ELSEVIER

Recent computational advances in the identification of allosteric sites in proteins

Shaoyong Lu, Wenkang Huang and Jian Zhang

Department of Pathophysiology, Key Laboratory of Cell Differentiation and Apoptosis of Chinese Ministry of Education, Shanghai JiaoTong University, School of Medicine (SJTU-SM), Shanghai 200025, China

Allosteric modulators have the potential to fine-tune protein functional activity. Therefore, the targeting of allosteric sites, as a strategy in drug design, is gaining increasing attention. Currently, it is not trivial to find and characterize new allosteric sites by experimental approaches. Alternatively, computational approaches are useful in helping researchers analyze and select potential allosteric sites for drug discovery. Here, we review state-of-the-art computational approaches directed at predicting putative allosteric sites in proteins, along with examples of successes in identifying allosteric sites utilizing these methods. We also discuss the challenges in developing reliable methods for predicting allosteric sites and tactics to resolve demanding tasks.

Introduction

Allosteric regulation, established almost 50 years ago, is a fundamental process by which distant sites within monomeric proteins or subunits of oligomeric proteins can communicate [1]. Intrinsically, binding of a ligand at an allosteric site topographically distinct from the orthosteric site regulates the functional activity of the protein through alteration of its conformation and/or dynamics. The widespread occurrence of allostery in the cell has motivated the development of two distinct but complementary methods to decipher how allostery works [2]: from a thermodynamic standpoint, allostery works via the population shift of an ensemble of protein conformations from the inactive to the active states, redistributing the conformational states toward the active conformation favored by the ligand [3–5]; from a structural viewpoint, allostery occurs by the propagation of strain energy created at the allosteric site by ligand binding to the functional site, which emphasizes the allosteric coupling (communication) between the allosteric and functional sites. This knowledge of the allosteric mechanism is extremely useful in facilitating allosteric discoveries and applications.

Allosteric modulators do not compete with orthosteric ligands and act primarily by modulating the affinity or efficacy of endogenous ligands [6]. Under certain circumstances, they have the potential to fine-tune protein activity even when the endogenous ligand occupies the orthosteric site on the same target. In contrast to classical approaches that design orthosteric ligands to interact with highly conserved orthosteric sites, targeting allosteric sites can endow allosteric modulators with greater selectivity, better physiochemical properties and fewer side effects [7]. In addition, in the case of class B G-protein-coupled receptors (GPCRs), allosteric modulators can hit their targets, whereas they are intractable to orthosteric manipulation as a result of the natural polypeptide ligands with varying lengths (ranging from 30 to 40 residues) outside classic oral drug-like space [8]. Thus, harnessing the allosteric modulation of protein function is now considered a novel approach in drug discovery [9–17].

Identification of allosteric sites in proteins of pharmaceutical interest is the first step in allosteric drug discovery. Until now, a vast number of allosteric sites deposited into the Allosteric Database (ASD) [18,19] have fortuitously been discovered through biochemical experiments, such as disulfide trapping [20], high-throughput screening [21] and fragment-based screening [22]. However, these experimental approaches face challenges regarding the fast-growing numbers of allosteric drug targets, and the biased chemical libraries that might not detect the potential allosteric sites. Alternatively, *in silico* methods that provide rapid platforms for identifying allosteric sites in proteins have been recognized to be valuable tools. A host of computational

Corresponding author:. Zhang, J. (jian.zhang@sjtu.edu.cn)

1359-6446/06/$ - see front matter http://dx.doi.org/10.1016/j.drudis.2014.07.012

approaches based on sequence, structure and dynamics have been developed for the prediction of allosteric sites [23–30].

In this review, we survey the recent advances in computational approaches developed for the prediction of allosteric sites. First, the state-of-the art computational approaches are briefly introduced, along with examples that utilize these tools to identify new allosteric sites in proteins. Finally, current challenges facing the computational prediction of allosteric sites and future perspectives are discussed.

Sequence-based prediction approaches

Statistical coupling analysis

Statistical coupling analysis (SCA) [31] is a sequence-based technique that uses a multiple sequence alignment (MSA) to identify networks of co-evolving residues in a protein family; such networks, also termed as protein sectors, provide a structural basis for allosteric communication between functional and allosteric sites [32–34]. Thus, allosteric sites can be predicted by identifying surface sites that are in direct contact with protein sectors. This assumption finds support in results from the computational prediction of allosteric communication of thermodynamically linked residues in the *Escherichia coli* dihydrofolate reductase [35] and PDZ domains [36], which are responsible for signal propagation within the protein structure, as confirmed through mutational analysis.

Very recently, Novinec *et al.* [23] used SCA to analyze the family of papain-like cysteine peptidases and constructed an MSA of 1239 catalytic domains from this family. Pairwise correlations were calculated between all pairs of residues in the alignment, and protein sector residues were identified through automated sector identification to uncover groups of co-evolving residues, which show a continuous network that surrounds and stretches throughout the molecule when mapped on the structure of cathepsin K, a member of the human peptidases. Alanine-scanning mutagenesis was performed to reveal that 14 out of 15 single-substitution mutants derived from the list of sector residues were involved in allosteric communication. Subsequently, potential ligand-binding sites on the surface of cathepsin K were predicted using AutoLigand, and eight potential allosteric sites were identified by filtering the predicted binding sites using the direct sector contact criterion. High-throughput docking of compound libraries to the eight potential allosteric sites of cathepsin K led to the identification of NSC13345, which was posited to occupy site 6 among the top 0.5% of solutions from the compound library and site 5 among the top 2%. Further X-ray crystallographic investigation of the cathepsin K–NSC13345 complex (PDB ID: 4LEG) determined that NSC13345 was bound to the flexible N terminus of cathepsin K, corresponding to the predicted site 6 distant from the active site (Fig. 1), which was thereby identified as a novel allosteric site in cathepsin K. Overall, these results were proved feasible by exploiting SCA to predict allosteric sites computationally through the identification of allosteric networks within a protein.

Structure-based prediction approaches

Allosite

We have constructed the largest dataset of known allosteric sites deposited into the ASD v2.0 [19]. The dataset contains 907 allosteric sites, of which 218 are unique allosteric sites occupied by 436 diverse chemical modulators. These unique allosteric sites are primarily distributed across several classes of therapeutic targets, including kinases (19.3%), ion channels (5.0%) and transcription factors (4.1%). The topological structures and physiochemical properties of these discovered allosteric sites are conductive for the development of a classification model to differentiate between allosteric and non-allosteric sites.

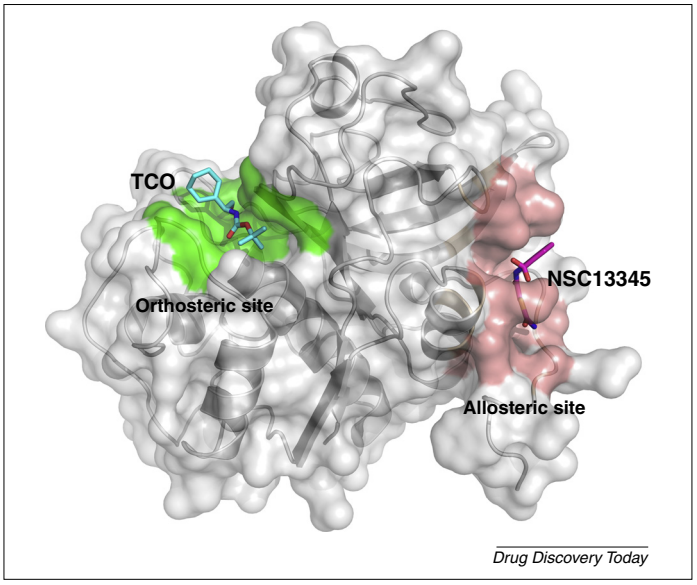

FIGURE 1

View of allosteric and orthosteric sites in cathepsin K (PDB ID: 4LEG). Regions of allosteric and orthosteric sites are highlighted in light pink and green on the surface, respectively, with the allosteric inhibitor NSC13345 (carbon atoms in magenta) and orthosteric inhibitor TCO (carbon atoms in cyan) represented by sticks. TCO is manually docked into the orthosteric site of 4LEG after superimposition with the PDB 1Q6K.

Recently, Huang *et al.* [24] selected 90 nonredundant allosteric sites from the 218 unique allosteric sites to develop a highly efficient, server-based model called Allosite (http://mdl.shsmu.edu.cn/AST) to predict allosteric sites. The training set consisting of 360 binding sites was classified into two groups: 72 allosteric sites and 288 non-allosteric sites, the latter of which were predicted by FPocket. The support vector machine classifier with 21 site descriptors achieved a sensitivity of >83% and a specificity of >96% in the fivefold cross-validation test; additionally, a prediction accuracy of 96.0% was obtained on an external test set composed of 18 allosteric sites and 231 non-allosteric sites. More importantly, based on the set of allosteric proteins where allosteric sites, as yet, undiscovered, Allosite has been proved to be capable of capturing putative allosteric sites in which several mutations adjacent to these sites and that affect the orthosteric functions of the proteins have been determined by biochemical experiments. Therefore, Allosite is a useful starting point for biologists and medicinal chemists in identifying the location of allosteric sites and, ultimately, allosteric drug design.

Normal-mode-analysis-based prediction approaches

Normal mode analysis (NMA) has the ability to provide global modes that bear functional significance. It not only captures most of the functional motions of quaternary structures but also unearths sites that could play a crucial part in mediating or propagating allosteric signals [37]. Hence, NMA has been used