PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Analysis and network representation of hotspots in protein interfaces using minimum cut trees

Nurcan Tuncbag, F. Sibel Salman, Ozlem Keskin, and Attila Gursoy*

Center for Computational Biology and Bioinformatics and College of Engineering, Koc University, Istanbul, Turkey

ABSTRACT

We propose a novel approach to analyze and visualize residue contact networks of protein interfaces by graph-based algorithms using a minimum cut tree (mincut tree). Edges in the network are weighted according to an energy function derived from knowledge-based potentials. The mincut tree, which is constructed from the weighted residue network, simplifies and summarizes the complex structure of the contact network by an efficient and informative representation. This representation offers a comprehensible view of critical residues and facilitates the inspection of their organization. We observed, on a nonredundant data set of 38 protein complexes with experimental hotspots that the highest degree node in the mincut tree usually corresponds to an experimental hotspot. Further, hotspots are found in a few paths in the mincut tree. In addition, we examine the organization of hotspots (hot regions) using an iterative clustering algorithm on two different case studies. We find that distinct hot regions are located on specific sites of the mincut tree and some critical residues hold these clusters together. Clustering of the interface residues provides information about the relation of hot regions with each other. Our new approach is useful at the molecular level for both identification of critical paths in the protein interfaces and extraction of hot regions by clustering of the interface residues.

Proteins 2010; 78:2283–2294.

Key words: hotspot; mincut tree; residue contact network; visualization; hot region.

INTRODUCTION

Protein–protein interactions (PPI) are crucial for almost all biological processes. Proteins act coherently in the cells and function in several processes by interacting with other molecules using the binding regions on their surfaces. The binding free energy distribution is not uniform in protein interfaces; instead, some residues contribute more to the binding, called hotspot. A hotspot can be identified experimentally by evaluating the free energy change upon mutating it to alanine.[1–3] Experimental information is available only for a limited number of complexes, which are deposited in databases, such as Alanine Scanning Energetics Database (ASEdb)[4] and Binding Interface Database (BID).[5] Hence, a need for computational methods arises. Several studies identify computational hotspots using energy-based models,[6,7] learning-based models,[8–12] and molecular dynamics-based methods.[13–15] Graph/network-based algorithms are also frequently used to analyze protein–protein interfaces, such as identification, organization, and packing of hotspots. Brinda et al.[16] used graph representation of homodimeric protein complexes and applied spectral analysis to the residue networks to predict hot residues. del Sol and O'Meara[17] used the small-world network approach to predict hot residues in protein–protein interfaces. In their work, highly central residues are considered and they stated that 77% of the predicted residues, conserved and buried ones, are either experimental hotspot or in direct contact with an experimental hotspot. Haliloglu et al.[18] applied Gaussian Network Model on several antigen–antibody and enzyme–inhibitor complexes to predict anchoring residues.

Hotspots are clustered into tightly packed regions, called hot regions. Contribution of distinct hot regions to stability is additive, whereas contribution is cooperative within clusters.[19,20] Schreiber and co-workers[20] showed this statement on the TEM1-β-lactamase and its inhibitor protein (BLIP) complex. On the other hand, Moza et al.[21] stated the necessity of long distance interactions for PPI and showed the cooperativity between distinct hot regions, which are 20 Å far-away from each other in T cell receptor-a bacterial superantigen complex. Analysis of the hot regions carries importance, especially in drug design. If distinct hot regions are energetically cooperative, a therapeutic agent designed for one region might be affected by the other

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

*Correspondence to: Attila Gursoy, Koc University, Center for Computational Biology and Bioinformatics and College of Engineering, Rumelifeneri Yolu, 34450 Sariyer Istanbul, Turkey. E-mail: agursoy@ku.edu.tr.
Received 9 December 2009; Revised 10 March 2010; Accepted 27 March 2010
Published online 9 April 2010 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22741

hot regions. On the other hand, this does not hold for independent hot regions.

Typical residue contact networks have a complex structure consisting of many residues (nodes) and interactions (edges). Hence, it is difficult to identify, which of these residues and interactions are critical in terms of stability and function of protein complexes. We propose a novel approach to generate a simple, yet informative representation of residue contact networks of protein interfaces. We assign knowledge-based potentials as edge weights of the network. Our approach constructs a minimum cut tree (mincut tree) from the weighted residue contact network. We propose algorithms to extract hotspots and their organization from the mincut tree. Based on the minimum-cut/maximum-flow theorem, the residues identified by our approach have maximal energetic flow. Mutating such residues would change the energy flow significantly. In this method, the residue contributing to several mincuts is the most important one in terms of energy. Computational tests on a nonredundant dataset of protein complexes, having experimental mutation data, indicate that the most connected residue in the mincut tree generally corresponds to an experimental hotspot and other critical residues are observed to form a subtree. This method does not focus only on hotspot prediction and it cannot identify all hotspots, rather it is used for the analysis of the organization of interface residues and the connection between residues. Further, we show how to use mincut trees to cluster residues corresponding to hot regions in protein interfaces.

METHODS

Dataset of experimental hotspots

Experimentally, a hotspot can be identified by evaluating the change in binding free energy upon mutating it to an alanine residue.[2] ASEdb is an information source for the hotspots obtained via alanine scanning mutagenesis experiments.[4] Another database, namely the BID, contains experimentally verified hotspots in interfaces collected from the literature.[5] In this work, we use the protein complexes deposited in these two databases. For the complexes in ASEdb, the residues whose change in binding free energy is at least 2.0 kcal/mol are considered as hotspots. For the complexes in BID, the residues whose interaction is "strong" are considered as hotspots. Thus, totally a nonredundant set of 38 protein complexes are examined.

Construction of weighted residue contact graph and mincut tree of a protein complex

An undirected weighted residue contact graph $G(N,E)$ consists of a node set N and an edge set E, with positive weights w_e, for all $e \in E$. In this graph, nodes represent interface residues and edges between them represent the contacts between pairs of residues. Two residues, one from each chain, are in contact if the distance between any two atoms belonging to two residues is smaller than the sum of their van der Waals radii plus a 0.5 Å tolerance. Also, two residues, within one chain, are in contact if the distance between the C^{α} atoms of these residues is smaller than 6 Å.

The weights of the edges in the graph are obtained from knowledge-based solvent-mediated potentials derived by Keskin et al. in 1998,[22] which are in good agreement with the residue frequencies obtained in a recent work.[23] The knowledge-based potentials have been shown to be useful in many threading, folding, and binding problems.[24–26] These potentials represent the interaction parameters between two residues in native proteins. A practical way to obtain these potentials is to extract them from frequencies of contacts between different residues in proteins with known three-dimensional (3D) structures.[27] We provide 210 distinct potentials (all possible pairs of 20 different aminoacids) in RT unit (R universal gas constant, T is temperature) for contacting residue pairs in the Supporting Information. All of the entries in this matrix are negative valued. In the residue contact network, the absolute value of the corresponding entry in the pair potential matrix is used as the edge weight (w_e).

A cut in a connected graph is defined by a partition of the node set into two sets, and consists of all edges that have one endpoint in each partition. Clearly, the removal of the cut disconnects the graph. The weight of a cut is the sum of the weights of the edges crossing the cut. For $s,t \in N$, an s-t cut is defined as a cut, which puts s and t into different node sets of the partition. A minimum weight s-t cut (min s-t cut) is a subset of edges with minimum total weight that separates the network into at least two disconnected sets of nodes. The problem of finding a min s-t cut can be efficiently solved using a maximum flow algorithm.[28] In the residue contact graph, the minimum weight cut between two residues illustrates the minimum total contact potential to separate these two residues into two disconnected subgraphs. Furthermore, min s-t cuts for all pairs of nodes can be represented by a mincut tree in a compact way so that both the weight of a min s-t cut in the graph and the corresponding partition is the same in the tree. Gomory and Hu showed that a mincut tree can be computed using only $n - 1$ min s-t cut computations (that finds the maximum flow from s to t), where n is the number of nodes.[29] To construct a mincut tree, G should be a connected network. If this is not the case, we take the largest connected component in G and perform our calculations on this graph. The algorithm to construct a mincut tree can be found in,[29] and alternatively in.[30,31] Here, we only demonstrate it with a simple example shown in Figure 1. First, all nodes are considered as a

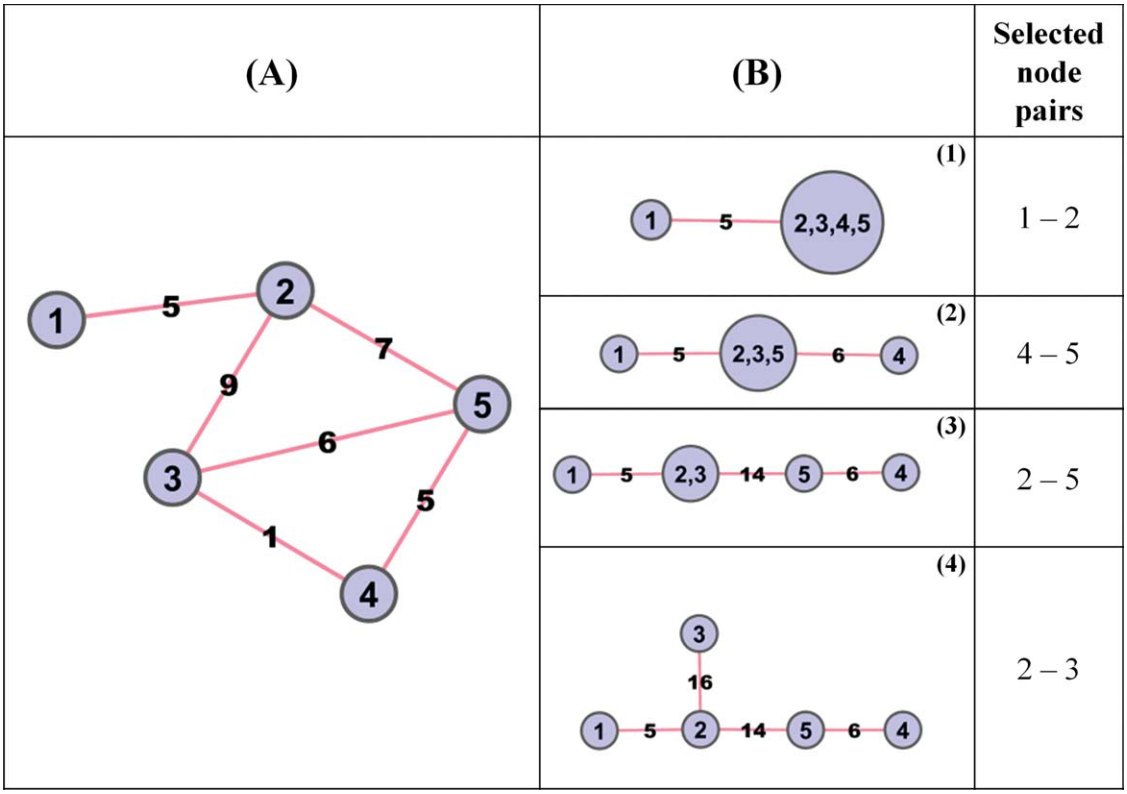

Figure 1

Illustration of construction steps of mincut tree with an example. Model network is shown in left panel (part **A**). (1) to (3) in part (**B**) are the intermediate steps to construct a mincut tree. The last column illustrates which node pairs are selected at each step. Here, nodes 1 and 2 are picked first, and a minimum 1–2 cut is found as 5. Then, from the remaining supernode, nodes 4 and 5 are selected and a minimum 4–5 cut is found as 6. (4) in part B is the resulting mincut tree of the network in part A. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

single node. Then, two nodes of N are picked to initialize the algorithm (here, nodes 1 and 2) and the minimum weight cut that separates 1 and 2 is found to split the node set N into subsets $S_1 = \{1\}$ and $S_2 = \{2, 3, 4, 5\}$, and has total weight 5. Next, another pair of nodes (nodes 4 and 5) is picked from S_2 and a minimum weight 4–5 cut is found, which has total weight 6. The algorithm continues until all subsets contain only one node. As a result, using $(n - 1)$ mincut calculations; the mincut tree is constructed (shown in part B4 in Fig. 1). We note that a mincut tree always exists for a connected graph but it does not need to be unique (because of the choice of nodes s,t in the s-t cut). However, we observed our procedure to be robust to different mincut trees in our computations.

Algorithm I: Determining the critical residue subtree

To understand the interconnections among hotspots, we identify a critical residue subtree in the mincut tree. In the initial step of the algorithm, the seed node, S, is chosen as the node with maximum total weight on edges incident on it in the tree. Then, we look at the neighbors of this residue and we take a neighbor if it has at least a degree of δ and at least a weighted degree of W_t. We set the degree threshold as $\delta = 3$ and the weight threshold as $W_t = (\sum_{1}^{n-1} W_e)/(n-1)$ in our computations, where W_e is the weight of the edges in the mincut tree and n is the number of nodes in the graph. The weighted degree of a node i is $dW(i) = \Sigma_{e \in E'} W_e$, where E' is the set of edges incident on i. Then, we check an adjacent node j in T if it exists. If $dW(j) \geq W_t$ and degree $(j) \geq \delta$, then the next node is j and it is added to the node list, L. Next, we go forward recursively by scanning the neighbors of the neighbors. At the end of the algorithm, if we cannot find any adjacent node passing the thresholds, then we output the node list, L, which corresponds to a subtree of the mincut tree. The simple steps of the algorithm are shown in **Algorithm 1.**

Algorithm 1 Critical subtree extraction algorithm.

Input: $G(N,E)$ ← a weighted undirected graph with weights w_e
Output: L, a list of nodes corresponding to a subtree in the mincut tree
$T \leftarrow$ mincut tree of G,
$\delta \leftarrow 3$, degree threshold
$W_i \leftarrow \left(\sum_{e=1}^{n-1} W_e\right)/(n-1)$, weight degree threshold
$S \leftarrow$ the node with maximum weighted degree
$L \leftarrow (S)$
$K \leftarrow (S)$
While $K \neq \emptyset$
  $i \leftarrow$ remove first node from K
  for all neighbors j of i
    if $dW(j) \geq W_t$ and degree $(j) \geq \delta$ and j is not in L
      append j to K
      append j to L
    end if
  end for
end while
Return L

Algorithm II: Iterative clustering of the interface residues

For clustering of the interface residues, we apply the iterative clustering algorithm in the work of Mitrofanova et al. to our problem.[32] Mitrofanova et al. use this algorithm to cluster unweighted network of PPI in yeast; in this way, they aim to identify protein complexes. In our work, our purpose is to generate residue clusters in protein interfaces, to see how residues are separated from each other along the iterations and to identify the relation between hot regions. For this purpose, we construct the bipartite residue graph of the interfaces. A bipartite graph is defined as a graph whose nodes can be divided into two disjoint sets such that every edge is between two nodes, one from each set. In the bipartite residue graph, $G(N,E)$ where $N = U \cup V$, U represents the set of contacting residues from one chain, whereas V represents the set of contacting residues from the other chain. So, only interchain contacts are represented in $G(N,E)$. We assign edge weights, w_e, as the absolute value of statistical residue contact potentials, as before. To find residue clusters, we follow an iterative procedure. After the construction of a mincut tree of the graph, we find the minimum value of the edge weights in the mincut tree (W_{min}) and remove the edges whose weight is equal to W_{min} from the mincut tree. Removal of an edge in the mincut tree corresponds to removal of a cut set from the residue contact network; in other words, separating the original network into at least two disjoint sets. At the ith iteration, the set of the connected components is represented as $G^i_{sub} = \{G^{i,1}, G^{i,2}, \ldots . G^{i,j}\}$ and the set of the subtrees is represented as $T^i_{sub} = \{T^{i,1}, T^{i,2}, \ldots . T^{i,j}\}$ where j is the number of the subnetworks and subtrees. Next, we find the minimum edge weight for each tree $W^i_{min} = \{W^{i,1}, W^{i,2}, \ldots . W^{i,j}\}$ and remove all minimum weight edges from the mincut trees in T^i_{sub}. Then, we reconstruct the contact network for all of the remaining subtrees and find mincut tree for each network. This iterative procedure continues until all subtrees have at least k connected residues where $k = 5$ in our computations. The simple steps of the algorithm are shown in **Algorithm 2.**

Algorithm 2 Clustering of the interface residues using a mincut tree.

Input: a weighted undirected bipartite graph, $G(N,E)$ with the edge weights w_e
Output: Node sets at the end of each iteration
Construct a mincut tree T of G
Find the minimum edge weight (W_{min}) and remove the edges whose weight is equal to W_{min} from T
While the size of each subgraph in $G^i_{sub} \geq k$ where $k = 5$
  $T^i_{sub} = \{T^{i,1}, T^{i,2}, \ldots . T^{i,j}\}$ is the set of subtrees after removal of the edges
  $G^i_{sub} = \{G^{i,1}, G^{i,2}, \ldots . G^{i,j}\}$ is the set of subnetworks after removal of the cut edges
  Construct the mincut tree of the subgraphs in G^i_{sub}; $T^{i+1}_{sub} = \{T^{i+1,1}, T^{i+1,2}, \ldots . T^{i+1,j}\}$
  Find minimum edge weight of each tree in T^{i+1}_{sub}; $W^{i+1}_{min} = \{W^{i+1,1}, W^{i+1,2}, \ldots . W^{i+1,j}\}$
  Remove all minimum weight edges from the mincut trees in T^{i+1}_{sub}
  $i = i + 1$
end while
return the set of subgraphs, G^{i-1}_{sub}

RESULTS

Significant interactions between residues are represented by highly weighted edges using pairwise contact potentials. A mincut tree represents the weakest connections with minimum absolute contact energy in a compact structure. Thus, the complex structure of the contact network of the residues containing n nodes and m edges is simplified and summarized by a tree with $n - 1$ edges. Hence, residue contacts and closely related parts of the network can be interpreted and visualized easily.

As an example, the Erythropoietin (EPO) receptor and EPO mimetic peptide complex is analyzed. EPO is a hormone participating in the regulation of proliferation and differentiation of immature erythroid cells. EPO mimetic peptide (EMP1) functions as a mimetic of EPO. There is a competition between EMP1 (pdbID:1ebp, chainC) and EPO to bind the EPOR (pdbID:1ebp, chainA). Despite the unrelated sequences of EMP1 and EPO, both can bind to EPOR and stimulate biological activity.[33] Their interface region (1ebpAC) is shown in Figure 2 both in structural representation (part A) and in graph representation (part B) (the edge weights are not shown in the

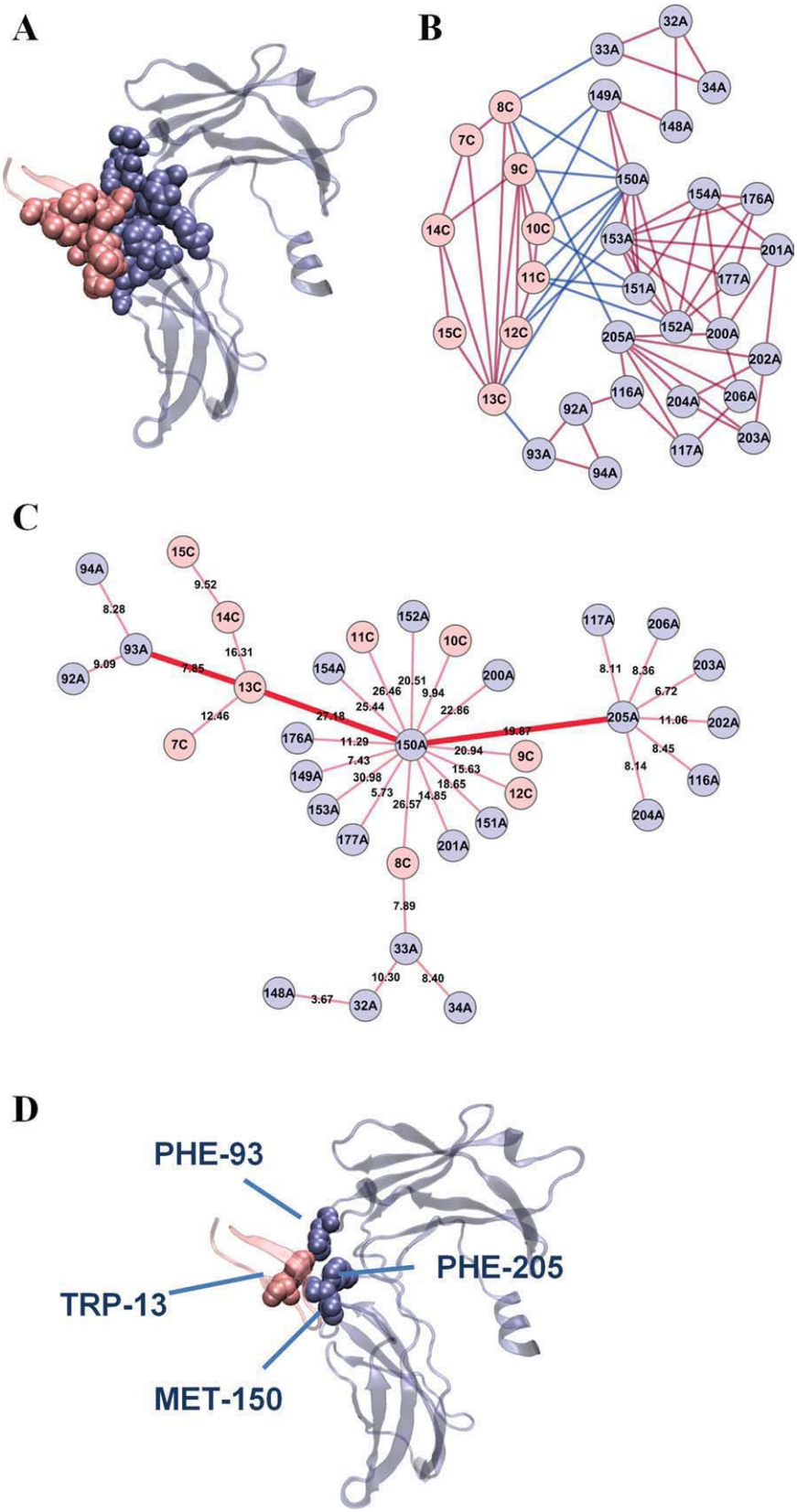

Figure 2

(**A**) An example of protein interface (1ebpAC) and (**B**) its residue interaction network. (**C**) Constructed mincut tree of the 1ebpAC interface. Nodes are colored according to the chains. (**D**) Mapping of the generated subtree on the 3D structure of 1ebpAC interface. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

figure.) The nodes in part B are colored according to the chains, and also the edges are colored according to the type of the contact (interchain or intrachain). Blue colored edges are the interchain contacts and the red edges are intrachain contacts. The mincut tree constructed for this residue network is illustrated in Figure 2 (part C). When we compare the network in part B and the tree in part C visually, we observe two advantages of the latter; (i) it is much easier to visually inspect the organization of the residues in the interface, (ii) it is more informative. When we check the most connected node in the mincut tree, we notice that this node corresponds to an experimental hotspot. Other important residues are consecutive in a path in the mincut tree. Experimental data in BID[5] indicate that residues 93A, 150A, 205A, and 13C are hot residues in the binding of EPO receptor and EPO mimetic peptide. In mincut tree of 1ebpAC interface, the most connected node is 150A. When we concentrate on the details of this analysis further, we observe that other experimentally determined key residues are sequenced in a subtree in this mincut tree. The bold lines in Figure 2 show the subtree where experimental hotspots are located. The hot residues (205A, 150A, 93A, 13C) form a path in the resulting mincut tree. This is an indication of the communication between the hotspots and also shows how information in one chain can be transmitted to another chain. Among the hotspots, 150A forms interaction with six residues in chain C and four residues in chain A. So, it is expected that this residue should be critical in binding. However, each of 205A and 93A interacts with a single residue; 13C interacts with two residues. The rest of their interactions are formed within the chain they are involved in. Therefore, it is not clear that these residues are hotspots just by looking at the interaction network shown in Figure 2(B). Algorithm 1 successfully points out these residues by locating them along the same subtree of the mincut tree.

Analyzing mincut trees for other protein complexes

We further construct a mincut tree for 38 complexes for which experimental hotspot information is available. We noticed a consistent trend that the most connected node in the mincut tree usually corresponds to an experimental hotspot. With this method, besides the visual compactness, critical residues can be identified as well. In Table I, the most connected node in the mincut tree is given for 38 complexes. In this table, the most connected residue corresponds to an experimental hotspot in 20 out of 38 protein complexes. For the remaining complexes, these residues are either a close neighbor of an experimental hotspot (in two proteins) or they are computational hotspots predicted by other methods, such as Hotpoint,[12] KFC[8] (in 13 proteins). The largest weighted degree node in the residue contact network is the residue with largest energetic contribution. In 27 out of 38 cases, the residue with largest energy contribution corresponds also to the most connected node in the mincut tree. In the remaining 11 cases, the largest weighted degree node

Table I
The Most Connected Node in the Mincut Tree for Several Complexes

Protein complex	Interface name	The most connected node in the mincut tree
Ribonuclease inhibitor–angiogenin complex	1a4yAB	**318A–TRP**
Growth hormone/receptor complex	1a22AB	179A–ILE[a]
Immunoglobulin heavy chain–tissue factor complex	1ahwBC	**156C–TYR**
Barnase–Barstar Complex	1brsAD	**39D–ASP**
E9 DNase–Im9 Complex	1bxiAB	**33A–LEU**
Chymotrypsin–BPTI Complex	1cbwCD	**15D–LYS**
Soluable tissue factor complex	1danTU	19T–PHE[a]
Immunoglobin heavy chain–peptide complex	1dn2AE	252A–MET[a]
Fv-Fv idiotope-anti-idiotope complex	1dvfBD	**98D–TYR**
Cell division protein ZipA/ FtsZ fragment complex	1f47AB	85B–PHE[a]
Immunoglobulin FC/Fragment B of protein A complex	1fc2CD	136C–LEU[a]
GP120/CD4 complex	1gc1GC	28C–TRP
Interferon gamma receptor/fab fragment complex	1jrhLI	**92L–TRP**
TEM1-β-lactamase-inhibitor complex	1jtgAB	**142B–PHE**
IGG1-kappa D1.3 FV complex	1vfbAB	36A–TYR[a]
Beta trypsin/inhibitor complex	2ptcEI	14I–CYS[a]
HyHEL-10 Fab heavy chain-lysozyme complex	3hfmHY	**33H–TYR**
HyHEL-10 Fab light chain-lysozyme complex	3hfmLY	**20Y–TYR**
Human Growth Factor–Receptor Complex	3hhrAB	**182A–CYS**
Calmodulin–Protein Kinase Complex	1cdlAE	**810E–ILE**
Numb Protein Complex	1ddmAB	199A–LEU
Ribonuclease inhibitor–ribonuclease A complex	1dfjEI	259I–TRP[a]
DES-GLA factor VIIA–peptide complex	1dvaHX	**34H–LEU**
Integrin–collagen complex	1dziAC	220A–LEU
EPO Receptor–EPO Mimetic Peptide Complex	1ebpAC	**150A–MET**
Bone morphogenetic protein-2/ receptor 1A complex	1es7AD	785D–PHE[a]
Blood coagulation factor VIIA/ soluable tissue factor complex	1fakLT	70L–CYS[a]
IGG1-Protein G complex	1fccAC	**27C–GLU**
Mms2/Ubc13 heterodimer	1jatAB	**8B–PHE**
HslUV protease/chaperone complex	1g3iAG	**443A–ILE**
Nidogen-1 with IG3 complex	1gl4AB	**429A–HIS**
Beta catenin/APC complex	1jppBD	424B–LEU
Phagocyte NADPH Oxidase complex	1k4uSP	505S—ILE[a]
alphaL I domain in complex with ICAM-1	1mq8AB	204B–LEU
IkappaBalpha/NF-kappaB complex	1nfiBF	254B–VAL[a]
MazE/MazF Complex.	1ub4AC	**458C–LEU**
Numb PTB domain-peptide complex	2nmbAB	199A–LEU[a]
p53 oligomerization domain complex	3sakAC	**23A–PHE**

Bold residues are experimental hotspots.
[a]Identified as hotspot by other prediction methods, such as Hotpoint[12] and KFC[8].

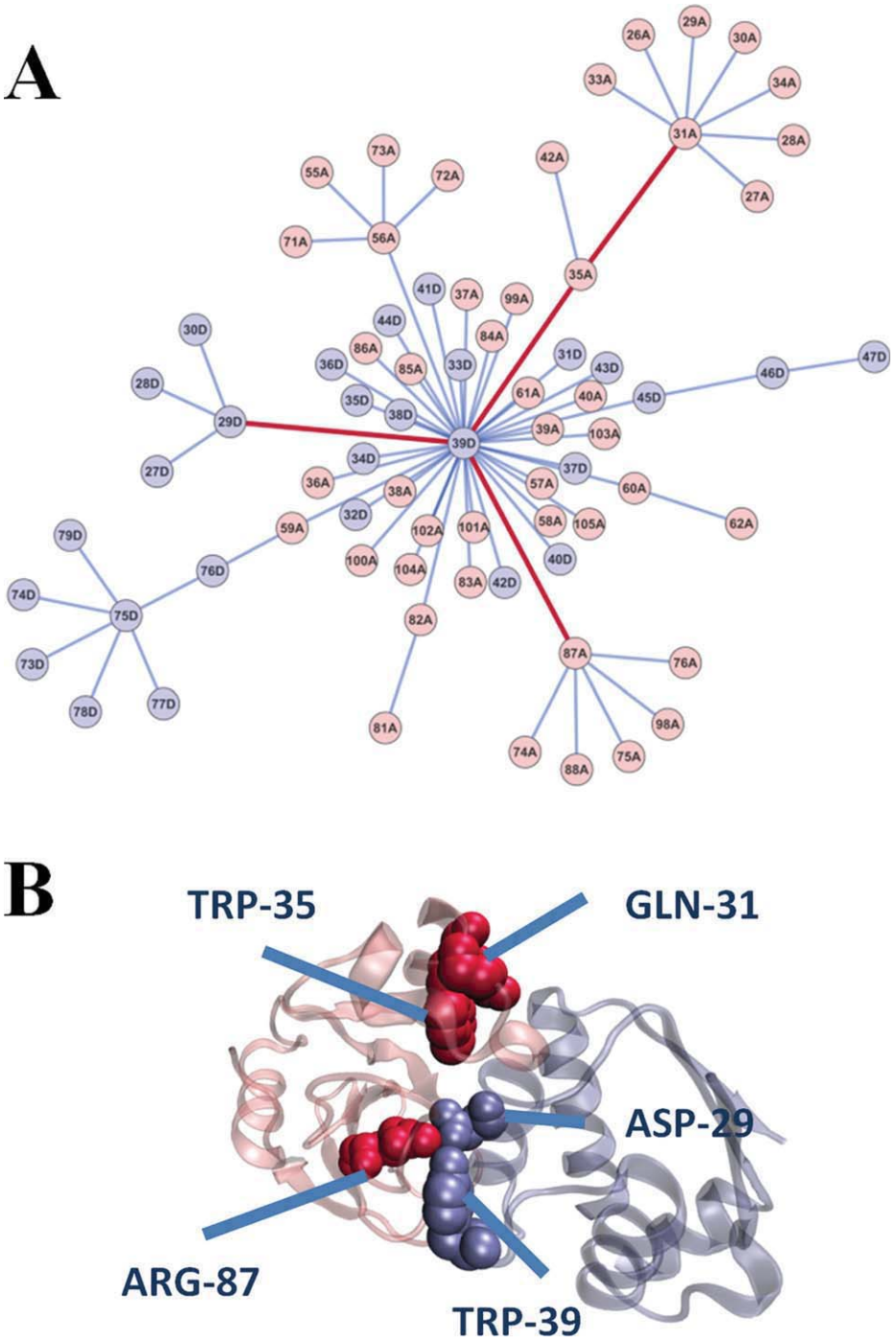

Figure 3

(**A**) Mincut tree for the 1brsAD interface. Nodes are colored according to the chains. The red colored bold edges represent the subtree of critical residues. The critical residues in this subtree are "*29D-39D-87A-31A-35A*". (**B**) Spatial illustration of this subtree on the protein complex. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

in the residue contact network does not correspond to the most connected residue in the mincut tree. In 6 out of 11 complexes, the mincut tree approach finds the hotspots; but in comparison, the largest weighted degree node in the residue contact network is a hotspot in only one complex. The advantage of the proposed approach is that besides the most connected node, a residue subtree is generated and several of the residues in this subtree either correspond to other hotspots or they are closely related to the residue with maximal flow, which may act

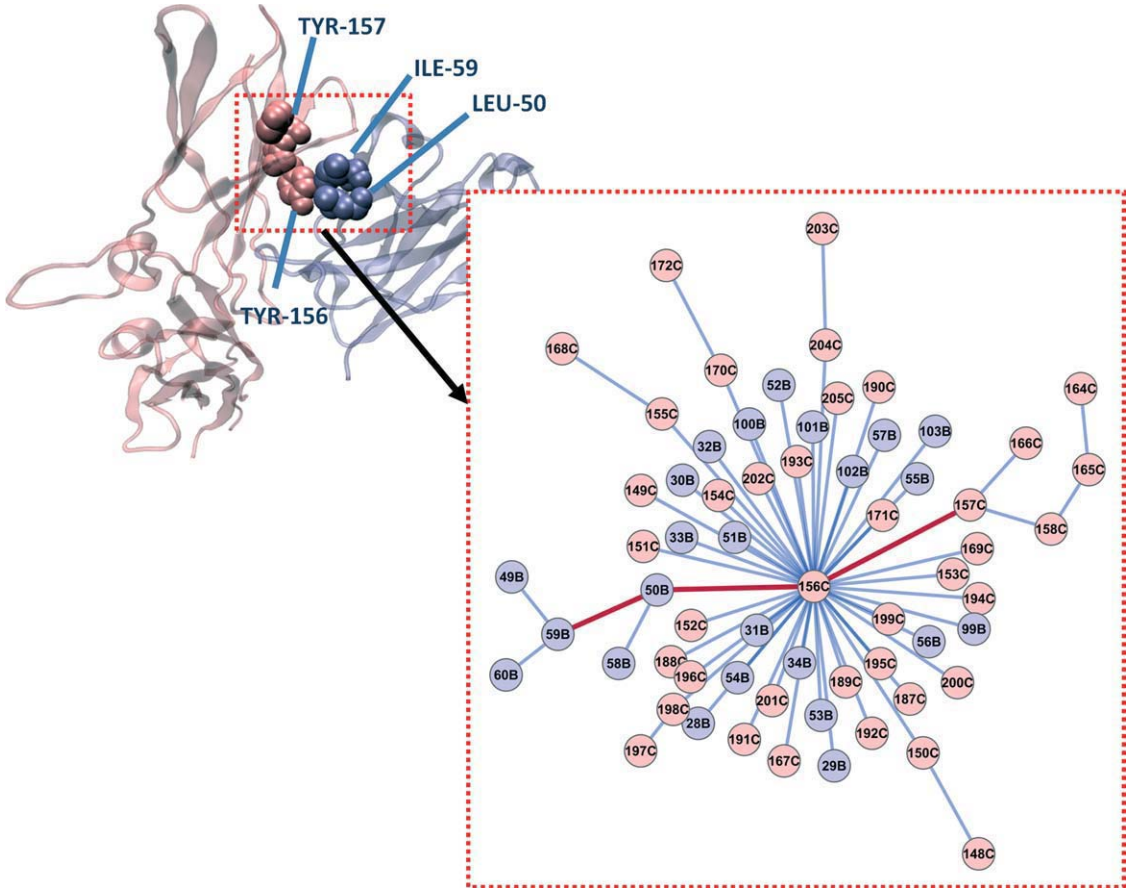

Figure 4

Mincut tree for the 1ahwBC interface and illustration of the extracted subtree on the 3D structure. Nodes are colored according to the chains. The red colored bold edges represent the subtree. This subtree forms a continuous residue path in 3D. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

cooperatively and form a continuous path in 3D. In a classical network representation, this information is hidden and the mincut tree approach brings it out. Furthermore, the original network is too crowded to visualize, whereas the mincut tree representation provides essential information in a simpler format.

One of the complexes in Table I is the barnase-barstar complex (pdb ID: 1brs). Barnase (chain A) is a ribonuclease enzyme. Barstar (chain D) inhibits barnase by blocking its active site. In this way, barstar stops barnase to damage the synthesized RNA.[34] A mincut tree is constructed for the barnase-barstar complex [see Fig. 3(A)] and the most connected node, 39D, found to be an experimental hotspot. Further, the generated subtree using Algorithm 1 consists of the nodes 29D-39D-31A-35A-87A. Within this subtree, all three residues in the 29D-39D-87A path are experimental hotspots; they are located closely in the 3D structure and form a region [shown in Fig. 3(B)]. The mincut tree brings these three residues together. When we go from the complicated overall residue contact network to the mincut tree, we can examine and interpret the organization of the hotspots. To justify the connection between these five residues, we take the single point and double point mutation information available for the barnase-barstar complex.[35] The observed changes in binding free energy are as follows, $\Delta\Delta G_{29D}$ = 3.4 kcal/mol, $\Delta\Delta G_{87A}$ = 5.5 kcal/mol, $\Delta\Delta G_{39D}$ = 7.7 kcal/mol, $\Delta\Delta G_{87A/39D}$ = 6.1 kcal/mol, $\Delta\Delta G_{87A/29D}$ = 8.0 kcal/mol. According to these energy values, the residues 87A and 39D are cooperative with each other, which cause an energy difference of 7.1 kcal/mol (difference between 5.5 + 7.7 kcal/mol and 6.1 kcal/mol); on the other hand, simultaneous mutation of 87A and 29D causes a difference of 0.9 kcal/mol. As mutation data related to the residues 31 and 35 in chain A are not known experimentally, we cannot comment about the relation between the distant residues 31A and 35A with the 29D, 39D, and 87A. However, from literature, we found that the residues 31A and 35A have significant effect on folding of barnase.[36,37] The effect of 31A is 1.1 kcal/mol.[37] The

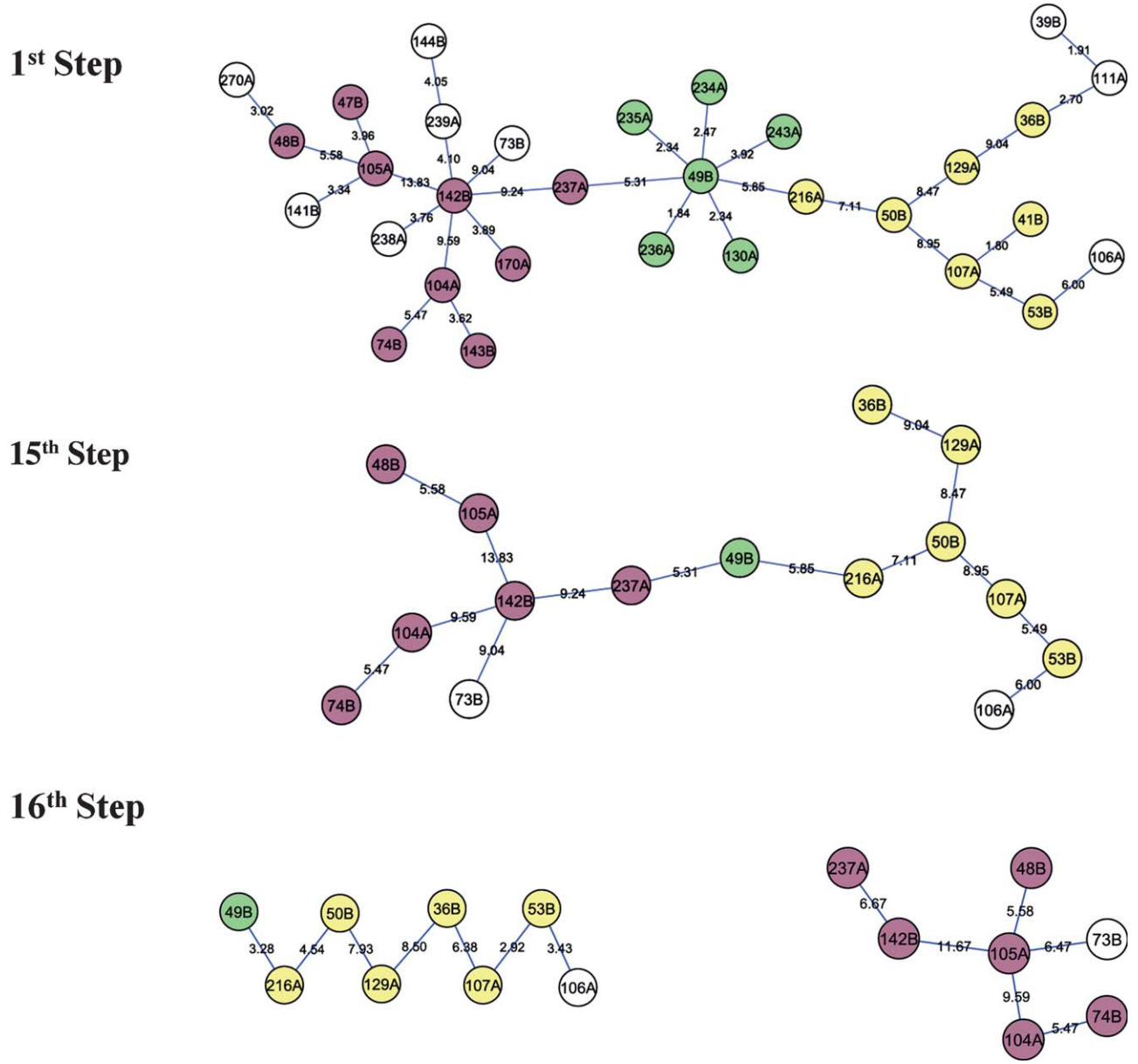

Figure 5

Mincut tree of the bipartite graph of the TEM1-BLIP complex at first, 15th and 16th steps of the iteration. Each color represents a cluster and the coloring scheme is the same as in Schreiber's work.[20]

mutation on 35A decreases more than 70% of the fluorescence intensity of barnase.[36] So, their mutations may also have effect on 29D, 39D, and 87A.

Another example is the immunoglobulin complex with tissue factor (pdbID:1ahw, between chain B and chain C).[38] The most connected node in the mincut tree is 156C, which is also an experimental hotspot defined in Alanine Scanning Database (ASEdb).[4] The subtree identified in the mincut tree is composed of the residues; 59B-50B-156C-157C [see Fig. 4]. The hotspot (156C) is surrounded by other residues in this path as seen in the 3D picture of the immunoglobulin-tissue factor complex and they form a region in the binding site. These residues play possibly an important role in binding (For further examples, see Table I).

Organization of hot regions

As discussed before, hotspots are clustered into hot regions,[19,39,40] and understanding the organization of hotspots and hot regions in protein binding sites is a major task for protein interaction prediction, as well as

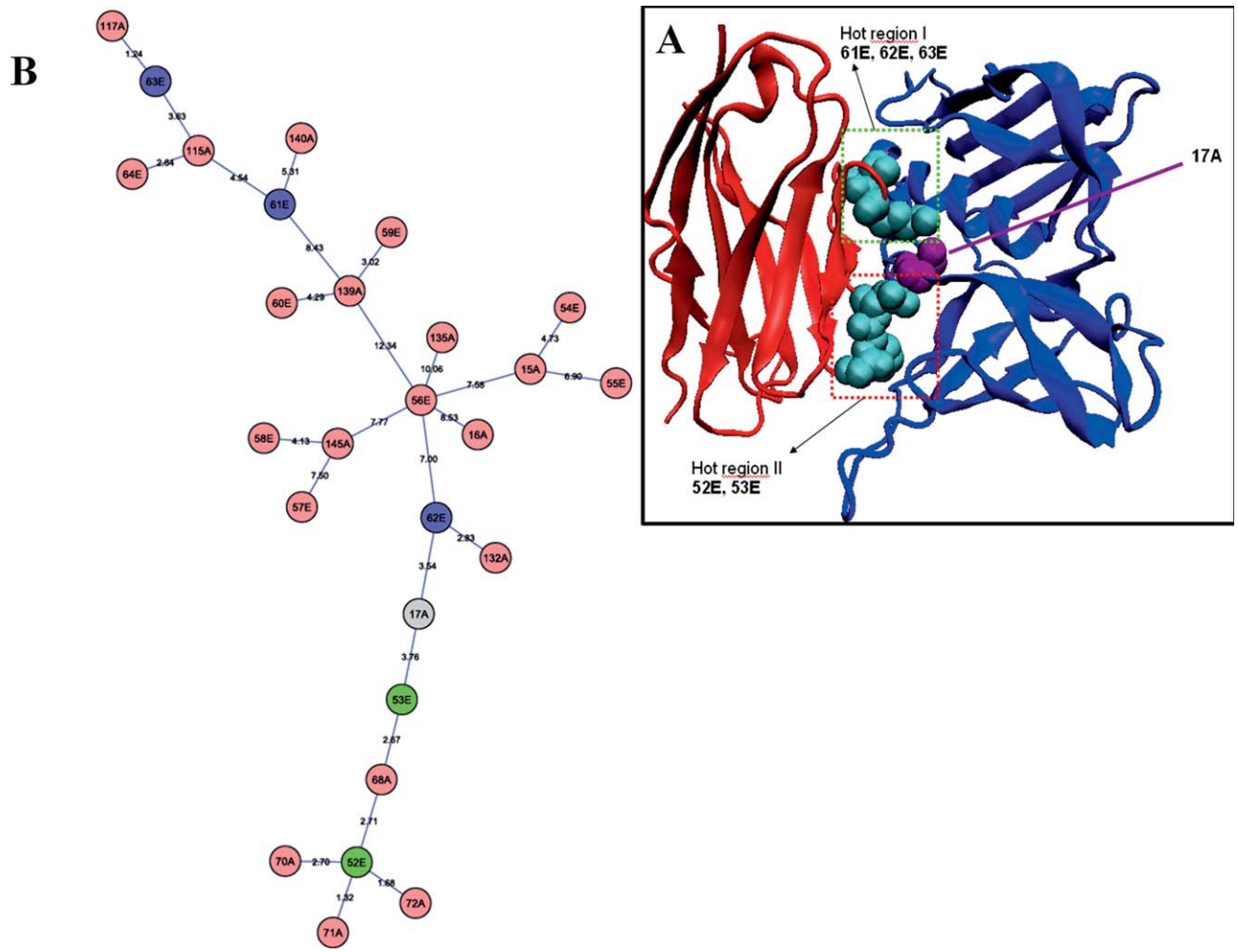

Figure 6

(**A**) Three dimensional structure of the hvb2.1-TSST1 complex. (**B**) Mincut tree of the bipartite graph of the hvb2.1-TSST1 complex.

the design of therapeutic agents. We have two contradicting cases available from the literature. The first one is the BLIP complex, which is analyzed by Schreiber and co-workers.[20] They stated that distinct residue clusters are energetically additive, but the residues within the same cluster are highly cooperative. The other one is the TSST1-hv-b2.1 complex whose distinct hot regions are energetically cooperative.[21] To analyze hot regions, we construct bipartite graphs of the residues. We apply the iterative clustering algorithm (Algorithm 2) to both examples, and analyze the correlation between the results. Here, our aim is to cluster the interface residues and to see which residues are important to bring clusters together along the iterations.

The clustering of the interface residues of TEM1-BLIP complex (pdb ID:1jtg, chain A and B, respectively), performed by Schreiber and co-workers, divides the interface region into five clusters (namely C1, C2, C3, C4, and C5). Using multiple mutagenesis analysis of two clusters (C1 and C2), they stated that these two clusters are energetically independent of each other, but the intracluster connections are cooperative. When we construct the bipartite graph of the residue network, we see four independent subgraphs. Two of the subgraphs correspond to C3 and C5; they are not connected to the largest connected component (see Fig. S1). For mincut tree analysis, we focus on the largest connected component. The largest connected component in the graph contains the residues in C1, C2, and C4. Using the mincut tree approach, we cluster the nodes and check the robustness of the clusters to deletion of edges in the minimum cut in the residue network iteratively as described in Algorithm 2. Here, we notice that 130A, 234A, 235A, and 243A are removed at the initial iteration steps (see Fig. 5), which corresponds to C1. All residues in C1, except 49B, are removed in the iterations, which imply that the minimum cuts connecting these residues to the network are weak within this cluster. When we continue the iterations, we end up with two clusters for

the largest connected component which correspond to C2 and C4, respectively. Further, we notice that two distinct clusters are connected to each other using residue 49B. The contact between 216A-49B-237A is robust to several edge cuts until the 16th iteration (see Fig. 5) and this part is almost the strongest part of the mincut tree. We hypothesize that the information flow from one cluster to another passes through 49B. The importance of this residue is not obvious from the original residue interaction network. However, mincut tree shows the critical role of 49B by showing that it connects two individual residue clusters. This result is in correlation with the experimental mutation data presented by Schreiber and co-workers. The change in binding free energy upon single point mutation of 49B is 7.5 kJ/mol. On the other hand, the change is 7.1 kJ/mol upon simultaneous mutation of 49B, 130A, 235A, 243A, and 234A. Both mutations have almost the same effect on binding. Further, the effect of single point mutations of 130A, 243A, and 234A (1.4, 5.3, and 4.3 kJ/mol, respectively) are not as large as the effect of 49B. Probably, the effect of 49B is dominant in simultaneous mutations and 49B is the most critical residue in C1. This residue also connects two other clusters and furthermore its connection with these clusters is robust to the edge removal. Thus, the mincut tree (shown in Fig. 5) clearly suggests a link between C2 and C4 through the residue 49B. Here, we state that the mutation of 49B may lead to structural rearrangements in C2 and C4. When 49B is mutated to alanine, the connection between C2 and C4 might be broken, according to the mincut tree. Thus, we suggest for an experimental analysis of the clusters C2 and C4 using multiple mutations. Analysis of the cooperativity between these two distinct clusters may be investigated further.

To check the cooperativity between distinct hot regions, Moza et al.[21] analyzed the interaction between hvb2.1 and TSST1. In the hvb2.1-TSST1 complex (pdb ID: 2ij0), there are two distinct hot regions on the hvb2.1 (chain E) surface, which are 52E and 53E in CDR2 loop and 61E and 62E in FR3. They stated that although these two regions are distant to each other by more than 20Å, they are highly cooperative. When we apply Algorithm 2 to the hvb2.1-TSST1 complex, we observe that these two hot regions are linked by the residue 17A on the surface of TSST1 (chain A) and this linkage is easily distinguished using the mincut tree (see Fig. 6). Although these two hot regions are spatially far away from each other, they are located in the loop regions, at the flexible parts of the hvb2.1 and they are connected via residue 17A on the partner protein TSST1. So, their cooperativity is expected when we analyze the mincut tree. Another observation is that when we apply iterative clustering algorithm, we obtain only one cluster and along the iterations residues are separated one-by-one from the main mincut tree. This result shows a strong connection between two hot regions in the TSST1-hvb2.1 complex.

CONCLUSIONS

Proteins interact through their binding sites. Several graph based algorithms are used to characterize and analyze PPI. In this work, we use mincut trees to visualize and analyze residue contact networks compactly. Edges in the contact network are weighted according to an energy function, namely knowledge-based potentials. Mincut tree representation highlights some central residues at first glance, which cannot be distinguished in classical network representation visually. This information provides us the most important node and the critical paths within the interface region. As the most connected residue in the tree usually corresponds to an experimental hotspot, and hotspots are sequenced along paths on the tree, we give an algorithm to find a subtree containing hotspot paths.

We also analyzed the dependency of the distinct residue clusters using some known protein complexes, such as TEM1-BLIP and hvb2.1-TSST1. We found some critical traces to explain the cooperativity between distinct residue clusters using a clustering algorithm that runs on a mincut tree. As a future direction, one may also analyze how hotspots are communicating with each other and how information in one chain is transmitted to another chain using the proposed algorithms.

Briefly, our new approach is useful for basic biological cases at the molecular level. A mincut tree simplifies the complex residue network between two interacting proteins visually. Using this tree, residue clusters and the critical residues in binding of two proteins can be identified computationally by efficient algorithms.

ACKNOWLEDGMENTS

This project has been supported by TUBITAK (Research Grant No 109T343 and 109E207). N.T. has been supported by TUBITAK fellowship.

REFERENCES

1. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
2. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science 1995;267:383–386.
3. Wells JA. Systematic mutational analyses of protein-protein interfaces. Meth Enzymol 1991;202:390–411.
4. Thorn KS, Bogan AA. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 2001;17:284–285.
5. 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. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics 2003;19:1453–1454.
6. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369–387.

7. Kortemme T, Kim DE, Baker D. Computational alanine scanning of protein-protein interfaces. Sci STKE 2004;2004:l2.
8. Darnell SJ, Page D, Mitchell JC. An automated decision-tree approach to predicting protein interaction hot spots. Proteins 2007;68:813–823.
9. Guney E, Tuncbag N, Keskin O, Gursoy A. HotSprint: database of computational hot spots in protein interfaces. Nucleic Acids Res 2008;36(Database issue):D662–D666.
10. Lise S, Archambeau C, Pontil M, Jones DT. Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. BMC Bioinformatics 2009;10:365.
11. Ofran Y, Rost B. Protein-protein interaction hotspots carved into sequences. PLoS Comput Biol 2007;3:1169–1176.
12. Tuncbag N, Gursoy A, Keskin O. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. Bioinformatics 2009;25:1513–1520.
13. Gonzalez-Ruiz D, Gohlke H. Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding. Curr Med Chem 2006;13:2607–2625.
14. Huo S, Massova I, Kollman PA. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. J Comput Chem 2002;23:15–27.
15. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. Proc Natl Acad Sci USA 2004;101:11287–11292.
16. Brinda KV, Kannan N, Vishveshwara S. Analysis of homodimeric protein interfaces by graph-spectral methods. Protein Eng 2002;15:265–277.
17. del Sol A, O'Meara P. Small-world network approach to identify key residues in protein-protein interaction. Proteins 2005;58:672–682.
18. Haliloglu T, Seyrek E, Erman B. Prediction of binding sites in receptor-ligand complexes with the Gaussian Network Model. Phys Rev Lett 2008;100:228102.
19. Keskin O, Ma B, Nussinov R. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol 2005;345:1281–1294.
20. Reichmann D, Rahat O, Albeck S, Meged R, Dym O, Schreiber G. The modular architecture of protein-protein binding interfaces. Proc Natl Acad Sci USA 2005;102:57–62.
21. Moza B, Buonpane RA, Zhu P, Herfst CA, Rahman AK, McCormick JK, Kranz DM, Sundberg EJ. Long-range cooperative binding effects in a T cell receptor variable domain. Proc Natl Acad Sci USA 2006;103:9867–9872.
22. Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL. Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. Protein Sci 1998;7:2578–2586.
23. Anashkina A, Kuznetsov E, Esipova N, Tumanyan V. Comprehensive statistical analysis of residues interaction specificity at protein-protein interfaces. Proteins 2007;67:1060–1077.
24. Bahar I, Jernigan RL. Coordination geometry of nonbonded residues in globular proteins. Fold Des 1996;1:357–370.
25. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. Principles of protein folding—a perspective from simple exact models. Protein Sci 1995;4:561–602.
26. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
27. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasichemical approximation. Macromolecules 1985;18:534–552.
28. West DB. Introduction to Graph Theory. USA: Prentice Hall; 1996.
29. Gomory RE, Hu TC. Multi-terminal network flows. J Soc Indust Appl Math 1961;9:551–570.
30. Flake GW, Tarjan RE, Tsioutsiouliklis K. Graph clustering and minimum cut trees. Internet Math 2004;1:385–408.
31. Goldberg AV, Tsioutsiouliklis . Cut tree algorithms: an experimental study. J Algorithm 2001;38:51–83.
32. Mitrofanova A, Farach-Colton M, Mishra B. Efficient and robust prediction algorithms for protein complexes using Gomory-Hu trees. Pac Symp Biocomput 2009;14:215–226.
33. Livnah O, Stura EA, Johnson DL, Middleton SA, Mulcahy LS, Wrighton NC, Dower WJ, Jolliffe LK, Wilson IA. Functional mimicry of a protein hormone by a peptide agonist: the EPO receptor complex at 2.8 A. Science 1996;273:464–471.
34. Buckle AM, Schreiber G, Fersht AR. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. Biochemistry 1994;33:8878–8889.
35. Schreiber G, Fersht AR. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. J Mol Biol 1995;248:478–486.
36. Loewenthal R, Sancho J, Fersht AR. Fluorescence spectrum of barnase: contributions of three tryptophan residues and a histidine-related pH dependence. Biochemistry 1991;30:6775–6779.
37. Serrano L, Sancho J, Hirshberg M, Fersht AR. Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. J Mol Biol 1992;227:544–559.
38. Huang M, Syed R, Stura EA, Stone MJ, Stefanko RS, Ruf W, Edgington TS, Wilson IA. The mechanism of an inhibitory antibody on TF-initiated blood coagulation revealed by the crystal structures of human tissue factor. Fab 5G9 and TFG9 complex J Mol Biol 1998;275:873–894.
39. Keskin O, Gursoy A, Ma B, Nussinov R. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? Chem Rev 2008;108:1225–1244.
40. Keskin O, Nussinov R. Similar binding sites and different partners: implications to shared proteins in cellular pathways. Structure 2007;15:341–354.