Computational Biology and Chemistry 49 (2014) 1–6

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Collective variable driven molecular dynamics to improve protein–protein docking scoring

Diego Masone[a,*], Solène Grosdidier[b]

[a] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Instituto de Ciencias Básicas (ICB), Universidad Nacional de Cuyo (UNCUYO), Mendoza, Argentina
[b] Integrative Biomedical Informatics Group, Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute (IMIM), DCEXS, Universitat Pompeu Fabra, Barcelona, Spain

ARTICLE INFO

Article history:
Received 22 October 2013
Received in revised form
28 December 2013
Accepted 28 December 2013
Available online 18 January 2014

Keywords:
Protein–protein
Docking
Scoring
Collective variable
Molecular dynamics

ABSTRACT

In biophysics, the structural prediction of protein–protein complexes starting from the unbound form of the two interacting monomers is a major difficulty. Although current computational docking protocols are able to generate near-native solutions in a reasonable time, the problem of identifying near-native conformations from a pool of solutions remains very challenging. In this study, we use molecular dynamics simulations driven by a collective reaction coordinate to optimize full hydrogen bond networks in a set of protein–protein docking solutions. The collective coordinate biases the system to maximize the formation of hydrogen bonds at the protein–protein interface as well as all over the structure. The reaction coordinate is therefore a measure for docking poses affinity and hence is used as scoring function to identify near-native conformations.

1. Introduction

As most biological processes involve macromolecular complexes, identifying and characterizing protein interactions as well as the network they create, is critical for understanding the molecular mechanisms within the cell. Protein–protein interactions are fundamental in most cellular processes as for example DNA replication or signal transduction. Knowledge of the structure and properties of protein–protein complexes is essential to understand how proteins function within the cell, in order to identify new targets for therapeutic applications and develop new approaches for drug discovery. Hence, solving the structure of protein–protein complexes might provide the basis for understanding how a biological signal is transmitted or how a biological function is performed (Smith and Sternberg, 2002).

Many docking algorithms rely on the divide-and-conquer strategy (Luo et al., 2010): first, an initial sampling of the configurational space of the interacting proteins is performed by an efficient algorithm specialized on generating docking candidates (or solutions), typically rigid-body or coarse-grained, based and optimized through Fast Fourier Transform (FFT). Second, a scoring step is performed to rank the candidates according to scoring functions that can offer different levels of complexity (Halperin et al., 2002). However, the arduousness of sampling and scoring is not equal, while modern supercomputers allow for very good sampling of the configurational space between two proteins, there are no efficient and accurate methods for refinement and scoring yet. Consequently, even though near-native poses could be generated, it is still extremely difficult to distinguish them from a pool of solutions, making docking protocols to produce significant amounts of false positives (Gabb et al., 1997; Chen et al., 2003). Therefore, efficient computational docking highly depends on the accuracy of the energy functions used to evaluate the strength of binding candidates.

Unfortunately, because of bad protein modeling (missing ions, heteroatoms or unrealistic inter-protein contacts generated by rigid-body docking) scoring is currently a serious issue. The energy function used for scoring analyzes the conformation of both proteins in complex molecule and outputs a value representing a total energy. This number is meaningless alone, but when used in a relative way to compare the evolution of the total energy along the conformational changes that the proteins suffered, it gives an idea of how stable the different docking conformations are. Then, computational studies of protein associations from an energetic point of view, are also important to comprehend their essential principles and thus to improve protein interaction modeling. Energetic landscapes represent at different smoothness the natural behavior of protein–protein interactions. Their knowledge

* Corresponding author at: Padre Jorge Contreras 1300, Parque General San Martín, Mendoza 5500, Argentina. Tel.: +54 261 4236003.
E-mail addresses: dmasone@fing.uncu.edu.ar, diegomasone@yahoo.com.ar (D. Masone).

1476-9271/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2013.12.003

allow to predict stable configurational and conformational spaces, such as the binding funnel. Unfortunately for mathematical simplifications, protein binding energy landscapes are extremely rugged surfaces (Camacho and Vajda, 2001; Hagler et al., 1974a; Hagler and Lifson, 1974b). Since interacting forces upon protein binding act at different length-scales, the spatial frequency spectrum shows rapid variations and though generates many wells of local minima. For this reason, many times search algorithms are confined to a small portion of the sampling space (Phillips et al., 2001), since characterizing every minima of the potential energy landscape of a geometry is a problem with tens to thousands of degrees of freedom (Crescenzi et al., 1998; Atkins and Hart, 1999; Calland, 2003). However, the energy funnel shape is generally observed in the vicinities of the bound form of protein–protein complexes, once optimal interface residue conformations have been reached (Camacho and Vajda, 2001). In general, to test the performance of force-fields in docking applications, the native complex and the lowest energy near-native solution generated by a docking approach are compared energetically (Dixon, 1997; Totrov and Abagyan, 1997). However, this procedure is artificial and may lead to incorrect conclusions (Verkhivker et al., 2000). In a complete docking protocol the possibility to correctly identify near-native conformations as the lowest energy ones stands on the ability of the sampling process to generate them (Diller and Verlinde, 1999). The gradually narrowing of the energetic landscape has a physical meaning in protein–protein docking: there are large amounts of unfavorable high energy poses while only a small part of the configurational space is energetically stable.

Among macromolecular biological interactions hydrogen bond networks are specially interesting due to their key role in protein 3D structure, and as a consequence, in molecular recognition specificity and protein function (Morozov et al., 2004). Thus the appropriate description of their energetics is of great interest in the fields of protein–protein docking and protein folding. In computer simulations the challenging problem emerging from using crystallographic protein structure from the PDB (www.pdb.org) is to position every hydrogen atom. Indeed, the optimization of full hydrogen bond networks requires force-fields to properly describe every possible hydrogen configurations and their interaction energies (Masone et al., 2013). But as observed in previous studies, molecular mechanics force fields show poor accuracy in describing hydrogen bond physics (Fabiola et al., 2002; Hu et al., 2003; Lii and Allinger, 1994, 1998; Morozov et al., 2004; Masone et al., 2012).

Molecular dynamics is the tool by excellence to exhaustively explore the protein potential energy landscape while simulating its flexibility. Fully flexible relaxation tends to increase the amount of recovered native contacts among sets of docking poses (Król et al., 2007a, 2007b). However, it is impossible to assure that extensive molecular dynamics simulations will result in good conformers suitable for docking. Moreover, determining the most important motions for binding purposes and then performing docking experiments may result in equivalent solutions as flexible docking (Cozzini et al., 2008). As pointed out by Alonso and collaborators (Alonso et al., 2006) molecular dynamics have shown to accurately, although expensively, refine a few selected candidates from a previous fast docking stage used to sample large configurational spaces. The full atomistic description in long, and though useful, time scales are still beyond classical molecular dynamics simulations for most biological systems due to the small femtosecond time steps needed for energy conservation. Moreover, crystallographic monomers of a protein–protein complex cannot provide enough information on how interface rearrangements will occur. Hence, in these cases classical molecular dynamics can achieve only a limited phase-space exploration (Tiwary and van de Walle, 2013).

Remarkably, collective variable driven molecular dynamics have shown to adequately reproduce complex conformational changes in biomolecules by accelerating rare events (Fiorin et al., 2013). When biasing the system with a previously chosen collective reaction coordinate, molecular dynamics simulations may surpass intrinsic limitations of the physical model and a more efficient statistical sampling can be performed. However, it is usually difficult to select the proper collective variable that adequately describes the macroscopic phenomena (Fiorin et al., 2013; Laio and Parrinello, 2002; Laio and Gervasio, 2008; Kumar et al., 1996).

The purpose of this work is to propose a collective coordinate to optimize hydrogen bond networks in protein–protein systems by driving molecular dynamics simulations. The collective coordinate is then a measure of the hydrogen bond formation in each docking solution and though is used as a scoring function.

Table 1
Selected complexes from Vakser et al. decoy. Complexi: pdb code of co-crystallized structure. Rec.ii: pdb code of unbound receptor structure. RMSDiii: C_alpha rmsd of unbound receptor and co-crystallized structure [Å]. Lig.iv: pdb code of unbound ligand structure. RMSDv: C_alpha rmsd of unbound ligand and co-crystallized structure [Å]. RMSDvi: The ligand RMSD of the best near-native solution [Å]. HITSvii: The number of near-native solution kept in each decoy set. H-bondviii: The ranking position of the first near-native conformation identified.

Complexi	Rec.ii	RMSDiii	Lig.iv	RMSDv	RMSDvi	Hitsvii	H-bondviii
1bvn	1hx0	0.63	1ok0	0.42	2.24	10	1
1tmq	1jae	0.77	1b1u	1.42	2.07	10	19
1ugh	1akz	0.61	1ugi	2.60	2.86	10	9
1xd3	1uch	2.45	1yj1	2.73	3.64	10	5
3sic	1sup	0.34	3ssi	0.78	3.54	10	2

2. Materials and methods

We used the DockGround (Liu et al., 2008) set of protein–protein solutions generated by Vakser and collaborators, freely available on-line (http://dockground.bioinformatics.ku.edu/) that provides 100 non-native and at least one near-native (ligand RMSD < 5 Å) solution generated by GRAMM-X (Tovchigrechko and Vakser, 2006) docking scan per complex for a total of 61 complexes. To select our complexes we chose three (1xd3, 1ugh, 3sic) that were not included in a previous study in hydrogen bond network optimizations (Masone et al., 2012) using the software PELE (Borrelli et al., 2005), but still containing 10 near-native solutions in the DockGround decoy. Other two complexes that did were studied before (1bvn, 1tmq) were also selected for validation purposes (see Table 1). In all of them the condition that the ligand RMSD to the crystal reference is 5 Å or less for at least one of the poses in the decoy was fulfilled. Previous studies in protein–protein (Masone et al., 2012) and protein–ligand interactions (Borrelli et al., 2010) show that refinement techniques can only return near-native top scores if at least one of the poses in the decoy is close enough to the crystallographic reference. Fig. 1 shows the configurational space explored by the docking scan for the 1bvn protein–protein system.

We performed molecular dynamics in GROMACS 4.5.5 (Hess et al., 2008) patched with PLUMED 1.3 (Bonomi et al., 2009) to use a collective reaction coordinate in order to drive the system into the formation of hydrogen bonds. The PLUMED code provides a variety of different collective variables to performed free-energy calculations. The collective variable we used, available in PLUMED 1.3, counts the number of intra-molecular hydrogen bonds between a group donors and acceptors and it is defined as follows:

$$s = \sum_{ij} \frac{1-(d_{ij}/r_0)^n}{1-(d_{ij}/r_0)^m} \tag{1}$$

where i counts over the group of donors and j over the group of acceptors. For d_{ij} distance calculations all donor–acceptor pairs were included and the user defined values where set to $r_0 = 2.5$, $n = 6$ and $m = 12$. As a general rule, the two monomers of a