Computational Biology and Chemistry 61 (2016) 23–38

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Comparison of non-sequential sets of protein residues

Leonardo D. Garma, André H. Juffer*

Biocenter Oulu, and Faculty of Biochemistry and Molecular Medicine, University of Oulu, PO Box 5400, FI-90014 Oulu, Finland

ARTICLE INFO

Article history:
Received 2 June 2015
Received in revised form
16 December 2015
Accepted 16 December 2015
Available online 25 December 2015

Keywords:
Structural alignment
Sequence-independent
Structural similarity
FAD
NAD
TM-SITE

ABSTRACT

A methodology for performing sequence-free comparison of functional sites in protein structures is introduced. The method is based on a new notion of similarity among superimposed groups of amino acid residues that evaluates both geometry and physico-chemical properties. The method is specifically designed to handle disconnected and sparsely distributed sets of residues. A genetic algorithm is employed to find the superimposition of protein segments that maximizes their similarity. The method was evaluated by performing an all-to-all comparison on two separate sets of ligand-binding sites, comprising 47 protein-FAD (Flavin-Adenine Dinucleotide) and 64 protein-NAD (Nicotinamide-Adenine Dinucleotide) complexes, and comparing the results with those of an existing sequence-based structural alignment tool (TM-Align). The quality of the two methodologies is judged by the methods' capacity to, among other, correctly predict the similarities in the protein-ligand contact patterns of each pair of binding sites. The results show that using a sequence-free method significantly improves over the sequence-based one, resulting in 23 significant binding-site homologies being detected by the new method but ignored by the sequence-based one.

1. Introduction

The overall structure of a protein is determined by its full sequence. Various types of interactions between residues on the one hand and between the residues and the solvent on the other hand guide the folding process. Once a protein acquires its native structure, its biologically relevant functions, including binding to other molecules and catalytic activity, are carried out by functional elements. Depending on the degree of which the mechanism of action of a particular function is known, these elements can be narrowed down to a specific chain in a multimer, a particular domain of the protein, or even to groups of residues localized in a particular region of the structure. As a consequence, the functions of a given protein may be identified by the presence of various residue patterns in its sequence (Whisstock and Lesk, 2003).

Ligand-binding sites (LBSs) in monomeric proteins are one of the most typical (and classical) examples of functional elements in proteins. They are located in a specific region of the structure, and are directly linked to the catalytic activity of enzymes. There is already a very large body of structures available for which biologically relevant protein–ligand interactions have been identified. However, there is no consensus on how the LBSs are defined on the basis of crystal or NMR structures. A number of databases that compile ligand–protein complexes exist (Hu et al., 2005; Liu et al., 2007).

Some of the experimentally determined structures include data about the LBSs. In some other cases, only the structure is available, with no immediate evidence available about which residues are involved in ligand binding. For protein–ligand complexes, the LBS residues can be defined using a simple ligand–residue distance criterion. For instance, in the LigASite database (Dessailly et al., 2008), any residue with an atom closer than a given cutoff distance to a ligand is considered to be part of the LBS. The cutoff distance should take into account that there is some uncertainty in the positions of atoms in the structures, while proteins are also inherently flexible. As a consequence, the protein–ligand interactions might not have the exact range as observed in a given structure or are simply a product of the crystallization process. A cutoff that is too short (e.g. based on ideal bond lengths) might leave out some meaningful residues or include irrelevant ones.

Other databases (Binkowski et al., 2003) instead opt for defining a ligand cavity. The definition of a cavity is based on similar principles as in LigASite, but relies on surface properties instead of distances to the ligand and does not assume that all the residues included in the cavity definition will be in direct contact with the ligand.

Proteins with the same function tend to share high sequence and/or overall structural similarity. However, proteins with

* Corresponding author.
E-mail address: andre.juffer@oulu.fi (A.H. Juffer).

http://dx.doi.org/10.1016/j.compbiolchem.2015.12.004
1476-9271/

divergent structures and sequence may still present homologous functional elements and be involved in similar functions (Bruns et al., 1997; Rosen et al., 1998). Generally, comparisons of sequence and structure are required for template-based predictions of protein functions. Many of the structural alignment tools currently available are restricted to three-dimensional structures that are continuous in sequence (Zhang and Skolnick, 2005; Hasegawa and Holm, 2009). Clearly, such comparisons are of little value when dealing with functional elements consisting of disconnected and sparsely distributed residues.

A number of methods have been developed to overcome this limitation by conducting sequence-independent structural alignments. Current methods for comparing two structures can be generally classified according to three different criteria: (i) the way the structures are represented, (ii) how the similarity between the structures is defined, and (iii) which method is used to find the superimposition of structures that maximizes the similarity. A representation of the structures can be obtained by employing all atoms in the structures (Gold and Jackson, 2006; Sehnal et al., 2012), C_α-atoms only (Nussinov and Wolfson, 1991; Wolfson and Rigoutsos, 1997), a combination of C_α-atoms and side chain averages (Ausiello et al., 2005), molecular surfaces (Coleman and Sharp, 2010) or functional groups (Konc and Janežič, 2010). Similarity between two structures can be established by employing different combinations of a number of factors, including geometry (Shatsky et al., 2002; Chang et al., 2004), surface shape (Chang et al., 2004), physico-chemical properties (Shulman-Peleg et al., 2005) or matching atom element types (Gold and Jackson, 2006). Finally, there are several methods for finding the superimposition that maximizes the similarity, the most common ones being geometric hashing (Nussinov and Wolfson, 1991; Wolfson and Rigoutsos, 1997; Chang et al., 2004; Gold and Jackson, 2006), maximum clique techniques (Konc and Janežič, 2010; Nguyen et al., 2011) and exhaustive searches (Ausiello et al., 2005; Feldman and Labute, 2010; Sehnal et al., 2012).

The method of the present work makes use of a C_α-atom representation of the structures. Feldman and Labute (2010) demonstrated that this is enough to develop a reliable method for comparing binding pockets. Moreover, in the vast majority of cases the C_α-atoms undergo very limited displacements (root mean square deviation ≤ 2 Å) even in the event of ligand binding (Brylinski and Skolnick, 2008), whereas for the side-chain atoms significant changes are observed (Najmanovich et al., 2000; Gaudreault et al., 2012). The similarity function of the present work is based on a combination of geometrical and physical–chemical similarities between structures, following an idea somewhat similar to the function used by CPASS (Powers et al., 2011), but without any notion of distances to ligand (therefore extending its use to any pair of structures, not only ligand-bound ones) and a rather different treatment of geometrical similarities. The similarity between pairs of structures is maximized by means of a genetic algorithm (Davis, 1991), which does not require the establishment of a pairing of residues or atoms prior to the structural comparison. In addition, it does not impose the optimal overlap of any pair of residues.

The BioLIP database (Yang et al., 2013) contains a non-redundant collection of all the structures of protein–ligand complexes available from the Protein Data Bank (Berman et al., 2002) (PDB). For each binding site in each complex, the database holds a list of residues that form the LBS and its associated so-called Shortest Sequence from First to Last binding residue (SSFL). The continuous substructure corresponding to the SSFL can be used instead of the full protein to perform a local structural alignment. This causes the alignment to be driven by the region surrounding the LBS, removing the influence of other regions of the protein far away from it. Such local alignments have already proven to be a much better approach for comparing LBSs (Yang et al., 2013).

TM-SITE (Yang et al., 2013) is a recently developed method for the prediction of LBSs. It makes use of ConCavity (Capra et al., 2009) to analyze the surface of a query protein and predict possible ligand cavities. The SSFLs of these predicted cavities are subsequently employed to perform structural alignments to the SSFLs of known LBSs. The outcome of the alignments is then evaluated according to a composite function that takes into account local and global similarities. The templates that produce high scores on the evaluation function are taken as putative templates for the prediction of the LBS residues.

A possible drawback of SSFLs is that they may contain residues other than the ones in the LBS, which in some cases may mean that the residues relevant to the comparison represent only a small fraction of the substructure used in the alignment. Thus, just like the alignment of SSFLs for comparing LBSs is an improvement over the global alignment by not taking into account regions distant from the LBS, the use of only those residues actually involved in ligand binding represents an even further increase of the sensitivity of such comparisons. This was already pointed out by Zhang et al. in Yang et al. (2013), but the same study claimed that such a method would result in a high false positive rate as a consequence of a too small number of residues involved in the comparisons, in reference to a previous work (Roy et al., 2012).

Any comparison of two structures or sets of possibly disconnected residues requires a definition of similarity. Clearly, the similarity can be measured using a virtually infinite number of different criteria. The simplest approach is to rely solely on geometry: the more residues of both structures can overlap, the more similar the structures are supposed. Due to the small size of protein functional sites, such a measure would provide limited information and only a few reliable assumptions about the shared properties of the two structures can be made. Alternatively, defining the similarity on the basis of the physico-chemical properties of those residues that can be superimposed, would make it a more strict criterion as it compares the local chemical environments of the structures. A main hypothesis of the present work is that such a definition of similarity allows for making more reliable assumptions as to whether or not two sets of non-sequential residues share common properties.

The purpose of the present work then is to introduce an alternative procedure for measuring the similarity between any two given sets of residues. The method introduced in this study is capable of handling interfaces or functional elements that are composed of disconnected and sparsely distributed residues. The similarity is defined in terms of specific common properties of the interfaces, as defined by chemical distance matrices. These matrices are built by establishing a distance between each pair of amino acid types, which reflects their differences on a number of given physico-chemical features (Grantham, 1974).

The method relies a non-deterministic search heuristic by means of a genetic algorithm (GA) (Davis, 1991) to generate superimpositions of one template structure over a query and finds the one that maximizes their similarity. The quality of the new methodology is demonstrated by comparing its results with those of TM-SITE (Yang et al., 2013). The cross-comparison of the two methods was performed by measuring the contact conservation rate (CCR), the ligand displacement (LD), and the value of a modified version of the TM-SITE scoring function (q^R_{str}) for each alignment. Various chemical distance matrices were employed.

2. Methods

The quality of the procedure for measuring the similarity between any two given sets of residues (forming substructures, interfaces, cavities, and so forth) was tested by employing the new method for an all-to-all comparison of two separate sets of ligand–protein complexes. Below we first present a very short