \| a_i - b_i \|^2} \quad (1)$$

where molecule poses $A=\{a_1, a_2, \ldots\ldots, a_n\}$ and $B=\{a_1, a_2, \ldots\ldots, a_n\}$ are defined by sets of Cartesian coordinates a_i and b_i of individual heavy (non-hydrogen) atoms. This formulation shows that the RMSD is calculated based on a predefined one-to-one correspondence between atoms in poses *A* and *B*. Although equivalent atoms can be found by matching atom indices, the presence of symmetric functional groups may result in inflated RMSD values (Allen and Rizzo, 2014). Several modified RMSD calculation methods were developed to handle symmetric molecules (Allen and Rizzo, 2014; Trott and Olson, 2010). These techniques re-index atoms dynamically instead of using the predefined order of atoms.

Further, a strong dependence of the RMSD on the number of atoms complicates the assessment of molecules with different sizes (Reva et al., 1998; Stark et al., 2003). On the other hand, the

* Corresponding author at: Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA.
E-mail addresses: yding8@lsu.edu (Y. Ding), yfang11@lsu.edu (Y. Fang), moreno@phys.lsu.edu (J. Moreno), ram@cct.lsu.edu (J. Ramanujam), jarrellphysics@gmail.com (M. Jarrell), michal@brylinski.org (M. Brylinski).

http://dx.doi.org/10.1016/j.compbiolchem.2016.08.007
1476-9271/

development and optimization of scoring functions for molecular docking often involves tuning force field parameters against diverse datasets of protein-ligand complexes. For example, weight factors can be adjusted to maximize the capability to recognize near native conformations amongst a large set of docking decoys (Brylinski and Skolnick, 2009a, 2008; Ding et al., 2015). An imprecise classification of near native and decoy conformations, e.g. by using a fixed RMSD threshold, may lead to suboptimal weight factors. Even though the number of ligand atoms can be taken into account by calculating the statistical significance of RMSD values (Reva et al., 1998; Stark et al., 2003), statistical testing is rarely employed in the development and optimization of docking algorithms and scoring functions.

Another issue is that ligand RMSD does not account for the protein environment (Kroemer et al., 2004). Depending on the ligand size and complexity, low RMSD values can be obtained even if key interactions with the protein are absent. Conversely, a substantial deviation from the experimental structure of a moiety that is irrelevant to binding (e.g., a solvent-exposed group) can notably increase the RMSD even when crucial binding features are recovered by docking calculations (Yusuf et al., 2008). To address this problem, the relative displacement error (RDE) (Abagyan and Totrov, 1997) was developed. The RDE down-weights large deviations, therefore, it is less sensitive to a small number of misplaced atoms compared to the RMSD. Nevertheless, similar to RMSD, the RDE takes no account of the protein environment.

Although conventional docking methods employ a single, static structure of the receptor, more recent approaches incorporate protein flexibility by docking against protein ensembles or using rotamer libraries for binding residue side chains (Chang et al., 2007; Lill, 2011; Meiler and Baker, 2006). The traditional ligand RMSD cannot be used to assess the accuracy of fully flexible molecular docking, where not only ligands, but also receptors change their internal conformations. For that reason, an alternative measure based on real space R-factors was proposed to compare electron density rather than to calculate the RMSD from Cartesian coordinates (Yusuf et al., 2008). Moreover, predicted binding modes can be visually inspected in order to identify key protein-ligand interactions recovered by docking calculations (Kroemer et al., 2004). However, the lack of automation makes this approach inapplicable to large datasets of docked ligand conformations.

The calculation of RMSD is straightforward and has a low computational complexity, therefore, it is still frequently used as the assessment measure, particularly across large datasets of protein-ligand complexes. Nevertheless, new techniques are highly desired to evaluate not only purely geometrical features, but also biological aspects of binding. On that account, we developed the Contact Mode Score (CMS), which effectively quantifies the similarity of ligand binding conformations. CMS compares the sets of interatomic contacts formed by a ligand and its receptor rather than ligand Cartesian coordinates. Such an approach also allows for the protein environment to be included in the assessment. Further, we developed the eXtended Contact Mode Score (XCMS), which provides a convenient template-based method to compare those protein-ligand complexes composed of different proteins and non-identical ligands. In contrast to the RMSD, CMS and XCMS are less dependent on the ligand size and have a well-defined statistical significance.

2. Materials and methods

2.1. Experimental datasets

Three datasets of protein-ligand complexes are used in this study. The first dataset was compiled from the *e*FindSite library (Brylinski and Feinstein, 2013) by clustering template proteins at 40% sequence identity using PISCES (Wang and Dunbrack, 2003), and then selecting representative chains that non-covalently bind small organic molecules at distinct locations. This procedure produced a set of 14 059 non-redundant structures of protein-ligand complexes, referred to as the *e*FindSite dataset, which was used to develop a mixed-resolution model of complex structures. In addition, we used the Astex/CCDC dataset (Nissink et al., 2002) comprising the high-quality experimental structures of 201 pharmacologically relevant proteins co-crystalized with drug molecules. The dependence of CMS and RMSD on the number of ligand atoms was examined against the Astex/CCDC dataset. Finally, the XCMS was developed and tested on the BioLiP database (Yang et al., 2013). BioLiP provides a comprehensive collection of protein-ligand complex structures curated specifically for studies focusing on biologically relevant interactions and template-based modeling approaches. From the entire database comprising 94 887 ligands bound to 71 359 proteins, we randomly selected 2200 protein-ligand complexes as query structures. In XCMS benchmarking, we searched the complete BioLiP database for non-identical templates for each query structure. A complex was used as the template if the Pocket Similarity score (PS-score) against the query pocket is <0.9, the fingerprint Tanimoto coefficient (1D-TC) against the query ligand is >0.5, and the number of ligand heavy atoms is greater than 6. Using these criteria produced a dataset of 802 058 query-template pairs to benchmark the XCMS. The PS-score measures the structural similarity of two ligand binding sites; it ranges from 0 to 1 with higher values indicating higher similarity (Gao and Skolnick, 2013a). 1D-TC employs 1024-bit molecular fingerprints to quantify the chemical similarity of two small molecules. The calculations of 1D-TC were conducted with OpenBabel (O'Boyle et al., 2011), which supports fingerprint indexing to accelerate searches against large databases.

2.2. Simulated datasets

In addition to experimental datasets, three sets of computer-generated structures were compiled for benchmarking purposes. The first simulated dataset is based on Astex/CCDC (Nissink et al., 2002) and it was prepared to assess the dependence of RMSD and CMS on the number of ligand heavy atoms. A series of systematic perturbations were applied to co-crystalized ligands, each comprising random translations and rotations about the x, y and z-axis of up to 0.02 Å and 5°, respectively. After each round of perturbation, RMSD and CMS were computed against the native conformation of a ligand. The second simulated dataset contains Metropolis Monte Carlo (MMC) trajectories constructed by GeauxDock (Ding et al., 2015) for Astex/CCDC complexes. GeauxDock employs a mixed-resolution representation of protein-ligand complexes and a hybrid scoring function comprising physics-, evolution-based energy terms and statistical potentials. GeauxDock effectively finds the near native structures of protein-ligand complexes by exploring low-energy configurations according to a dimensionless scoring function. Here, binding ligands were initialized at random conformations and GeauxDock simulation engine (Ding et al., 2015) was used to generate docking trajectories through 800 MMC cycles. The CMS was calculated for each accepted conformation against the ligand bound in the crystal complex structure.

The last simulated dataset was built on BioLiP (Yang et al., 2013) to benchmark RMSD, CMS and XCMS using predicted and random ligand conformations. First, query ligands were randomized within receptor binding pockets to produce a set of 2200 random conformations of query ligands. Subsequently, each randomized ligand was re-docked to the protein with AutoDock Vina (Trott and Olson, 2010). The docking box was set to an optimal size based on the radius of gyration of the ligand (Feinstein and Brylinski, 2015)