Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

Cryptic binding sites on proteins: definition, detection, and druggability

Sandor Vajda[1,2], Dmitri Beglov[1], Amanda E Wakefield[1,2], Megan Egbert[1] and Adrian Whitty[2]

Many proteins in their unbound structures lack surface pockets appropriately sized for drug binding. Hence, a variety of experimental and computational tools have been developed for the identification of cryptic sites that are not evident in the unbound protein but form upon ligand binding, and can provide tractable drug target sites. The goal of this review is to discuss the definition, detection, and druggability of such sites, and their potential value for drug discovery. Novel methods based on molecular dynamics simulations are particularly promising and yield a large number of transient pockets, but it has been shown that only a minority of such sites are generally capable of binding ligands with substantial affinity. Based on recent studies, current methodology can be improved by combining molecular dynamics with fragment docking and machine learning approaches.

Addresses
[1] Department of Biomedical Engineering, Boston University, Boston, MA 02215, United States
[2] Department of Chemistry, Boston University, Boston, MA 02215, United States

Corresponding authors: Vajda, Sandor (vajda@bu.edu), Whitty, Adrian (awhitty@bu.edu)

Current Opinion in Chemical Biology 2018, 44:xx–yy

This review comes from a themed issue on **Next generation therapeutics**

Edited by **Adrian Whitty** and **Peter J Tonge**

https://doi.org/10.1016/j.cbpa.2018.05.003

Introduction

Many proteins have small-molecule binding pockets that are not easily detectable in the ligand-free structures. These cryptic sites require a conformational change to become apparent. A cryptic site can therefore be defined as a site that forms a pocket in a ligand-bound structure, but not in the unbound protein structure [1••]. It has long been well-known that proteins are dynamic objects, and that their binding sites may change conformation upon ligand binding [2]. However, finding and utilizing cryptic or hidden binding sites has received growing attention during the last few years [1••,3•,4,5••,6,7,8,9,10], seemingly motivated by two factors. First, many biologically relevant drug targets lack appropriately sized pockets in their unbound structures to support the strong binding of drug-sized ligands [7,11]. It has been suggested that cryptic sites can provide previously undescribed pockets, potentially enabling targeting of proteins that would otherwise be considered undruggable [7], and thereby expanding the 'druggable genome' [11]. In keeping with this idea, some of the pockets that bind small molecule inhibitors of protein–proteins interactions, a class that include many such challenging targets, are cryptic [12]. Cryptic sites located away from the main functional site of a protein, but which can modulate the activity of the protein allosterically, are also potentially useful [13•], particularly if the main functional site cannot be targeted with sufficient specificity [14]. Targeting a distal site also has the potential to give a different pharmacological profile [15]. The second factor contributing to the increased interest in cryptic sites is the availability of improved methodology for identifying such sites, particularly molecular dynamics and Markov state simulation methods that are now computationally feasible [3•,8,9,16,17,18,19]. More specialized computational tools have also been developed that integrate molecular dynamics simulations [7,9] with fragment docking [5••] and machine learning approaches [1••]. Despite this high level of recent interest in cryptic sites, a review of recent publications suggests that a number of questions concerning the definition, identification, and druggability of cryptic sites, and their potential value for drug discovery, are either not fully answered, or have been answered in conflicting ways. Here we identify and discuss some of these questions, with emphasis on problem areas that need further work.

When is a binding site cryptic?

Intuitively, a binding site is cryptic if it can be identified in the ligand-bound but not in the unbound structure of a protein. This definition is far from rigorous, however, since it depends both on the method of searching for the sites, and on the particular unbound structure(s) considered. To develop a benchmark set of proteins with cryptic sites, Cimermancic *et al.* [1••] screened over 20,000 unbound–bound protein pairs from the Protein Data Bank using two pocket detection algorithms, Fpocket [20] and ConCavity [21]. Both algorithms derive scores that reflect the putative capacity of pockets to bind small molecules. They averaged Fpocket's residue

druggability scores and ConCavity's residue scores over residues within 5 Å of the ligand, to form a pocket score with values between zero (no pocket) and 1 (well-formed binding site). This composite measure primarily depends on the volume of the pocket, but also includes other factors such as residue polarity and evolutionary conservation. Cryptic sites were defined as sites with an average pocket score of less than 0.1 in the unbound form of the protein and greater than 0.4 in the bound form. Using these criteria, together with manual inspection, Cimermancic *et al.* [1••] selected 93 pairs in which each unbound structure had a site considered cryptic due to its low pocket score, and each bound structure had a functionally relevant ligand bound at the site. The resulting CryptoSite set is very useful for testing cryptic site prediction algorithms, and the definition of a cryptic site that these authors developed could provide the basis for a community standard as other groups test newly developed methods.

A potential issue with the approach of Cimermancic *et al.* [1••] is that, to determine whether a site can correctly be considered cryptic, it is important to consider the full range of conformations available to the protein in the absence of ligand. Basing the structural comparison on a single unbound structure ignores the ensemble of conformations available to the unbound protein, especially important at mobile regions such as potential cryptic sites [16]. This consideration raises the following question: Can a site be properly considered as cryptic if it is absent in just one or a very few unbound structures, even if it is fully formed in other unbound structures? Such behavior suggests that, although the protein can adopt conformations in which the pocket is absent, it also has accessible conformations in which the pocket is present. In some cases these bound-like conformations may even represent the most abundant state of the protein in the absence of ligand. An alternative, more stringent definition is for a pocket to be considered cryptic only if it is absent in all, or nearly all, unbound structures of the protein, such that it cannot be reliably identified in the absence of a bound ligand, and likely does not exist in any large fraction of the conformational states available to the unbound protein.

Beglov *et al.* [22••] have investigated how broader consideration of the conformations available to the unbound protein would affect the set of cryptic sites identified by Cimermancic *et al.* [1••]. To each protein pair in the CryptoSite set, they added all unbound structures in the Protein Data Bank having at least 95% sequence identity. The number of such additional unbound structures varied from zero to 498 per protein, resulting in an extended CryptoSite dataset that included 4950 structures rather than the original 186. Inclusion of these additional unbound structures revealed that bound-like pockets are at least partially formed in some unbound structures for close to 50% of the 93 proteins in the CryptoSite set [22••]. For example, in the original CryptoSite set the protease beta-secretase 1 (BACE-1) is represented by unbound and bound structures 1W50 and 3IXJ. The unbound structure 1W50 has a low Fpocket druggability score because the loop comprising residues 71–74 is far from the active site, making the pocket too open to score as druggable (Figure 1a). The loop is closed down on the inhibitor in the bound structure 3IXJ, resulting in a well-formed pocket that binds the isophthalamide ligand with high affinity. The analysis of 52 structures of unbound BACE-1 in the extended CryptoSite set reveals that, in these structures, the pockets in question are almost evenly distributed between conformations resembling the unbound and bound forms, with druggability scores varying between 0.2 and 0.6 (Figure 1c). Thus, it is arguable whether this site should be considered as cryptic.

In many of the 93 proteins the analysis of the structures in the extended CryptoSite set revealed some degree of spontaneous shift toward the ligand-bound conformations at the binding site, but with the distribution of observed conformations heavily weighted toward the unbound state [22••]. For example, the original CryptoSite set includes 2GFC and 2JDS as an unbound–bound pair of structures for the cAMP-dependent protein kinase known as Protein Kinase A (PKA). In the unbound structure, 2GFC, the activation loop (which has the sequence SFG rather than the DFG segment seen in many kinases) protrudes into the active site, closing the pocket (Figure 1b). The site opens when an inhibitor binds (PDB ID 2JDS, see Figure 1b). With few exceptions, in the unbound structures the SFG loop resides in the partially hydrophobic outer region of the kinase active site (Figure 1c). Although the pocket is fully formed in a few structures, this is because these structures also contain bound allosteric modulators far from the active site, for example, myristolic acid in the structure 4DFZ. Thus, it is probably reasonable to consider this PKA pocket cryptic.

Proteins in which the cryptic site is completely missing in all unbound structures seem to be rare. A classic example of this type is TEM β-lactamase, in which an elongated cryptic site was discovered serendipitously when crystals revealed two small molecules from the crystallization buffer bound between helices 11 and 12 [23]. In the bound structure (PDB code 1PZO) the position of the shorter helix shifts, opening a substantial crevice (Figure 1d). However, no opening of the site is seen in any unbound structure. Engineered variants of the protein exist in which helix 11 is unfolded, resulting in partial pocket opening [24], but in the wild-type enzyme the site is definitely cryptic.

The conclusion as to whether a site is cryptic may also depend on the resolution of the available structures. An example is protein tyrosine phosphatase 1B (PTP1B), which has a cryptic allosteric site close to its C-terminus [25,26]. The unbound structures 2F6V and 1SUG of