FEBS Letters 588 (2014) 326–333

journal homepage: www.FEBSLetters.org

A computational tool to predict the evolutionarily conserved protein–protein interaction hot-spot residues from the structure of the unbound protein

Neeraj J. Agrawal [a], Bernhard Helk [b], Bernhardt L. Trout [a,*]

[a] Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E19-502b, Cambridge, MA 02139, USA
[b] Novartis Pharma AG, CH-4057 Basel, Switzerland

ARTICLE INFO

Article history:
Received 23 September 2013
Revised 1 November 2013
Accepted 4 November 2013
Available online 12 November 2013

Edited by Wilhelm Just

Keywords:
Molecular simulation
Spatial interaction map
Protein engineering
Protein interaction
Structure-based tool

ABSTRACT

Identifying hot-spot residues – residues that are critical to protein–protein binding – can help to elucidate a protein's function and assist in designing therapeutic molecules to target those residues. We present a novel computational tool, termed spatial-interaction-map (SIM), to predict the hot-spot residues of an evolutionarily conserved protein–protein interaction from the structure of an unbound protein alone. SIM can predict the protein hot-spot residues with an accuracy of 36–57%. Thus, the SIM tool can be used to predict the yet unknown hot-spot residues for many proteins for which the structure of the protein–protein complexes are not available, thereby providing a clue to their functions and an opportunity to design therapeutic molecules to target these proteins.

1. Introduction

It is estimated that a human protein–protein interaction (PPI) interactome is composed of as many as 650000 different PPIs, and understanding these interactions is expected to lead to new therapeutic targets [1]. Proteins are the work-horse of the cellular machinery, and the formation of specific protein complexes led by specific PPIs underpins many cellular processes. Aberrant PPIs, either through the loss of a function or through the formation and/or stabilization of a protein–protein complex at an inappropriate time or location, are implicated in many diseases such as cancer and autoimmune diseases. Elucidating the regions of the protein that drive the PPI helps in understanding the protein function and in designing drugs that target the regions that are involved in the PPI [2,3].

Over the past decade, a large number of protein structures have been solved, and the number of solved structures of protein–protein complexes has been also increasing. These structures of the complexes yield information on the residues that are present in the protein–protein binding regions. These residues constitute the structural epitope of the protein. However, not all of the residues that are present in the binding region contribute equally to the binding energy of the complex. In pioneering work on the binding of human growth hormone (GH) to its receptor, Cunningham et al. identified a region of energetically important residues on the protein surface that were critical to the binding [4]. Following their work and other experiments, it became evident that only a few of the binding-region residues contribute a major component of the binding energy. These residues, which constitute the functional epitope, are termed hot-spot residues. Although a qualitative definition of hot-spot residues is straightforward, consensus on the quantitative definition of hot-spot residues is still lacking. One of the definitions of a hot-spot residue can be construed as the residue that contributes more than a certain threshold (e.g., 2.5 kcal/mol [5]) to the binding energy of the PPI. Because direct experimental measurements of the contributions of individual residues to the protein–protein binding free energy are currently very tedious, an operational definition of a hot-spot residue is often used. Operationally, a hot-spot residue can be defined as a residue that, when mutated to alanine, leads to at least some given increase (e.g., 10-fold) in the protein–protein dissociation constant (K_D).

Experimentally, site-directed mutagenesis has been widely used to analyze how protein–protein interfaces function. In this method, subsets of the protein residues are systematically mutated, mostly one at a time, and the effect of each mutation on the protein–protein binding energy is analyzed. The preferred residue to mutate to is alanine because the alanine amino acid lacks a

* Corresponding author. Fax: +1 617 253 2272.
E-mail address: trout@mit.edu (B.L. Trout).

0014-5793/$36.00
http://dx.doi.org/10.1016/j.febslet.2013.11.004

side chain beyond the β-carbon. Hence, the binding assays performed in conjunction with (alanine) mutagenesis identify hot-spot residues as defined by the operational definition. In these experiments, it is tacitly assumed that the mutation of a residue to alanine does not lead to structural perturbations of the protein. In fact, Rao et al. have aptly demonstrated the limitation of such an assumption [6]. In their experiments, although the mutation F19A led to a significant reduction in the binding strength of human prolactin to its receptor, residue F19 cannot be considered to be a hot-spot residue because the F19A mutation is accompanied with significant structural changes [6]. In experiments in which site-directed mutagenesis is restricted to only surface-exposed residues, as identified from the protein structure, the chances of protein structure perturbation upon mutation greatly diminishes.

On the computational front, a few tools have been developed to identify hot-spot residues. All of these bioinformatics tools, which have been trained over a dataset, can be broadly classified into two categories: tools that are based on the structure of the protein–protein complex and tools that are based on the sequence/structure of the unbound protein. The first category includes energy-based tools [7–11], and machine learning-based tools such as PCRPi [12], KFC [13], MINERVA [14], HotPoint [15], and others [16]. While these tools can identify hot-spot residues with great accuracy, the requirement of the protein–protein complex structure severely limits the application of such tools, and these tools cannot be employed to predict hot-spot residues when the structure of the protein–protein complex is unavailable. The other category of computational tools identifies hot-spot residues by using the sequence or structure of the unbound protein alone. Tool such as ISIS [5,17] is designed to identify protein-protein interaction hot-spot residues using an unbound protein structure and/or sequence. The majority of other sequence-based computational tools, e.g., PredUs [18], meta-PPISP [19] and ConSurf [20], are designed to identify protein-protein binding-region residues. Another tool, called FTMAP [21], has been designed to predict hot-spot residues of small molecule ligand interactions with a protein by using the structure of the protein. Readers are directed to reviews [22,23] from the laboratory of Nussinov on the available computational tools for predicting the binding-region residues. In this article, we present a new method for the prediction of the hot-spot residues from the structure of the unbound protein. We also compare our method to other methods (ISIS [5,17], meta-PPISP [19], PredUs [18,24] and ConSurf [20]), which also use the sequence/structure information of only the unbound protein to predict the hot-spots/binding region residues of the protein.

Recently, our group developed a tool that was called the spatial-aggregation-propensity (SAP) to identify aggregation-prone regions in proteins [25]. SAP is a measure of the dynamic exposure of hydrophobic patches on the protein surface. The SAP tool can also predict, using the unbound protein structure, the binding regions in a protein [26]. Thus, a patch of exposed hydrophobic residues that is indicated by a high SAP value of the region is a good indicator of a protein binding region. Furthermore, previous work on the detection of hydrophobic patches on the surfaces of proteins has also shown the utility of finding hydrophobic patches for identifying protein binding regions [27]. Recently, Kozakov et al. also demonstrated that protein hot-spots are characterized by regions that are patterned with hydrophobic and polar residues [28]. With this background in mind, we developed a computational tool called the spatial-interaction-map (SIM).

We apply the SIM tool to a number of proteins, to predict their hot-spot residues. By design, the SIM tool can be applied to a single (i.e., static) structure of the protein and to multiple structures of the protein. When the SIM tool is applied to a static structure, we refer to it as sSIM; when the SIM tool is applied to multiple structures, we refer to it as dSIM. We compare the SIM-predicted residues with the experimentally known hot-spot residues and the experimentally known binding-region residues; we also compare ISIS, PredUs, meta-PPISP and ConSurf in terms of their ability to predict hot-spot and binding-region residues for these proteins. Because a few previous studies on the characterization of protein–protein interfaces have cast doubt on the utility of hydrophobicity for the prediction of the protein–protein interface [29–31], we also compare our predictions obtained by using SIM against predictions obtained by performing simple hydrophobic analysis. For benchmarking purposes, we also report the results that were obtained when all of the exposed residues were considered to be hot-spot and binding-region residues.

For validation of our computational method, we resort to the experimentally known hot-spot residues and binding-region residues of evolutionarily conserved protein–protein interactions. Publicly available databases such as ASEdb [32] and BID [33] contain a repository of experimentally known hot-spot residues, and the HotSprint [34] database contains a repository of computational hot-spot residues. However, quite a large number of protein–protein interactions contained in ASEdb and BID belong to an antigen-antibody interaction, which is an evolutionarily non-conserved interaction. Furthermore, these databases do not necessarily provide information on the known binding-region residues. For most of the protein–protein interactions, a number of binding-region residues still lack experimental data that can be used for classifying them as hot-spot or non-hot-spot residues. This lack of information can affect the performance of a computational method when the reported method's accuracy is based on the ratio of the correctly predicted hot-spot residues to the total number of predicted residues. To account for this lack of experimental information, we calculate the accuracy and the theoretical maximum accuracy (defined as the accuracy when all of the binding region residues that have not been experimentally tested are assumed to be hot-spot residues) for each method. Hence, in this work, we study only those proteins for which the protein–protein interaction is evolutionarily conserved and for which both the hot-spot residues and the binding-region residues are experimentally known.

In this report, we show the results for IL-13 protein. The results for other proteins, specifically IL-2, growth hormone receptor, IL-15, growth hormone, Fc-domain of an IgG1, erythropoietin, IL-13Rα1 and EGFR, are given in the Supporting information, Section S4. For all of these proteins, we report any concerns on the quality of the experimental data in the Supporting information tables.

2. Methods

2.1. Spatial-interaction-map (SIM)

The input to the spatial-interaction-map (SIM) tool is a fully atomistic three-dimensional structure of the protein (see Supporting information, Sections S1.1 for the details on the methods used to obtain protein structure and perform molecular simulations). sSIM indicates SIM computed on a single protein structure, and dSIM indicates SIM computed over multiple structures of the protein. These multiple structures of the protein are generated using molecular-dynamics simulations. Calculations to perform SIM analysis can be divided into four steps. Step I: using the structure of the protein, we assign an effective-hydrophobicity value to each of the residues of the protein. The effective hydrophobicity Φ_{eff} of the i^{th} residue is defined as [25]: