The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009
Copyright © 2009 ORSC & APORC, pp. 296–303

Align Protein Surface Structures to Identify Evolutionally and Structurally Conserved Residues*

Lin Wang[1,†] Ji-Guang Wang[2] Luonan Chen[3]

[1]Computer Science and Information Engineering College, Tianjin University of Science and Technology, Tianjin 300222, China
[2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[3]Department of Electronics, Information and Communication Engineering, Osaka Sangyo University, Osaka 574-8530, Japan

Abstract Protein-protein interface underlies the protein protein interaction. Alanine mutation of protein-protein interface residues has shown that the distribution of binding free energy is not average among the interface residues. Actually, there are hot spots in the protein interfaces that contribute most binding energy. Here we provide a new method based on integer quadratic programming that systematically aligns protein surface structures shared by a set of proteins. This method incorporates protein sequence and structure data, and can correctly identify residues having evolutional and structural conservation between different proteins. It is sequence order independent, so can unravel the evolutional similarity between distant proteins. Furthermore, it can be used to predict hot spots with ROC area AUC=0.6. Compared with most hot spot prediction methods, our method does not need prior knowledge for the structure of protein complex or even the structure of the binding partner.

Keywords Hot spots; Protein surface; Residue conservation; Integer quadratic programming

1 Introduction

Protein-protein interaction plays a key role in the signal transduction network and metabolic network, and protein-protein interface is the region of interaction between two non-covalently linked protein molecules. Protein-protein interfaces ($\sim$ 1,500-3,000Å^2) are much larger than protein-ligand interfaces ($\sim$ 300-1,000Å^2). In addition, the protein-protein interfaces are generally flat and often lack the grooves and pockets presenting at the protein-ligand interfaces [1]. Experiments show that the binding energies among the residues in the protein-protein interfaces are not evenly distributed, and only a small fraction of interface residues named hot spots are responsible for the binding [2]. Alanine scanning mutagenesis is a main experimental technique for recognition of hot spots. It

*The authors are separately supported by the NSFC grants 10631070, 60873205, 10801131, the grant kjcx-yw-s7 from CAS, and 2006CB503905 from MST of China.

†Email: linwang@amss.ac.cn.

has been shown that structurally conserved residues at protein-protein interfaces correlate with the experimental alanine-scanning hot spots and hot spots are tightly packed to be 'hot regions' [3]. In addition, interface residues are evolutionally conserved compared with other surface residues [4]. Based on these features, we provide a method PSAlign that systematically aligns the surfaces of two proteins by exploiting protein sequence and structure to identify the residues having structural and evolutional similarities. Specifically due to the local surface structure is more related to protein function but the large size of protein surface, we first partition the protein surfaces into several local structures, then use residue network alignment method that is based on an integer quadratic programming to align these protein local structures. When it is used to align the surfaces in a protein set, we find structurally and evolutionally conserved residues can distinguish hot spots from other surface residues. As a major advantage, our prediction method for hot spots does not use the structure of protein complex or even the structure of the binding partner which is in contrast with most hot spot prediction methods nowadays.

2 Method for aligning protein surfaces

Given two protein structures, we use protein surface alignment method PSAlign to align their protein surfaces by incorporating sequence and structure data. The main stages of PSAlign are presented in Figure 1 and are detailed in the following.

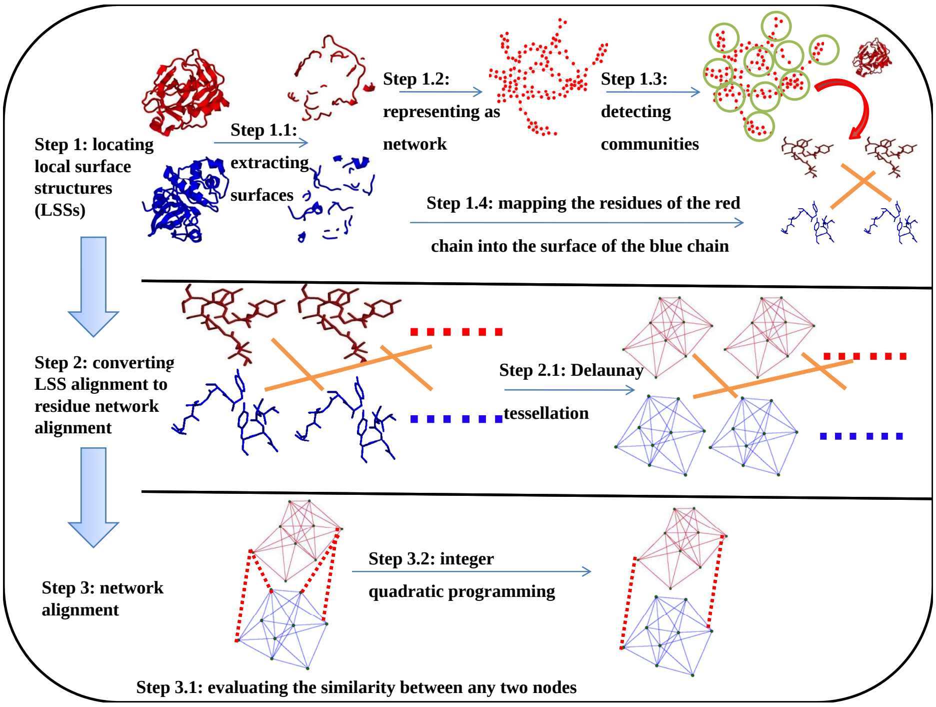

Step 1: locating local surface structures (LSSs)
Step 1.1: extracting surfaces
Step 1.2: representing as network
Step 1.3: detecting communities
Step 1.4: mapping the residues of the red chain into the surface of the blue chain
Step 2: converting LSS alignment to residue network alignment
Step 2.1: Delaunay tessellation
Step 3: network alignment
Step 3.2: integer quadratic programming
Step 3.1: evaluating the similarity between any two nodes

Figure 1: Overview of the method PSAlign for aligning two protein surfaces.

2.1 Locating local surface structures of two proteins

Owing to the local surface structure is more related to protein function but the large size of protein surface, we locate the local surface structures of two proteins 1 and 2 through the following four steps.

Firstly we extract surfaces of two proteins 1 and 2. For two proteins 1 and 2, a protein residue is defined as a surface residue if its solvent accessible surface area is at least 16% of the nominal maximum area of this amino acid type [5].

Secondly we represent the surface of protein 1 as a network. Specifically we construct a surface network of protein 1 by connecting any two surface residues if their distance is not more than 7Å.

Thirdly we identify communities of the surface network of protein 1 and locate local surface structures of protein 1. In practical we use Markov Cluster Algorithm (MCL [6], the parameter I=1.4) to partition the surface network of protein 1 into several communities. Then each community is projected into a local surface structure so that protein 1 is located as several local surface structures.

Finally we locate local surface structures of protein 2. We perform sequence alignment of protein chains 1 and 2 using ClustalW [7], then project the residues from different local surface structures of protein 1 into the surface of protein 2 so that the surface of protein 2 can be accordingly partitioned into several local surface structures.

2.2 Converting the local surface structure alignments into residue network alignments

At the next stage we convert each local surface structure alignment into a residue network alignment. Here we use the coordinates of CA atoms of each local surface structure to construct a corresponding Delaunay tessellation network. Delaunay tessellation network, i.e. tetrahedron mesh partition of a point set, requires each tetrahedron's circumsphere does not contain any other points. It is a basic problem in computational geometry, and is often used to describe protein 3D structure [8].

2.3 Aligning the residue networks through integer quadratic programming

Here we construct an integer quadratic programming model to align two residue networks with the following two steps.

Firstly we evaluate the similarity of two nodes from two different residue networks. For two proteins containing the two local structures, based on PSI-BLAST we use their protein sequences respectively to search the non-redundant protein sequence databases for 3 times to obtain two position-specific scoring matrices (PSSMs). So each residue from each local structure corresponds to a scoring profile and a frequency profile of 20 amino acids. For two nodes a and b from two different networks, i.e. two residues from two local structures, their scoring profiles are denoted as s_a and s_b, and frequency profiles are denoted as f_a and f_b, then we define the similarity score of two nodes as: $w_{ab} = \sum_{i=1}^{20} s_a(i) * f_b(i) + \sum_{i=1}^{20} s_b(i) * f_a(i)$.

Then we apply an integer quadratic programming for residue network alignment. For two local structures (1) and (2) from two proteins, and their corresponding residue networks, we exploit sequence evolutionary information and structural information, and design the following integer quadratic programming to align their residue networks.

$$\begin{aligned}
max \quad & \lambda \textstyle\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} w_{ij}x_{ij} - (1-\lambda)\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{k=1}^{N_1} \sum_{l=1}^{N_2} \lvert d_{ik} - d_{jl} \rvert^2 x_{ij}x_{kl} && (1)\\
s.t. \quad & \textstyle\sum_{i=1}^{N_1} x_{ij} \leq 1 \quad for \quad j = 1,...,N_2 && (2)\\
& \textstyle\sum_{j=1}^{N_2} x_{ij} \leq 1 \quad for \quad i = 1,...,N_1 && (3)\\
& x_{ij} \in \{0,1\} && (4)
\end{aligned}$$

where N_1 and N_2 are respectively the number of residues in two local structures (1) and (2); w_{ij} is the similarity score for residue i in local structure (1) and residue j in local structure (2), which is based on evolutionary information; d_{ik} is the Euclidean distance between the residues i and k in local structure (1) if there is an edge in its corresponding network, otherwise it is 0; d_{jl} is the Euclidean distance between the residues j and l in local structure (2) if there is an edge in its corresponding network, otherwise it is 0; λ is a parameter and is used to tune the weights of two objective functions. The solution $x_{ij} = 1$ implies the residue i is aligned with residue j, otherwise $x_{ij} = 0$.

We relax the integer quadratic programming as quadratic programming, and further use interior algorithm to solve it [9]. For the alignment result, if $x_{ij} > 0.5$, we let residue i align with residue j. In addition, we score the aligned residues i and j with $s_{ij} = w_{ij}$.

3 Uncovering hot spots in protein-protein interfaces

Here we use available protein structure data and alanine scanning mutagenesis data to show that those structurally and evolutionally conserved residues are likely to be hot spots.

We firstly retrieve the complexes with alanine mutations deposited in the ASEdb database. Then for each such complex, we retrieve the complexes created by molecules with the same molecule name in the PDB and the same family id in SCOP as the PPI family members. This data is compiled from Shulman-Peleg et al [10], PDB and SCOP. Furthermore, from the PPI family members we identify the corresponding chains having the same molecule name and the same family id as a protein family. Table 1 shows the 19 protein chains from 11 protein complexes and their associated protein families. In alanine scanning mutagenesis if a residue is mutated as alanine and its change of binding energy $\geq$2 kcal/mol, the residue is defined as a hot spot [3]. The 19 protein chains include 36 surface hot spots and 239 surface non-hot spots. Next we use PSAlign to align the surfaces whose proteins are in the same protein family and to identify hot spots.

Given the protein chains in the same family, now we predict hot spots based on PSAlign. We retrieve the protein surface from each protein chain and use surface in mutated chain to compare with other surfaces in the same family to find evolutionally and structurally conserved residues. Specifically for each residue i in the surface of the mutated chain, we give a score $s\_ratio(i)$ with the following procedures:

(1) Use PSAlign to align the surface of the mutated chain with the surfaces of other chains of the same family, and obtain the score s_{ij} for the residue i in mutated chain aligned with residue j in other chains of the same family respectively.

Table 1: Protein surface alignment dataset. Column 1 is the protein chains with experimental alanine mutations. Column 2 details the protein family members having the same functional description by PDB and the same family id by SCOP.

PDBid	Protein family members
1brsA	1brsA 1b2sA 1b27A 1x1uA 1b2uA
1brsD	1brsD 1b2sD 1b27D 1x1uD 1b2uD
1a4yA	1a4yA 1z7xY 1dfjI 2bexA
1a4yB	1a4yB 1z7xZ 1dfjE 2bexC
3hhrA	3hhrA 1a22A 1axiA 1hwgA 1hwhA
3hhrB	3hhrB 1a22B 1axiB 1hwgB 1hwhB
1bxiA	1bxiA 1emvA 1fr2A 1mz8A 1ujzA
1gc1C	1gc1C 1g9nC 1rzkC 1rzjC 1g9mC
3hfmY	3hfmY 1ua6Y 1j1oY 1j1pY 1uacY 1ic7Y 1c08C 1nbyC
3hfmL	3hfmL 1ua6L 1j1oL 1j1pL 1uacL 1ic7L 1c08A 1nbyA
3hfmH	3hfmH 1ua6H 1j1oH 1j1pH 1uacH 1ic7H 1c08B 1nbyB
1vfbC	1vfbC 1a2yC 1fdlY 1g7hC 1g7iC 1kipC
1vfbA	1vfbA 1a2yA 1fdlL 1g7hA 1g7iA 1kipA
1vfbB	1vfbB 1a2yB 1fdlH 1g7hB 1g7iB 1kipB
1dfjI	1dfjI 1a4yA 1z7xY 2bexA
1danL	1danL 1fakL 1wunL 1wtgL 1wqvL
1danT	1danT 1fakT 1wunT 1wtgT 1wqvT
1danU	1danU 1wunT 1wtgT 1wqvT
1cbwI	1cbwI 1tawB 1ca0I 1ejmB 3tgkI 1fakI 1p2kI 1f7zI

(2) Take the sum of the aligned scores for residue i and multiply it by the percent of residues in other chain aligned with residue i, i.e.,

$$s(i) = (\sum_{k=1}^{K_1} s_{ij_k})\frac{K_1}{N-1}$$

where K_1 is the number of residues in other chains that are aligned with i, s_{ij_k} is the score for the residue i aligned with residue j_k in other chains, and N is the number of chains in the family.

(3) Finally, normalize score $s(i)$ with the average score of aligned residues in mutated chain, i.e.

$$s\_ration(i) = \frac{s(i)}{\sum_{k=1}^{K_2} s(i_k)}$$

where K_2 is the number of residues i_k having $s(i_k) > 0$ in the surface of the mutated chain.

We differentiate hot spots from less important residues by setting a cutoff for s_ratio. If s_ratio of a residue is not less than the cutoff, we regard the residue as a hot spot. In practical, if $s\_ratio \geq 0.9$ the residue is defined as a hot spot. In the integer quadratic programming, we let the parameter $\lambda = 0.9$. In the following, we validate PSAlign for hot spot prediction through two cases and the whole dataset, and compare it with the state-of-the-art method Consurf [11]. Consurf calculates evolutionary conservation score within a homologous protein family for each residue in the protein chain. The results show that PSAlign performs better than Consurf for hot spot prediction.

3.1 PPI between E colicin DNases and immunity proteins

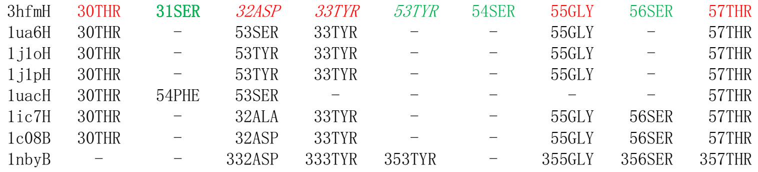
The E colicin DNases (1bxiA, 1emvA, 1fr2A, 1mz8A, 1ujzA) are bacterial toxins killing target microbial cells through random degradation of chromosomal DNA. Immunity proteins (1bxiB, 1emvB, 1fr2B, 1mz8B, 1ujzB) interacts with the E colicin DNases to resist the infection of microorganism. Here we use PSAlign to align the surface of 1bxiA with the surfaces of 1emvA, 1fr2A, 1mz8A, 1ujzA respectively to predict the hot spots of 1bxiA. MCL partitions the surface of 1bxiA into 5 communities. It is observed that 14 of 15 interface residues in the surface of 1bxiA fall into 2 communities, i.e. 7 interface residues fall into one community with community size 15 while the other 7 interface residues fall into another community with community size 10. There are 17 surface residues having records for the change of binding energy upon alanine mutations. The predicted results show that PSAlign (accuracy=0.65, sensitivity=0.5, specificity=0.73) performs better than Consurf (accuracy=0.59, sensitivity=0.5, specificity=0.64) for hot spot prediction.

3.2 PPI between HYHEL and HEL

HYHEL (3hfmH, 1ua6H, 1j1oH, 1j1pH, 1uacH, 1ic7H, 1c08B, 1nbyB) interacts with HEL (3hfmY, 1ua6Y, 1j1oY, 1j1pY, 1uacY, 1ic7Y, 1c08C, 1nbyC). MCL partitions the surface of 3hfmH into 12 communities. We find that the interface residues of 3hfmH almost fall into one community. There are total 8 interface residues in the surface and among them 7 residues falling into the community with community size 9. Figure 2 shows the alignment results that the community of 3hfmH consisting of interface residues compares with the corresponding communities of 1ua6H, 1j1oH, 1j1pH, 1uacH, 1ic7H, 1c08B, 1nbyB respectively based on PSAlign. There are 5 surface residues having records for the change of binding energy upon alanine mutations. Residue 32ASP is structurally and evolutionally conserved with s_ratio=2.4, and it is indeed a hot spot with $\Delta\Delta G$=2.0. In addition, hot spot 33TYR with $\Delta\Delta G$=6.0 has s_ratio=0.93. From the alignment result we can observe that 32ASP and 33TYR have sequence order independent evolutional and structural conservation, which can not be detected by sequence alignment. The performance of PSAlign for 3hfmH is that accuracy=0.6, sensitivity=0.5, specificity=1. In contrast, Consurf scores are all 1 for the mutated residues, and it does not give prediction for the hot spots if we take Consurf score=6 as a cutoff, so its accuracy=0.2, sensitivity=0, specificity=1.

3.3 PPIs in ASEdb

Table 2 presents the predicted results of PSAlign and Consurf on the whole dataset. It is obvious that the performance of PSAlign is better than Consurf. Furthermore, between hot spots and non-hot spots there is a statistically significant difference in s_ratio (P-value=0.05, Wilcoxon rank sum test) compared with insignificant difference in Consurf score (P-value=0.28, Wilcoxon rank sum test). Finally, we generate ROC curves of PSAlign and Consurf based on different cutoffs with s_ratio and Consurf score. Figure 3



3hfmH	30THR	**31SER**	*32ASP*	*33TYR*	*53TYR*	54SER	55GLY	56SER	57THR
1ua6H	30THR	-	53SER	33TYR	-	-	55GLY	-	57THR
1j1oH	30THR	-	53TYR	33TYR	-	-	55GLY	-	57THR
1j1pH	30THR	-	53TYR	33TYR	-	-	55GLY	-	57THR
1uacH	30THR	54PHE	53SER	-	-	-	-	-	57THR
1ic7H	30THR	-	32ALA	33TYR	-	-	55GLY	56SER	57THR
1c08B	30THR	-	32ASP	33TYR	-	-	55GLY	56SER	57THR
1nbyB	-	-	332ASP	333TYR	353TYR	-	355GLY	356SER	357THR

Figure 2: The local structure of 3hfmH, i.e. the community consisting of interface residues, compares with the corresponding local structures of 1ua6H, 1j1oH, 1j1pH, 1uacH, 1ic7H, 1c08B, 1nbyB respectively. Residues marked by red are conserved ($s\_ratio \geq 0.9$), marked by green are partial conserved ($s\_ratio < 0.9$). The residues 32, 33 and 53 shown as italic are hot spots ($\Delta\Delta G \geq 2$), the residue 31 shown as bold is non-hot spot ($\Delta\Delta G < 2$).

illustrates that PSAlign (ROC area AUC=0.6) performs better than Consurf (AUC=0.56) for prediction of hot spots. When only sequence evolutionary information is used, i.e. the parameter $\lambda = 1$, the results show worse performance than that of $\lambda = 0.9$. Hence, the structural conservation indeed reflects the feature of the hot spots.

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad sensitivity = \frac{TP}{TP+FN}$$

$$specificity = \frac{TN}{TN+FP} \quad precision = \frac{TP}{TP+FP}$$

$$f-measure = \frac{2*precision*sensitivity}{precision+sensitivity}$$

Table 2: Prediction of hot spots with PSAlign and Consurf [11]. The sensitivity and the specificity are calculated with a s_ration of 0.9 and Consurf score of 6.

Method	accuracy	sensitivity	specificity	precision	f-measure
PSAlign($\lambda = 0.9$)	0.65	0.5	0.67	0.19	0.28
Consurf	0.63	0.47	0.65	0.17	0.25
PSAlign($\lambda = 1$)	0.6	0.47	0.65	0.16	0.24

4 Discussion and Conclusion

Here we provide a systematic method that can find evolutionally and structurally conserved residues in protein surface. It incorporates protein sequence and structure data, and is based on an integer quadratic programming model. As an application, it can be used for location of hot spots in protein surface. Its main advantage is that it does not need the structure of protein complex or even the structure of the binding partner for hot spot prediction. With the development of structural biology and the availability of multiple structures of functionally related proteins, PSAlign is expected to become increasingly useful.

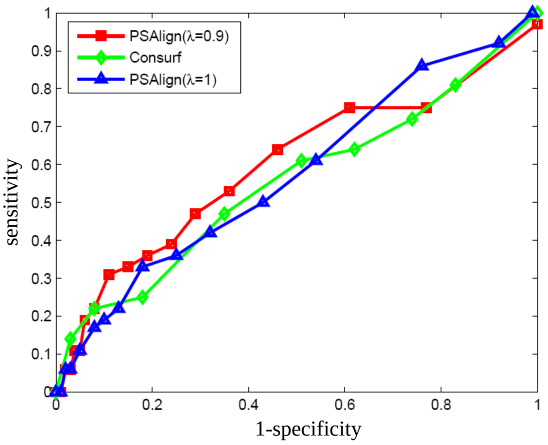


Figure 3: ROC curves of PSAlign and Consurf

References

[1] Wells JA and McClendon CL. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, 450(7172):1001–1009, 2007.

[2] Wells JA. Systematic mutational analyses of protein-protein interfaces. *Methods Enzymol*, 202:390–411, 1991.

[3] Keskin O, Ma B, and Nussinov R. Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology*, 345:1281–1294, 2005.

[4] Zhou HX and Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Structure, Function, and Genetics*, 44(3):336–343, 2001.

[5] Chung J, Wang W and Bourne PE. Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins*, 62:630–640, 2006.

[6] Van Dongen S. Graph Clustering by Flow Simulation. *PhD thesis, University of Utrecht*, May, 2005.

[7] Larkin MA, Blackshields G, and Brown NP et al. ClustalW and ClustalX version 2. *Bioinformatics*, 23(21):2947–2948, 2007.

[8] Xie L and Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc. Natl. Acad. Sci. USA*, 105(14):5441–5446, 2008.

[9] Li Zhenping, Zhang Shihua, Wang Yong, Zhang Xiang-Sun and Chen Luonan. Alignment of molecular networks by integer quadratic programming. *Bioinformatics*, 23(13):1631–1639, 2007.

[10] Shulman-Peleg A, Shatsky M, Nussinov R and Wolfson HJ. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biology*, 5:43, 2007.

[11] Glaser F, Pupko T, Paz I, Bell R, Bechor-Shental D, Martz E and Ben-Tal N. ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics*, 19:163–164, 2003.