Computational Biology and Chemistry 68 (2017) 6–11

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

A new search subspace to compensate failure of cavity-based localization of ligand-binding sites

Kalpana Singh, Tapobrata Lahiri*

Department of Applied Science, Indian Institute of Information Technology, Indian Institute of Information Technology, Deoghat, Jhalwa, 211012, Allahabad, India

ARTICLE INFO

Article history:
Received 3 December 2015
Received in revised form 27 April 2016
Accepted 30 January 2017
Available online 31 January 2017

Keywords:
Ligand-binding site localization
Alternative search subspace
Protein atomic cluster roughness

ABSTRACT

The common exercise adopted in almost all the ligand-binding sites (LBS) predictive methods is to considerably reduce the search space up to a meager fraction of the whole protein. In this exercise it is assumed that the LBS are mostly localized within a search subspace, cavities, which topologically appear to be valleys within a protein surface. Therefore, extraction of cavities is considered as a most important preprocessing step for finally predicting LBS. However, prediction of LBS based on cavity search subspace is found to fail for some proteins. To solve this problem a new search subspace was introduced which was found successful to localize LBS in most of the proteins used in this work for which cavity-based method MetaPocket 2.0 failed. Therefore this work appeared to augment well the existing binding site predictive methods through its applicability for complementary set of proteins for which cavity-based methods might fail. Also, to decide on the proteins for which instead of cavity-subspace the new subspace should be explored, a decision framework based on simple heuristic is made which uses geometric parameters of cavities extracted through MetaPocket 2.0. It is found that option for selecting the new or cavity-search subspace can be predicted correctly for nearly 87.5% of test proteins.

1. Introduction

Localization of ligand-binding site (LBS) of a protein is a challenging task till today. LBS are important for many reasons, e.g., structure-based drug design (Sotriffer and Klebe, 2002), explanation of cause of diseases and protein function (Konc and Janežič, 2014), and for study of side effects of drugs (Xie et al., 2016). LBS are usually (Laskowski et al., 1996) found on the surface of proteins in clefts, pockets and cavities of largest size and depth (Lewis, 1991). In this perspective, many LBS prediction approaches are centered on search of pockets and cavities (Yu et al., 2010; Weisel et al., 2007; Brady and Stouten, 2000) notwithstanding of the fact that sometimes LBS are also found outside deep cavities (Nisius et al., 2012). These methods mainly relies on extraction of a search subspace that includes deep cavities which was reported to have success rate in around 83% cases (Laskowski et al., 1996), however over the time, cavity finding approaches became more robust to include LBS with more success rate on studied data sets. There are various cavity-finding methods based on the strategies which are purely geometrical or geometrical with added physicochemical properties or solely energy based or evolutionary and threading based or consensus methods like MetaPocket 2.0 (based on meta approaches adding results of other methods) (Krivák and Hoksza, 2015).

Continuing with this trend, inthis study, 198 drug-target complexes mentioned in work of Zhang et al. (2011) were utilized along with other datasets. The success rate of finding top three cavities as LBS was reported as 74% in this drug-target dataset (Zhang et al., 2011). However considering all cavities as a search subspace for LBS, it was found that overall success rate was 88% for the same dataset. Therefore, it appeared that there still remained scope for development of a method which could include proteins for which LBS was found outside the deep cavities.

Furthermore, the reason of occasional failure of resultant cavities to contain LBS within them may be explained on the basis of the surface-depth criteria following which they are obtained. Considering surface-depth as main criteria for selecting a search subspace, cavity, one might yield deep yet smooth valleys within the protein surface. However, the work of Pettit and Bowie (1999) showed that the roughest patch of the protein surface is the most probable zone to contain LBS. In this context, it remains a matter of interest to study the application of the roughness as a criterion at the initial phase of extraction of a search subspace. It appears to be further necessary since mechanistic formalism also favors

* Corresponding author.
E-mail addresses: rs108@iiita.ac.in (K. Singh), tlahiri@iiita.ac.in (T. Lahiri).

http://dx.doi.org/10.1016/j.compbiolchem.2017.01.013
1476-9271/

gripping, attaching or adhering of an object to a rough surface (i.e., increased amount of surface) against a comparatively smooth surface. Taking cue from this fact, this paper primarily targeted to utilize the potential of roughness of spatial distribution of protein atoms as main signature to extract search subspace for localization of LBS.

In this direction, a new method of obtaining search subspace was investigated which was primarily cavity-independent where geometric roughness in the atomic spatial distribution within a protein was utilized. It was followed by introduction of a new parameter, Protein Atomic Cluster Roughness (PACR) in the next step. Further, PACR was utilized to get some specific atomic clusters that were merged to obtain a search subspace referred as Grand Merge Cluster (GMC) construct. This subspace was further refined through taking its intersection with the protein surface to obtain the final search subspace which was further studied for its effectiveness to address the issue of improving localization of LBS. The newly derived subspace was found to offer a comparable and somewhat complementary partition for better localization of LBS together with subspaces extracted by MetaPocket 2.0 (Zhang et al., 2011) in comparison to that solely obtained through MetaPocket 2.0.

The complementary nature of effectiveness of the new subspace vis-à-vis the cavity based method adopted in MetaPocket 2.0 also brought in challenge to identify the cases which were going to be complementary in nature. It boiled down to the problem of screening of the proteins for which cavity subspace might fail to include LBS. In this regard, machine learning as well as simple heuristics based methods were applied taking input as various geometric properties of the cavities obtained from MetaPocket 2.0 to decide whether the result was incorrect and hence the new subspace should be pressed in.

2. Material and methods

2.1. Collection of datasets

210 protein-ligand complexes were taken from the PLD database along with datasets of 48 bound (Holo) and unbound (Apo) structures mentioned in work of Huang and Schroeder (2006). Moreover, 198 drug-target complexes mentioned in work of Zhang et al. (2011) were also used in this study.

2.2. Derivation of protein atomic clusters

For this, first, all protein atoms were clustered to obtain Protein Atomic Clusters utilizing *K*-means clustering following the algorithm described by Forgy (1965). In this method the atomic positions described by X, Y, and Z-Cartesian coordinates of the protein atoms taken from pdb file were clustered. As described in the later Section (2.2.2) *K* was heuristically chosen as 19.

2.2.1. Determination of invariant initial cluster centers

Atomic depths (i.e., distances) of all atoms to CG (Center of gravity)of the protein were measured following the concept of residue depths (Tan et al., 2013) and were sorted in ascending order to pick *K* number of initial cluster centers (i.e., atoms) spread in such a manner so that index-distance between two successive atoms would be almost constant.

2.2.2. Determination of K

As shown in Fig. 1, *K* was chosen from the maximum value of Fractional Binding Site Residues, *F* from *K* versus *F* plot where, $F = (B/T) \times B$, *B* and *T* were % of LBS residues and% of total atoms in GMC averaging out from 26 bound proteins (mentioned in work of Pettit and Bowie (1999)) respectively. *F* fulfilled the desired objective of maximization of *B/T* indicating maximum inclusion of LBS in minimum new search subspace weighted by the factor *B*.

2.3. Derivation of protein atomic cluster roughness (PACR)

PACR of *i-th* cluster, $PACR_i$ was calculated following the method of Singha et al. (2006) as shown below:

$$PACR_i = \sqrt{\frac{1}{N_i}\sum_{j}^{N_i}(D_j^i - \overline{D_j^i})^2}$$

Where, D_j^i and $\overline{D_j^i}$ were the distance of j-th atom of i-th cluster from the CG of the protein and mean of all such distances measured over i-th cluster respectively, and, N_i is number of atoms within this cluster.

2.4. Derivation of grand merged cluster (GMC) based on PACR

The task of constructing GMC was adapted from the work of Pettit and Bowie (1999) where two protein atomic clusters with maximum PACRs were first merged into one cluster, termed as merged cluster (MC). Distance of the maximally distant atom from the center of MC was calculated and referred as D. Then the space formed by atoms within the distance 2/3rd of D (chosen heuristically after many trials) from the CG of the protein (referred as Core Space) was merged with MC to form GMC.

2.5. Recruitment of GMC to derive new R-subspace for localization ofLBS

Surface part of GMC was considered as final new R-subspace referred as Rough Subspace (R-subspace) to localize LBS whereas,

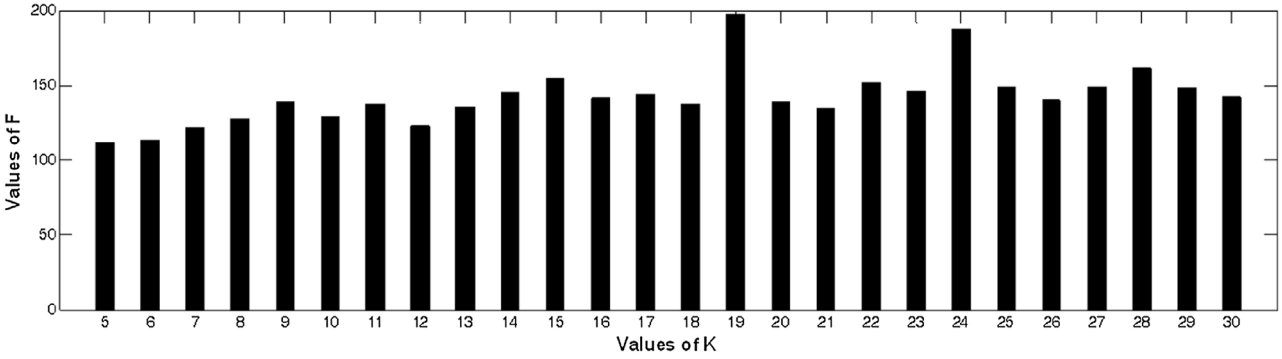

Fig. 1. Bar diagram of values of *F* (Fractional Binding Site Residue within GMC) for each value of *K* (number of clusters).