A compound-based computational approach for the accurate determination of hot spots

Lincong Wang,^{1*} Yaqin Hou,² Haihua Quan,² Weiwei Xu,² Yongli Bao,^{2*} Yuxin Li,² Yuan Fu,¹ and Shuxue Zou¹

¹The College of Computer Science and Technology, Jilin University, Changchun, Jilin, China

²National Engineering Laboratory for Druggable Gene and Protein Screening, Northeast Normal University, Changchun, Jilin, China

Received 6 February 2013; Revised 27 May 2013; Accepted 1 June 2013

DOI: 10.1002/pro.2296

Published online 14 June 2013 proteinscience.org

Abstract: A plethora of both experimental and computational methods have been proposed in the past 20 years for the identification of hot spots at a protein–protein interface. The experimental determination of a protein–protein complex followed by alanine scanning mutagenesis, though able to determine hot spots with much precision, is expensive and has no guarantee of success while the accuracy of the current computational methods for hot-spot identification remains low. Here, we present a novel structure-based computational approach that accurately determines hot spots through docking into a set of proteins homologous to only one of the two interacting partners of a compound capable of disrupting the protein–protein interaction (PPI). This approach has been applied to identify the hot spots of human activin receptor type II (ActRII) critical for its binding toward Cripto-I. The subsequent experimental confirmation of the computationally identified hot spots portends a potentially accurate method for hot-spot determination *in silico* given a compound capable of disrupting the PPI in question. The hot spots of human ActRII first reported here may well become the focal points for the design of small molecule drugs that target the PPI. The determination of their interface may have significant biological implications in that it suggests that Cripto-I plays an important role in both activin and nodal signal pathways.

Keywords: protein–protein interaction; hot spot; hot-spot prediction; protein–ligand docking; protein–protein docking; scoring function; activin receptor; Cripto-I; activin signal pathway and nodal signal pathway

Abbreviations: ActRIIA, activin receptor type IIA; ActRIIB, activin receptor type IIB; ALK4, activin type I receptor or activin-like kinase 4; SAA, solvent accessible surface area; H-bond, hydrogen bond; PPI, protein–protein interaction; VDW, van der Waals.

Additional Supporting Information may be found in the online version of this article.

Lincong Wang and Yaqin Hou contributed equally to this work.

*Correspondence to: Lincong Wang, The College of Computer Science and Technology, Jilin University, Changchun, Jilin, China. E-mail: wlincong@gmail.com (or) Yongli Bao, National Engineering Laboratory for Druggable Gene and Protein Screening, Northeast Normal University, Changchun, Jilin, China. E-mail: baoyl800@nenu.edu.cn

Introduction

The delicate balance between the binding affinity and specificity of a protein–protein interaction (PPI), perfected through the evolutionary process, is a prerequisite for the proper functionality of all the biological processes such as metabolic and signal transduction pathways. A deviation in either affinity or specificity of an interaction from its norm in healthy individuals could lead to diseases. Consequently, the modulation of PPIs using therapeutic agents has become one of the most active drug discovery efforts.¹ A starting point of such an effort is an intermolecular interface (binding site) with atomic details. Over the past 20

years, a variety of experimental methods and computational approaches have been developed for the detection, quantification, and prediction of PPI sites.^{2,3} It has been found that most of the studied PPI sites share common features with each other including (1) a small percentage of the residues in the binding site contribute disproportionately to the binding, and these residues are called hot-spot residues (or hot spots in short)^{4,5}; (2) they are composed of a hydrophobic core surrounded by an O-ring of charged/polar residues^{4,5}; and (3) there exists amino acid preference in hot spots.⁵ The focus of the present article is on hot-spot determination, and thus in the following we only review the current state of their experimental measurement and *in silico* prediction. Previously, given a complex structure their hot spots have been identified experimentally through alanine scanning mutagenesis followed by the measurement of the binding affinity of each mutant.⁴ However, large-scale alanine scanning is rather costly. Furthermore, for the majority of PPIs there may exist no complex structures, even no apo-structures for any of the two interacting partners. Consequently, in the last 10 years or so different computational approaches have been proposed for hot-spot prediction including both structure-based^{6–19} and sequence-based^{8,20} ones. However, the overall accuracy of sequence-based methods remains low, so most of the current approaches require a complex structure that is determined either experimentally or computed using a protein–protein docking program. Given a protein–protein interface, these approaches perform alanine scanning *in silico* to predict hot spots through the estimation of the binding affinities of the mutants generated for all the residues at the PPI site using an empirical scoring function. The following survey focuses on them.

A computational approach for hot-spot prediction consists of at least two components either explicitly or tacitly: a scoring function devised to distinguish hot-spot residues from regular surface residues and a search algorithm designed to find all the hot spots. The previous search algorithms are often heuristic in nature and adopted from machine learning such as neural networks,¹⁶ support vector machine and its variants.^{13,19} Even when a protein–protein docking program is used to find all the binding sites and poses, the search may still not be exact if the docking program itself relies on a heuristic search.⁸ A search algorithm is employed for both the tuning of the parameters in a scoring function and hot-spot identification. In theory, a scoring function is an approximation to the exact physics for a PPI expressed by $\Delta G = \Delta H - T\Delta S$ where ΔG is the binding Gibbs free energy. Neither enthalpy ΔH nor entropy ΔS could be measured or computed precisely. The explicit forms of current scoring functions may differ but all of them use a very crude approximation to ΔS and a somewhat more accurate approximation to

ΔH . Part of ΔS could be estimated through a quantification of the desolvation effect by computing the difference in solvent accessible surface area (SAA) before and after the binding. According to how ΔH and ΔS are approximated and sequence information is used, current scoring functions could be roughly divided into four categories.

1. Energy-based functions that consist of such terms as hydrogen bond (H-bond), salt bridge, van der Waals (VDW) and electrostatic interactions as well as SAA^{6–8,19};
2. Knowledge-based that include such terms as atomic (residue) contact and distance-dependent pair potential,⁹ the shape of the binding site, and the local structural environment such as structural conservation.¹⁰ These terms are typically mined from either a PPI or hot-spot database¹¹;
3. Sequence-based that have such terms as the degree of residue conservation¹² and a sequence profile that includes neighboring residues;
4. Ad hoc functions that include terms from both energy-based and knowledge-based. For example, such a function may combine an SAA term with residue conservation,¹³ residue contact,¹⁴ and pair potential of interface residues¹⁵ or distance-dependent pair potential,⁹ or combine shape specificity with atomic contact, H-bond, and salt bridge.¹⁶

In contrast to the terms in an energy-based function those in either a knowledge-based or sequence-based function have neither clear physical meaning nor proper energy unit. The parameters in these functions such as their weighting factors are learned through a training process using a database of complex structures either determined experimentally^{6,17–19} or predicted by a protein–protein docking program such as ICM.⁸ With an ad hoc scoring function that includes terms from different categories, it is tricky to determine its parameters and to assign proper energy unit to its terms. Though the individual terms in an energy-based function may have clear physical meaning, the function itself is still a very crude approximation to the overall Gibbs free energy ΔG of binding. The other scoring functions, though whose terms lack clear physical meaning, have been adopted mainly because of their good performance in practice. The development of both scoring functions and search algorithms relies heavily on experimental alanine scanning databases such as ASEdb,²¹ BID,²² and ISIS²³ for their parameter optimization and prediction verification.

Although alanine scanning mutagenesis is an established technique for hot-spot identification, the results are somewhat difficult to interpret quantitatively as a point mutation to a single type is not always sufficient to quantify a residue's contribution

Figure 1. The five top-scored alantolactone binding sites on human ActRIIB. A ternary complex structure of human bone morphogenetic protein 2 (BMP2, magenta), the extracellular domain of bone morphogenetic protein receptor type IA (cyan), and the extracellular domain of ActRIIB (green) (PDBID 2H62). The five sites are shown with an alantolactone molecule drawn as thick sticks. Sites III and IV are located on the interface between the extracellular domain of bone morphogenetic protein receptor type IA and extracellular domain of ActRIIB. Sites II and V are located near the lower part of ActRIIB and thus it is likely that if the human Cripto-I binds there, they may interact directly or collide with bone morphogenetic protein receptor type IA. Site I with the best overall score is located on the top right of ActRIIB, where if Cripto-I does bind, it will likely make contact with neither BMP2 nor bone morphogenetic protein receptor type IA. The figure is prepared using the program pyMol.

to the binding. In fact, it is rather tricky to quantify the contribution of any individual residue to either a PPI or protein–ligand interaction through point mutation without in the same time measuring the structural and dynamical changes brought about by the mutation.²⁴ Thus, it is not straightforward to translate the knowledge of alanine scanning into the design of small chemicals that target the PPI because of the lack of specific and explicit information about the particular intermolecular interactions that have been modulated by the point mutation.

In this article, we present a novel computational strategy for hot-spot determination through the docking into a set of proteins homologous to one of the two interacting partners of a compound known to be capable of disrupting a PPI. The rationale behind our approach is in line with the recent findings that PPI sites have increased propensity for small molecule (ligand) binding.^{25,26} We have developed an ad hoc scoring function that includes intermolecular hydrogen bonding propensity, intermolecular VDW energy and SAA as well as residue conservation first to predict PPI sites and then to distinguish hot spots from the other surface residues. In our approach, the

search for a PPI site is achieved through the docking of the known compound into a set of homologous structures while the hot spots themselves are determined by evaluating their specific interactions with the compound and the degree of conservation. Compared with the previous structure-based computational methods for hot-spot identification, our approach offers several advantages. First, only a homologous structure of one of the two interacting proteins is required. Second, because it starts with experimental observations, the prediction is more reliable, as having been confirmed experimentally through point mutation studies and by a comparison with a prediction made with a frequently used protein–protein docking program, ZDock.²⁷ Third, the explicit atomic interactions between the docked compound and binding site residues make it possible to mutate the key residues according to the particular type of interaction they engage in, rather than to mutate all of them to a single type residue as in an alanine scanning approach, and thus the quantification of a particular residue's contribution to the PPI should be more precise than what could be possibly achieved through alanine scanning. Fourth, as our strategy is based on a competitive mechanism between the ligand binding and protein binding, the explicit types of interactions identified by our approach could become the starting point for the optimization of lead compounds that target the PPI. Our novel approach has been applied to predict the hot spots of human activin receptor type II (ActRII) critical for its binding toward Cripto-I. Both ActRII and Cripto-I are involved in nodal and activin signal pathways. However, it has not been proved with certainty whether Cripto-I binds directly to ActRII or whether the interaction between Cripto-I and ActRII is mediated by activin. The experimental confirmation of human ActRII and Cripto-I binding interface and ActRII hot spots first reported in this article provides concrete evidence for their direct interaction.

Results

Our novel computational approach for hot-spot determination using a small molecule, alantolactone, as a probe has been successfully applied to the identification of residues of human ActRII critical for its binding to Cripto-I. In the following, we first describe the ActRII and Cripto-I interface determined for the first time by our approach. Then, we present a quantitative experimental assessment of the identified hot spots using a mammalian two-hybrid system with luciferase as a reporter.

Computational identification of human ActRII PPI site and hot spots

Out of all the possible alantolactone binding sites (Fig. 1) on the set of ActRIIA homologs (Table I) output from the docking program GOLD, a site (Site I)

Table I. The List of Four Homologous Protein Structures used in the Docking Study.

Protein complex	PDBID	Sequence identity (extracellular)	Sequence identity (whole)
BMP-2/ALK-3/ActRIIB ^a	2H62	62	62
BMP-2/ALK-3/ActRIIB ^a	2H64	63	62
ActRIIB	2QLU	35	77
BMP-2/ALK-6	3EVS	29	29

The four structures, three of them are complexes, are selected from the PDB based on sequence similarity of both their extracellular domain and the whole sequence.

^a The two complexes have similar structures with different resolutions.

on human activin receptor IIB (ActRIIB, PDBID 2H62, 62% sequence identity to human ActRIIA) has the best (lowest) overall score as determined by our scoring function (Table II, and Supporting Information S1). This site is composed of residues W45, R46, N47, G50, T51, I52, V74, Y82, F95, T96, and H98 of ActRIIB (Fig. 2). Assuming that both ActRIIA and ActRIIB interact with alantolactone in a similar manner and thus share hot spots, we could use the ActRIIB interface to infer the contributions of the corresponding residues in ActRIIA to its binding toward Cripto-I.

We start with an analysis of the top-scored ActRIIB alantolactone binding interface in terms of the type of interaction the residues in the interfaces could possibly engage in. In the ActRIIB binding interface, there exist an H-bond between O34 of alantolactone and HE1 of W45 and two H-bonds between O29 of alantolactone and the carboxamide group of N47. Both HE1 of W45 and the carboxamide group of N47 are exposed without the bound compound. A comparison of human ActRIIA and

Figure 2. The top-scored alantolactone binding interface on human ActRIIB. The ball-and-stick model represents alantolactone while the line model depicts the atoms in the binding site of ActRIIB that are within 6 Å of at least one atom of the docked compound. These atoms are from residues W45, R46, N47, G50, T51, I52, V74, Y82, F95, T96, and H97. Three dash lines indicate, respectively, intermolecular H-bonds between O29 and O34 of alantolactone and the carboxamide group of N47 and HE1 of W45. Together with H97, their contributions to the binding have been assessed by single point mutation. The atoms C, O, N, and H in the protein are colored, respectively, in green, red, blue, and white while the C, O, and H atoms of alantolactone are colored, respectively, in cyan, magenta, and olive. The figure is prepared using our own molecule visualization program.

ActRIIB sequence reveals that both residues are conserved (Fig. 3). We hypothesize that the side chains of both residues could participate in hydrogen bonding with Cripto-I while the binding of an alantolactone renders them unavailable for Cripto-I binding. With their statuses as hot spots having

Table II. The Ranking Scores of the Five Binding Sites.

Site	VDW (E_v)	H-bond energy(E_h)	Surface SAA of bound ligand (S_{aa})	Sequence conservation (S_c)	Is there an O-ring (O_r)	S_T
I	-0.43	-0.79(2) ^a	-0.36	10/11 ^b	-0.5 (yes)	-2.99
II	-1.01	-0.30(1)	-0.16	9/14	0.0 (no)	-2.11
III	-0.38	0.0(0)	-0.47	11/14	0.0 (no)	-1.64
IV	-0.16	0.0(0)	-0.69	6/6	0.0 (no)	-1.85
V	-0.50	0.0(0)	-0.23	10/13	0.0 (no)	-1.50

The total score, S_T , and its individual terms, $S_T = E_v + E_h + S_{aa} + S_c + O_r$, are computed as follows. The protein-ligand VDW term, E_v , is computed using CHARMM force field parameters²⁸ for the protein and AMBER force field parameters for the ligand²⁹. The protein-ligand H-bond energy, E_h , is computed using the parameters as implemented in the program DSSP³⁰ a weighting factor of 10^{-4} is applied to the raw H-bond energy in order to bring its numerical values into the range -1.0 and 0.0. The term that approximates the desolvation effect, S_{aa} , is defined as the ratio of the change in SAA versus the total SAA of all the hydrophobic atoms in the binding site, $S_{aa} = \frac{A_b - A_u}{A_u}$, where A_u is the total surface SAA of all the hydrophobic atoms of the unbound ligand, that is 278.6 Å² for alantolactone, A_b is the total surface SAA of all the hydrophobic atoms of the bound alantolactone. S_c is defined as the minus of the percentage of the conserved residues with respect to all the residues in the binding site; O_r is assigned -0.5 if at least one residue in the binding site is a part of an O-ring, otherwise, is assigned 0.0.

^a The first number is H-bond energy while the second is the number of H-bonds.

^b The first number is the number of the conserved residues in the binding site while the second is the total number of the residues. The binding site residues are those that have at least one atom to be within 6 Å of at least one atom of the docked compound. The lower the total score S_T is, the better.


```

6  AILGRSETQECLFFNANWEKDRTNQTGVEPCYGDKDKRRHCFATWKNIS
+  G +ET+EC+++NANWE +RTNQ+G+E C G++DKR HC+A+W+N S
SGRGEAETRECIYYNANWELERTNQSGLERCEGEQDKRLHCYASWRNSS

50 GSIEIVKQGCLWDDINCYDRTDCVEKKDSPEVYFCCCEGNCNEKFSYF
G+IE+VK+GCWLDD NCYDR +CV +++P+VYFCCCEGN CNE+F++
GTIELVKKGCWLDDFNCYDRQECVATEENPQVYFCCCEGNFCNERFTHL

```

Figure 3. The sequence similarity between human ActRIIA and ActRIIB as determined by BLAST. The target sequence is ActRIIA with ActRIIB as a query sequence.

been confirmed experimentally through point mutation study, we extend into the neighboring region of the above ligand binding site to search for additional hot spots. The Cripto-I binding interface of an ActRII is expected to include the alantolactone binding interface as a subregion, and most likely extend further into the neighboring region. A visual inspection of the possible extensions indicates that if the above alantolactone binding region is expanded into its right side as shown in Figure 4, then the Cripto-I binding interface of ActRIIA would be formed by a valley of hydrophobic residues surrounded by an O-ring composed of polar/charged residues, a common structural feature in many PPI interfaces.⁵ Moreover, the extended binding site includes W45, Y82, and R94, the three types of residues that have been

Figure 4. The binding site isosurface of ActRIIB (ActRIIA) toward Cripto-I. The color scheme for both the protein and ligand is the same as that in Figure 2 except that the carbon atoms in protein surface are colored in white. The isosurface itself is computed by the marching cube algorithm while each point is colored according to the color of the protein atom that is the closest to the point. The residues colored in green are, respectively, W45 (covered by alantolactone), N47, S49, R94, and H97. The PPI site is formed by an O-ring of polar residues as shown in red color and a hydrophobic center as shown in white color. The small balls and sticks depict, respectively, protein atoms and bonds. The figure is prepared using our own molecule visualization program.

reported to occur frequently in PPI sites.⁵ In this extended Cripto-I binding site, N47 is a part of an O-ring while W45 in the valley. Other residues in the O-ring include both S49 and R94 with their side chains possibly forming either an H-bond or salt-bridge with Cripto-I because their side chains extend far into the solvent. In particular, S49 is conserved and its hydroxyl group could not form any intramolecular H-bonds. The residue in ActRIIA corresponding to R94 of ActRIIB is K94, both of them are positively charged and thus possibly form a salt-bridge with Cripto-I, while Y97 of ActRIIA corresponds to H97 of ActRIIB, both of them are aromatic and thus possibly engage in a hydrophobic or aromatic-aromatic interaction with Cripto-I. Y97(H97) covers the left side of the alantolactone binding site (Figs. 2 and 4) while K94(R94) is located at its upper left (Fig. 4). All of them may contribute to Cripto-I binding and thus are selected for experimental verification and quantitative assessment as detailed later.

Experimental validation and quantification of hot-spot residues

The binding affinities of both the wild-type and five mutants of ActRIIA toward Cripto-I have been measured by a mammalian hybrid system with luciferase as a reporter (Fig. 5). The data show that N47V mutant retains less than 20% affinity for the wild-type. Though on average 70% affinity remains for W45F mutant, none of the measured values have more than 90% of the affinity for the wild type

Figure 5. The luciferase activities of five mutants. The X-axis labels, respectively, the wild-type ActRIIA and the five mutants while the Y-axis denotes their binding affinity as measured by the luciferase reporter assay (a mammalian two-hybrid system assay) with the average value for the wild-type being normalized to 1.0. Each measurement of a mutant affinity is repeated for at least three times and the reported value is their average with the range indicated by a bar. The three mutants, W45F, N47V, and S49A whose highest activities are still lower than that for the wild-type, are shown with **.

(Fig. 5). The W45F and N47N mutants are designed to disrupt their capability to form a hydrogen-bond. It is possible that W45F mutant retains its aromatic–aromatic interaction with Cripto-I and thus has smaller reduction in affinity compared with that of N47V mutant. We predict that both of them are hot spots though W45 is likely to be a weak one. To quantify the contributions of the three residues S49, K94, and Y97 at the extended ActRIIA PPI site, S49 is mutated to an alanine to remove its hydrogen bonding propensity, K94 to a methionine to cut out its capacity to form a salt-bridge and Y97 to an alanine to disrupt a possible aromatic–aromatic interaction or to reduce a possible hydrophobic interaction with Cripto-I. Compared with the binding affinity of the wild-type, the affinity of S49A is reduced by 60%, that of Y97A by 10% while the binding affinity of R94M remains more or less the same. Thus, we conclude that S49 and possibly Y97(H97) are hot spots while K94(R94) may be not.

Discussion

We first present the rationale behind our scoring function on which our computational approach for hot-spot identification depends. Next, we compare our approach with previous structure-based computational methods for hot-spot identification. We then discuss the biological implications of the identified hot spots of human ActRII critical for Cripto-I binding and highlight the significance of our computational approach for the design of small molecule drugs targeting a PPI. Finally, we discuss the limitations and possible extensions of our approach.

The scoring function

Our approach assumes that there exist only homologous structures for one of the two interacting proteins. Consequently, in contrast with the computational alanine scanning methods that require a complex structure, in our case no PPI energies such as intermolecular VDW or H-bond energy could be computed. On the other hand, given a pose from a docking program, the protein–ligand interaction energy is readily available. With the assumption that both the ligand and protein compete for the same site, it is possible to use protein–ligand binding affinity to estimate protein–protein binding affinity. Thus, for the purpose of PPI site identification, we have developed an ad hoc scoring function, $S_T = E_v + E_h + S_{aa} + S_c + O_r$, that is composed of two types of terms that occur respectively in most protein–ligand and protein–protein scoring functions. The first three terms, E_v , E_h and S_{aa} , occur frequently in many scoring functions for both protein–ligand docking^{31,32} and protein–protein docking as well as hot-spot identification^{6–8} because of their clear physical meaning; while the next two terms, S_c and O_r , are typical in most knowledge-based scoring

functions for PPI ranking and hot-spot identification¹² though their concrete forms for term O_r may differ from our expression. The ad hoc combination of terms from both protein–ligand interaction and PPI seems to be odd but the idea is consistent with the previous findings that a PPI site has high propensity for ligand binding.^{25,26} The practice of combining physical terms with nonphysical ones is not new: in structural biology especially in NMR structure determination a typical target function includes both molecule-mechanics terms and pseudo-energy terms that assign a penalty to restraint violation.^{33,34} However, it is challenging to determine weighting factors and the associated energy units for individual terms in such an ad hoc function without first training it against a database of PPIs that could be inhibited by chemicals. At present, we are not aware of the existence of any such databases. Consequently, to keep the scoring function as simple as possible the numerical value for each term is scaled to the range between -1.0 and 0.0 except for VDW energy E_v whose value is scaled to the range from -1.0 to $+1.0$.

The binding site search method and hot-spot identification

Our approach assumes that the ligand and protein share the same binding site, thus it is possible to use a protein–ligand docking program to search for a PPI site. The initial set of binding sites is identified by the program Ligsite³⁵ and then the set is pruned by the protein–ligand docking program GOLD³¹: if GOLD fails to output any pose for the site, it is removed from any further considerations. The accuracy of our PPI site search and hot-spot determination relies critically on the accuracy of GOLD that itself employs a genetic algorithm to optimize a protein–ligand scoring function. To reduce the possibility of missing the correct binding sites, three rounds of docking experiments are performed with each to be initialized with a different center for a binding site. If any pose is output from GOLD, the site is kept for hot-spot identification. When a pose is used to identify hot spots, only its key intermolecular interactions such as intermolecular hydrogen bonding is used in order to minimize the effect of the possible low accuracy of the docked poses on hot-spot determination.

Comparison with the previous structure-based computational methods for hot-spot identification

Our novel computational strategy determines the hot spots of a PPI interface through the docking of a compound known to be capable of disrupting a PPI into a set of proteins homologous to only one of the two interacting partners. Compared with previous computational methods, our approach offers several

advantages. First, no complex structures not even the exact structures of any individual proteins are required. In contrast, the previous structure-based hot-spot prediction methods require either a complex structure or the structures of individual proteins, for the latter a protein–protein docking must first be performed to search for the possible interfaces between the two proteins.³⁶ Second, because our approach starts with an experimental observation, the predicted hot spots are more accurate than those achieved by the previous methods as being confirmed experimentally through point mutation study. For the purpose of comparison, we have used a protein–protein docking program ZDock²⁷ with the default parameters provided by the program to dock an ensemble of NMR solution structures of mouse Cripto-I³⁷ into human ActRIIB (PDBID 2H62) used in our protein–ligand docking experiment. None of the poses output from ZDock could correctly identify any of the above hot spots (Supporting Information S2). Third, the specific intermolecular interactions with atomic details between the docked ligand and individual residues obtained by our approach make it possible to mutate the involved residues according to the specific type of interaction they engage in while in the same time to keep all their other physical–chemical properties as intact as possible. In contrast, the experimental alanine scanning itself and the computational methods built upon it mutate every residue to an alanine, consequently the quantification of a particular residue's contribution to the PPI may be less precise than that obtained from our approach.

Implications for nodal and activin signal pathways and tumorigenesis

It has been shown that nodal signals via ActRIIA/ActRIIB and ALK4 (activin-like kinase 4 or activin type I receptor), and is essential for mesoderm formation and subsequent organization of left–right axial structures in early embryonic development. Specifically, nodal requires Cripto-I as a coreceptor for signaling via ActRIIA/ActRIIB and ALK4. Although previous studies have indicated that Cripto-I could form a trimer with activin and ActRII,^{38,39} and may interact directly with human ActRIIA,⁴⁰ it has not been proved with certainty whether Cripto-I binds directly to ActRIIA/ActRIIB or whether the interaction between Cripto-I and ActRIIA/ActRIIB is mediated by activin as no complex structures have been reported and neither has the interface between Cripto-I and ActRIIA/ActRIIB nor the interface between Cripto-I and activin been mapped experimentally. Through a screening effort using a mammalian two-hybrid system, alantolactone has been identified previously⁴⁰ as a possible candidate capable of disrupting Cripto-I and ActRIIA interaction. The conclusion is based on a co-

PI experiment of Cripto-I and ActRIIA and the following experimental observation: alantolactone could induce activin/SMAD3 signaling in human colon adenocarcinoma HCT-8 cells. If Cripto-I could bind directly to ActRIIA, it should play an essential role in both activin and the nodal signal pathways. It has been well documented that activin controls diverse cellular processes ranging from tissue patterning during embryogenesis to the control of homeostasis, cell growth, and differentiation in many types of adult tissues. The malfunction of the activin signaling is associated with multiple pathological states including reproductive disorders and carcinogenesis. Thus, the role of Cripto-I in tumorigenesis should be consistent with the facts that (1) it could inhibit the activin signaling and (2) alantolactone could reverse the inhibition through the induction of activin/SMAD3 signaling by blocking the Cripto-I and ActRII binding. However, up to now neither is known whether alantolactone binds to either ActRIIA or Cripto-I nor has the alantolactone or Cripto-I binding site on ActRIIA been characterized. The alantolactone binding site, the interface between ActRIIA and Cripto-I and the hot spots on ActRII all determined for the first time by our computational approach provide the concrete evidence that supports a direct though possible transient interaction between human ActRII and Cripto-I and further supports the view that Cripto-I may play an important role in the activin signal pathway as well.

The significance for the design of small molecules that modulate PPIs

The delicate balance between the binding affinity and specificity of a PPI plays an essential role in the development of many diseases. Consequently, the modulation of PPIs using small molecules has become one of the most active drug discovery efforts. However, compared with the structure-based design of small molecules that target an enzyme or a ligand-binding protein, the design of small molecule PPI inhibitors is more challenging because of (a) the lack of complex structures, (b) the lack of definite information about both hot spots and the type of interaction at atomic level, and (c) the unique characteristics of a typical PPI interface such as being noncontiguous, flat, and devoid of a deep binding site, and its surface area being much larger than the potential binding area of a typical small molecule.⁴¹ Our computational approach may make it easier to design small molecule inhibitors because it could determine hot spots with high accuracy even when neither a complex structure nor the structures of any individual proteins are available and because the probe itself is a small molecule that has been experimentally proved to be capable of disrupting the PPI. Moreover, the hot spots and the detailed information about the PPI determined by our

approach should be particularly useful for lead optimization with the probe itself as a scaffold and being guided by the specific types of intermolecular interactions. In our particular case, with the previous experimental observation that alantolactone exhibits antiproliferative function specific to tumor cells with almost no toxicity to normal cells at a concentration of up to 5 $\mu\text{g/mL}$,⁴⁰ alantolactone may well become a promising lead for the design of an entirely new class of anticancer agents that target the activin signal pathway.

The limitations of our approach and possible extensions

In the current implementation of our approach, only limited flexibility of the shared ligand and protein binding site is taken into consideration as the current version of GOLD only lets the methyls of the binding site and the protein atoms that could possibly participate in hydrogen bonding with a ligand atom have multiple conformations. A possible remedy is first to use molecular dynamics simulation to explore the mobility of the binding site and then to dock the compound into the representative trajectories, and finally to rank all the possible binding sites. Our approach assumes that the ligand and the protein compete for the same binding site, and thus it will fail if the ligand achieves its inhibition through an allosteric effect. On the other hand, our approach could be used to determine whether the ligand achieves its inhibition through an allosteric mechanism as in this case, the point mutation of a ligand binding site residue will not affect the PPI in question. A possible disadvantage of our approach is that it requires the existence of both a compound capable of inhibiting the PPI and an apo-form structure of a protein homologous to one of the two interacting partners and thus only the hot spots on the protein that has a homologous structure could be determined. Finally, though the newly developed scoring function and the computational approach itself have been applied successfully to determine the hot spots of ActRII toward Cripto-I, for better assessment of both the scoring function and computational approach itself, they should be tested on more PPI systems that could be interrupted by a ligand. In particular, the fine tuning of our ad hoc scoring function including the determination of its weighting factors and the associated units requires a database of PPIs that could be disrupted by known compounds.

Materials and Methods

We first present our novel computational approach for hot-spot determination. Then we illustrate its application to human ActRIIA that binds to Cripto-I. Finally, we describe the experimental verification of our computational results.

The computational approach

Our approach does not require any complex structures, instead it searches for hot spots using as a probe a compound known to be capable of modulating the interaction between the two proteins whose hot spots we are seeking to identify. Briefly, our approach proceeds as follows. Starting with the structures of a set of proteins homologous to only one of the two binding partners, we use a protein–ligand docking program to dock the probe into all the possible binding sites on all the homologs. These binding sites are identified initially with the program Ligsite.³⁵ Those sites into which the program failed to dock the compound are removed from further considerations while all the remaining sites are ranked using our newly developed scoring function. The scoring function is composed of intermolecular H-bond energy, intermolecular VDW energy, and the change in solvent SAA upon ligand binding as well as two additional terms that represent, respectively, the degree of residue conservation and whether the ligand binding site is a part of an O-ring structure characteristic of most known PPI sites. Finally, the conserved residues in the top-ranked site that make important contributions to the ligand binding are predicted to be hot spots. In the following, we present our approach in detail.

The selection of protein structures for docking study. There exist no complex structures for ActRIIA and Cripto-I. Instead, we use a set of crystal structures of activin receptors (Table I) that are similar to ActRIIA at the sequence level.

Protein–ligand docking for PPI site identification. We first use the program Ligsite³⁵ to identify all the possible binding sites on each structure in the set of homologs. Then, we perform several rounds of docking studies on each binding site of each homolog using the program GOLD³¹ initialized with three distinct centers manually picked for a binding site. GOLD requires a user to pick a point as the center of a binding site and the final poses depend, sometime critically, on the location of the center. We use the default parameters as provided by GOLD. Specifically, the number of final poses output from GOLD is decided by the program itself using its default fitness score. During the docking process, except for methyl groups and those atoms that possibly participate in hydrogen bonding with the ligand, all the other atoms in a binding site are kept to be rigid while the ligand itself is flexible.

The scoring function for binding site ranking. The following scoring function, S_T , has been developed to rank all the binding sites on all the

homologs output from the above protein–ligand docking study:

$$S_T = E_v + E_h + S_{aa} + S_c + O_r$$

where E_v and E_h are, respectively, the intermolecular protein–ligand VDW and the protein–ligand H-bond energies, S_{aa} is the ratio of the change in SAA versus the total SAA of the hydrophobic atoms, $S_{aa} = \frac{A_b - A_u}{A_u}$, where A_u is the total surface SAA of all the hydrophobic atoms of the unbound ligand, A_b is the total surface SAA of all the hydrophobic atoms of the bound ligand. S_c is the percentage of the conserved residues with respect to all the residues in the binding site, while O_r is assigned -0.5 if there exists an O-ring in the ligand binding site, otherwise, is assigned 0.0 . If there are several docking poses for a binding site, we compute S_T using the pose with the best GOLD fitness score.

All the terms in S_T are computed using our own implementation. The protein–ligand VDW term, E_v , is computed using CHARMM force field parameters²⁸ for the protein and the AMBER force field parameters for the ligand.²⁹ The term S_{aa} that approximates the desolvation effect of both protein–ligand and protein–protein binding is computed by following a procedure for SAA computation in the program DSSP.³⁰ The protein–ligand hydrogen bond term E_h is also computed using the parameters in the program DSSP,³⁰ and as its numerical values are rather large, they are divided by a constant to bring them into the range -1.0 to 0.0 (0.0 if there exist no H-bonds) in agreement with the numerical values for those of S_{aa} and S_c that are in the range -1.0 to 0.0 . To keep the scoring function as simple as possible, the numerical value for every term is scaled to the range between -1.0 and 1.0 . For example, the numerical value for O_r is set to either -0.5 or 0.0 . Except for VDW energy E_v , the numerical values for all the other terms are negative. The residues used in S_c computation are those that have at least one atom to be within $6 \text{ \AA}$ of at least one atom of the docked compound.

Hot-spot identification and the design of point mutation. Our approach assumes that a ligand accomplishes its inhibition of a PPI through a competitive mechanism, thus a hot spot is also pivotal for protein–ligand binding. In other words, a residue in the top-ranked site that has important interactions with the docked compound such as an intermolecular H-bond or an aromatic–aromatic interaction is predicted to be a hot spot if it is also conserved and a part of an O-ring structure. These predicted hot spots are then assessed by point mutation study. After an initial binding region having been confirmed experimentally, we extend into its neighbor-

ing regions to further search for the conserved residues that could potentially form a salt-bridge or hydrogen bond or participate in an aromatic–aromatic interaction with a protein residue as a protein–protein interface is likely to be larger than a protein–ligand interface especially when the ligand itself is a small chemical. Point mutation is performed on these predicted hot spots as well to assess their contributions to the binding. The design of point mutation itself is based on the type of the interaction this particular residue is expected to participate in when it binds to the protein. For example, if a residue is expected to form an intermolecular H-bond, we mutate it to a residue that keeps as much the same as possible the other physical and chemical properties while the capacity of its side chain to form an H-bond is being removed.

Experimental verification and quantification of hot spots

The effect of a single point mutation on human ActRIIA-Cripto-I interaction is assessed using a mammalian two-hybrid system with luciferase as a reporter.

Cell line and cell culture. Human embryonic kidney cells (HEK293T) were obtained from the Cell Bank of the Chinese Academy of Sciences (Shanghai, China). The cells were maintained in Dulbeccos modified Eagles medium (DMEM, Gibco), supplemented with 10% fetal bovine serum (FBS, TBD, China), containing 100 U/ml penicillin and 100 mg/ml streptomycin and incubated in a 37°C incubator with 5% CO_2 .

Plasmids construction. An N-truncated cDNA sequence, that is, with its signal peptide sequence deleted, of either human ActRIIA or Cripto-1 is cloned into a pBind vector (Promega). The five mutants, W45F, N47V, S49A, K94M, and Y97A, are generated using a Site Directed Mutagenesis Kit by following a protocol provided by the manufacturer (Beyotime, Jiangsu, China). Briefly, the five DNA constructs (pBind-ActIIA W45F, pBind-ActRIIA N47V, pBind-ActRIIA S49A, pBind-ActRIIA K94M, and pBind-ActRIIA Y97A) are first created and then amplified by PCR using the respective primers. All the five constructs are then transferred into *E.coli* cells for culture and finally their cDNAs are extracted from the cells and sequenced to confirm the mutation.

Luciferase reporter assay. HEK 293T cells are grown for about 24 h until they achieve a 60–70% confluence when they are ready for cotransfection by plasmids. The HEK293T cells are then cotransfected by three types of plasmids ($1 \mu\text{g}/\text{well}$) having respectively human pAct-Cripto-1 DNA, pG5luc luciferase DNA, and either the wild-type pBind-ActRIIA DNA

construct or one of the above five point mutation constructs. The cotransfection is achieved using a Calcium Phosphate Transfection Kit (Beyotime, Jiangsu, China) according to the manufacturer's recommended protocol. After the cotransfection, the HEK 293T cells are further cultured for an additional 24 to 48 h prior to a luciferase reporter assay. The assay itself proceeds as follows. The cells are first harvested and washed twice using phosphate-buffered saline (PBS), and lysed in 85 μ L cold lysis buffer (25 mmol/L glycylglycine, 1% Triton X-100 [v/v], 15 mmol/L MgSO_4 , 4 mmol/L ethylene glycol tetra-acetic acid [EGTA], and 1 mmol/L dithiothreitol, pH 7.8). Then, luciferin (100 μ L, 25 mmol/L, BD Monolight; BD Biosciences, Pharmingen, San Jose, CA), potassium salt (20l 0.5M KH_2PO_4) and 5 μ L assay cocktail (1 mol/L adenosine triphosphate, 15 mmol/L KH_2PO_4 , 15 mmol/L MgCl_2 , pH 7.8) are added to each 45 μ L cell lysate. Finally, the luciferase activity and β -Galactosidase activity are measured, respectively, using a FLUOstar OPTIMA fluorescence spectrometer (BMG LABTECH, Offenburg, Germany) and a β -galactosidase assay system (Promega Biosciences, San Luis, Obispo, CA).

Conclusions

The hot spots of a protein quantify its binding affinity toward another protein at the individual residue level with atomic details, and thus their identification is the starting point for the design of small chemicals targeting the PPI. However, the experimental determination of hot spots is time-consuming and resource demanding while the accuracy of the current computational methods for hot-spot identification remains low. In this article, we presented a novel structure-based computational approach for hot-spot determination through the docking of a compound known to be able to inhibit the particular PPI into a set of protein structures homologous to only one of the two interacting partners. The subsequent experimental confirmation of the predicted hot spots suggests its potential as a general method for hot-spot determination while the specific types of interactions identified by the approach could be used to guide lead optimization with the compound itself as a scaffold. The confirmation also demonstrates the adequacy of our ad hoc scoring function for hot-spot determination. The interfaces between human ActRIIs and Cripto-I as well as the hot spots on ActRIIs first determined by our approach support the view that Cripto-I plays an important role in both activin and nodal signal pathways.

REFERENCES

1. Thanos CD, DeLano WL, Wells JA (2006) Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc Natl Acad Sci* 103:15422–15427.
2. Shoemaker BA, Panchenko AR (2007) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol* 3:e42.
3. Shoemaker BA, Panchenko AR (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol* 3:e43.
4. Clackson T, Wells J (1995) A hot spot of binding energy in a hormone-receptor interface. *Science* 267:383–386.
5. Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol* 280:1–9.
6. Kortemme T, Baker D (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci* 99:14116–14121.
7. Moreira J, Fernandez-Recio IS, Ramos MJ (2007) Computational alanine scanning mutagenesis—an improved methodological approach. *Nat J Comput Chem* 28: 644–654.
8. Grosdidier S, Fernandez-Recio J (2008) Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics* 9:447–460.
9. Krger DM, Gohlke H (2010) Drugscore PPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions. *Nucl Acids Res* 38(Suppl 2):W480–W486.
10. Geppert T, Hoy B, Wessler S, Schneider G (2011) Context-based identification of protein-protein interfaces and hot-spot residues. *Chem Biol* 18:344 – 353.
11. Ma B, Elkayam T, Wolfson H, Nussinov R (2003) Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci* 100:5772–5777.
12. Hu Z, Ma B, Wolfson H, Nussinov R (2000) Conservation of polar residues as hot-spots at protein interfaces. *Proteins* 39:331–342.
13. Guney E, Tuncbag N, Keskin O, Gursoy A (2008) Hot-Sprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res* 36(Suppl 1):D662–D666.
14. Tuncbag N, Gursoy A, Keskin O (2009) Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics* 25:1513–1520.
15. Cukuroglu E, Gursoy A, Keskin O (2012) HotRegion: a database of predicted hot spot clusters. *Nucleic Acids Res* 40:D829–D833.
16. Darnell SJ, Page D, Mitchell C (2007) An automated decision-tree approach to predicting protein interaction hot spots. *Proteins: Struct Funct Bioinformatics* 68: 813–823.
17. Kortemme D, Kim T, Baker D (2004) Computational alanine scanning of protein-protein interfaces. *Sci STKE* 219: pl2.
18. Guerois J, Nielsen R, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* 320:369–387.
19. Lise S, Archambeau C, Pontil M, Jones D (2009) Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. *BMC Bioinformatics* 10:365.
20. Ofra Y, Rost B (2007) Protein-protein interaction hot-spots carved into sequences. *PLoS Comput Biol* 3:e119.
21. Thorn KS, Bogan AA (2001) A database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* 17: 284–285.
22. Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C, Ellison B, Galer S, Shapley J, Fuentes E, Tsai J (2003) The binding interface

- database (bid): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics* 19:1453–1454.
23. Ofra Y, Rost B (2007) ISIS: interaction sites identified from sequence. *Bioinformatics* 23:e13–e16.
24. Wang L, Li Y, Yan H (1997) Human cellular retinoic acid binding proteins: quantitative analysis of the ligand binding properties of the wild-type and site-directed mutants. *J Biol Chem* 272:1541–1547.
25. Davis FP, Sali A (2010) The overlap of small molecule and protein binding sites within families of protein structures. *PLoS Comput Biol* 6:e1000668.
26. Zerbe BS, Hall DR, Vajda S, Whitty A, Kozakov D (2012) Relationship between hot spot residues and ligand binding hot spots in protein–protein interfaces. *J Chem Inform Model* 52:2236–2244.
27. Chen R, Li L, Weng Z (2003) Zdock: an initial-stage protein-docking algorithm. *Proteins: Struct Funct Bioinformatics* 52:80–87.
28. MacKerell AD, Bashford D, Bellott, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102:3586–3616.
29. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. *J Comp Chem* 25:1157–1173.
30. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.
31. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 267:727–748.
32. Böhm HJ (1991) The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J Comput-Aided Mol Des* 6:61–78.
33. Wüthrich K (1986) *NMR of proteins and nucleic acids*. New York: InterScience.
34. Wang L, Zou S, Wang Y (2012) Algorithmic challenges in structure-based drug design and NMR structural biology. *Front Electr Electron Eng* 7:69–84.
35. Hendlich M, Rippmann F, Barnickel G (1997) Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* 15:359–363.
36. Fernandez-Recio J, Totrov M, Abagyan R (2004) Identification of protein–protein interaction sites from docking energy landscapes. *J Mol Biol* 335:843–865.
37. Calvanese L, Saporito A, Marasco D, D'Auria G, Minchiotti G, Pedone C, Paolillo L, Falcigno L, Ruvo M (2006) Solution structure of mouse Cripto CFC domain and its inactive variant Trp107Ala. *J Med Chem* 49:7054–7062.
38. Gray PC, Harrison CA, Vale W (2003) Cripto forms a complex with activin and type ii activin receptors and can block activin signalling. *Proc Nat Acad Sci* 100:5193–5198.
39. Kelber JA, Shani G, Booker EC, Vale WW, Gray PC (2008) Cripto is a noncompetitive activin antagonist that forms analogous signaling complexes with activin and nodal. *J Biol Chem* 283:4490–4500.
40. Shi Y, Bao YL, Wu Y, Yu CL, Huang YX, Sun Y, Zheng LH, Li YX (2011) Alantolactone inhibits cell proliferation by interrupting the interaction between cripto-1 and activin receptor type ii a in activin signaling pathway. *J Biomol Screen* 16:525–535.
41. Toogood PL (2001) Inhibition of protein–protein association by small molecules: approaches and progress. *J Med Chem* 45:1543–1558.