Journal of Biotechnology 159 (2012) 123–134

Contents lists available at SciVerse ScienceDirect

Journal of Biotechnology

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

Structure-based computational analysis of protein binding sites for function and druggability prediction

Britta Nisius, Fan Sha, Holger Gohlke*

Department of Mathematics and Natural Sciences, Institute of Pharmaceutical and Medicinal Chemistry, Heinrich-Heine University Düsseldorf, Germany

ARTICLE INFO

Article history:
Received 2 October 2011
Received in revised form 2 December 2011
Accepted 6 December 2011
Available online 14 December 2011

Keywords:
Computational analysis
Druggability
Protein binding sites
Protein function prediction
Protein structure

ABSTRACT

Protein binding sites are the places where molecular interactions occur. Thus, the analysis of protein binding sites is of crucial importance to understand the biological processes proteins are involved in. Herein, we focus on the computational analysis of protein binding sites and present structure-based methods that enable function prediction for orphan proteins and prediction of target druggability. We present the general ideas behind these methods, with a special emphasis on the scopes and limitations of these methods and their validation. Additionally, we present some successful applications of computational binding site analysis to emphasize the practical importance of these methods for biotechnology/bioeconomy and drug discovery.

1. Introduction

The interactions of a protein with other molecules, *e.g.*, ligands, nucleic acids, or other proteins, are critical to its biochemical function. Usually, not all residues on a protein's surface participate in these interactions; rather the interactions occur at defined locations, the protein binding sites. Thus, the identification and characterization of these protein binding sites is crucial to understand molecular interactions and recognition. The binding of a molecule to a protein's binding site depends on their physico-chemical and shape complementartiy. Size, buriedness, and flexibility of the binding site are additional key factors for molecular recognition. The role of these key factors will be described in more detail below in this section.

Due to the importance of protein binding sites in molecular recognition and interactions, various approaches aiming at the structure-based computational *binding site analysis* (BSA) have been developed in recent years. In this review, we initially focus on BSA methods to perform *binding site comparison* (BSC). By applying such methods, one can detect binding sites in a set of protein structures that are similar to a given binding site. From a biotechnological/bioeconomical point of view, this characterization of binding pockets allows de-orphanization of (biochemical) protein function by comparing binding pockets of multiple proteins and inferring the function of the orphan protein from the most similar protein(s). This is valuable for the identification of novel enzymes, which can subsequently be used for biocatalytic compound transformation, aiming at more sustainable production pathways (see Section 2). From a pharmaceutical–medicinal chemistry point of view, these methods also help to rationalize and predict cross-target drug interactions and toxicity (see Section 2). Another important question in drug discovery relates to whether a protein binding site is amenable to binding drug-like molecules, *i.e.*, "*druggability*" *prediction* (DP). BSA methods that aim at answering this question are reviewed in Section 3. Rather than presenting algorithmic details of all available methods, we aim at presenting the general ideas of selected methods together with practical applications. Finally, we discuss the scopes and limitations of the presented methods in Section 4.

1.1. Diversity of protein binding sites

Since proteins are able to interact with a wide range of molecules, the binding sites involved in these interactions are diverse: The active site of an enzyme is often characterized by a particularly large and deep cleft, whereas protein–protein interfaces are usually flat and unstructured (Laskowski et al., 1996). Furthermore, even binding sites of enzymes can vary significantly. For instance, the binding site of endonuclease is a spherical cavity containing a deeply buried ligand, whereas the binding site of ribonuclease is an elongated groove containing a rather exposed ligand (Fig. 1).

Abbreviations: BSA, binding site analysis; BSC, binding site comparison; DP, druggability prediction; PPI, protein–protein interaction.

* Corresponding author at: Universitätstraße 1, 40225 Düsseldorf, Germany. Tel.: +49 211 81 13662; fax: +49 211 81 13847.
E-mail address: gohlke@uni-duesseldorf.de (H. Gohlke).

0168-1656/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2011.12.005

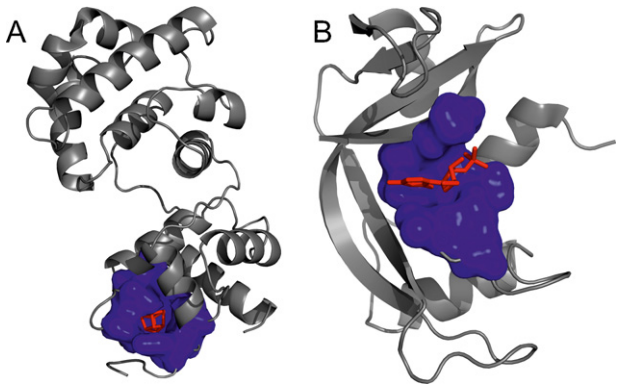

Fig. 1. Diverse enzyme binding sites. The 3D structures of endonuclease (shown in A, PDB code 2ABK) and ribonuclease (shown in B, PDB 1ROB) and their ligand binding sites (highlighted in blue) are shown. Endonuclease has a spherical cavity with a deeply buried ligand, whereas ribonuclease has an elongated binding site and a rather exposed ligand.

For enzyme binding sites, the largest cavity is most often the active site (Laskowski et al., 1996; Liang et al., 1998). Furthermore, while larger proteins tend to have *more* binding sites, they do not necessarily have *larger* binding sites (Liang et al., 1998). Binding sites involved in protein–ligand interactions are often characterized by the presence of regions with very low and other regions with very high structural stability. Thus, these binding sites exhibit a dual character (Luque and Freire, 2000). The stable part of the binding site usually contains residues involved in interactions requiring a well-defined stereochemical arrangement, *e.g.*, hydrogen bonds. By contrast, the flexible part of the binding site enables an adjustment of the binding pocket's shape to ensure an optimal fit and buriedness of the ligand and/or the accommodation of structurally different ligands. Furthermore, the low stability regions are proposed to play a crucial role in the transmission of information from an allosteric binding site to a catalytic binding site. Thus, shape and size of a ligand binding site are crucial determinants of its recognition power. However, geometrical complementarity alone is not sufficient to fully account for molecular recognition (Kahraman et al., 2007). Additionally, a physico-chemical complementarity is important. For instance, it was shown that some specific amino acids (Arg, His, Trp, and Tyr) occur substantially more frequently in protein binding sites than in the entire protein (Villar and Kauvar, 1994). Furthermore, the amino acid composition among protein binding sites can vary significantly, *e.g.*, neuraminidase has a highly charged binding site whereas the binding site of avidin contains no charged residue (Hou et al., 2011).

Interfaces involved in protein–protein interactions (PPIs) are typically flat and significantly larger than protein–ligand binding sites. Additionally, these larger protein–protein interfaces are most often composed of multiple epitopes that are not sequentially connected. The epitopes can be divided into "functional epitopes", which actually contribute to binding, and additional "structural epitopes" (Grimme et al., in press). In fact, mutagenesis studies revealed that only a small subset of all amino acids flanking the protein–protein interface significantly contribute to binding affinity. These residues are called "hotspots" (Bogan and Thorn, 1998). Furthermore, protein–protein interfaces show a significantly higher degree of inherent flexibility and plasticity than protein–ligand binding sites (Grimme et al., in press).

Thus, since proteins are involved in complex and diverse molecular interactions, a full characterization of protein binding sites requires a detailed analysis of the various factors contributing to molecular recognition. Up to now there is no standard definition of what constitutes a binding site, which represents a major complication in BSA (Perot et al., 2010). However, since the relative importance of the key factors significantly varies for different binding sites, it is indeed very difficult to develop procedures that are generally applicable across diverse sites (Henrich et al., 2009).

1.2. Protein flexibility

Another complication in BSA arises from the flexibility of proteins, which enables a range of possible movements, from single side-chain rotations to drastic structural rearrangements (Ahmed et al., 2007; Cozzini et al., 2008). Thus, it is not always sufficient to use just one static structure for BSA: First, protein flexibility and plasticity can allow for the opening of novel binding sites (so-called "transient" or "cryptic" pockets (Eyrisch and Helms, 2007; Metz et al., in press)) that may not be detectable in the one single structure selected for analysis. Second, binding sites can also change their sizes and shapes upon binding. This is in line with the "conformational selection model" (Tsai et al., 1999), which proposes that, from various rapidly interconverting conformations of the unbound protein, that conformation is picked by a binding partner that has a binding site most complementary with the characteristics of the partner. Thus, based on this model, it has to be assumed that a binding site's shape and size strongly depends on the interacting ligand and, therefore, cannot be analyzed independently of the ligand (Ma et al., 2002). Hence, whenever knowledge about moving protein parts is available, it should be included in the analysis of protein binding sites. This knowledge can be gained from experimental information, *e.g.*, multiple structures solved by crystallography or an ensemble of structures determined by NMR, as well as from computational approaches such as molecular dynamics simulations, graph-theoretical approaches, or normal mode analysis (Ahmed et al., 2007; Cozzini et al., 2008).

1.3. Computational approaches for binding pocket identification

Before protein binding sites can be analyzed by computational means, their location on the surface of a protein has to be identified. Unless a co-crystallized ligand readily provides this information, the detection of potential binding pockets is the first step in computational BSA. Many computational methods have been developed with that aim, which, given a 3D structure of a protein, scan the surface for cavities or pockets that most likely represent binding sites. Since several recent reviews provide detailed insights into these methods (Henrich et al., 2009; Laurie and Jackson, 2006; Perot et al., 2010), we only summarize the main findings as to their detection performance here.

In general, pocket prediction methods can be divided into two categories: energy- and geometry-based algorithms. Energy-based methods aim at finding pockets by computing the interaction energy between protein atoms and a small-molecule probe. By contrast, geometry-based methods try to detect solvent accessible regions that are embedded in the protein surface solely using geometric criteria. A recent comparison of energy- and geometry-based algorithms revealed that, in general, both types of binding pocket detection algorithms exhibit a very good performance (Schmidtke et al., 2010). Especially for *holo* structures, the performance of the compared methods is very similar because all methods correctly predict around 95% of the known binding sites, even though the underlying methods are rather diverse. Still, in a large-scale prediction of potential binding pockets, geometry-based algorithms were found to have some inherent advantages over energy-based algorithms because the former are faster and more robust against structural variations or missing atoms/residues in the input structures (Schmidtke et al., 2010). Notably, many pocket detection algorithms are freely available via web-servers, or are accessible via commercial software packages (see Table 1 in Perot et al., 2010 for a detailed overview).