International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:3, No:9, 2009

Identification and Analysis of Binding Site Residues in Protein-Protein Complexes

M. Michael Gromiha, Kiyonobu Yokota, and Kazuhiko Fukui

***Abstract*—**We have developed an energy based approach for identifying the binding sites and important residues for binding in protein-protein complexes. We found that the residues and residue-pairs with charged and aromatic side chains are important for binding. These residues influence to form cation-π, electrostatic and aromatic interactions. Our observation has been verified with the experimental binding specificity of protein-protein complexes and found good agreement with experiments. The analysis on surrounding hydrophobicity reveals that the binding residues are less hydrophobic than non-binding sites, which suggests that the hydrophobic core are important for folding and stability whereas the surface seeking residues play a critical role in binding. Further, the propensity of residues in the binding sites of receptors and ligands, number of medium and long-range contacts, and influence of neighboring residues will be discussed.

***Keywords*—**Protein-protein interactions; energy based approach; binding sites; propensity; long-range contacts; hydrophobicity.

I. Introduction

PROTEIN-PROTEIN interactions are important for most of the cellular processes in life. Hence, understanding the mechanism of protein-protein recognition at molecular level is of practical interest and has direct applications to functional genomics. The two major approaches to this problem are large-scale studies on protein-protein interaction networks and investigations on the general principles of recognition and prediction of their binding sites. Unraveling the mechanism of protein-protein recognition is a fundamental problem and it would aid in function prediction and drug design.

The availability of numerous numbers of protein-protein complexes enables researchers to analyze the binding sites in terms of amino acid composition, preference of residues, secondary structures, solvent accessibility, electrostatic patches, hydrophobic contacts, hydrogen bonding networks and so on [1-7]. The mapping of protein-protein interactions on protein sequences suggests that the hotspots can be predicted from amino acid sequences [8]. Furthermore, protein-protein interactions have been studied in terms of efficient clustering, stability calculations, conformational changes and docking simulations [9-11]. The concepts of protein-protein interactions in terms of experimental techniques, databases, organization, cooperativity and prediction of protein-protein and domain interactions have been reviewed in detail [12-15].

On the other hand, several methods have been proposed for identifying the binding sites in protein-protein complexes. Jones and Thornton [16] used surface patches for predicting protein-protein interaction sites. The combination of sequence and structural features as well as the information on nine consecutive residues, secondary structure of the central residue and average properties based on solvent accessibility, protrusion and depth has also been employed for detecting the binding sites from amino acid sequence [17]. Shulman-Pelag et al. [18] constructed a method based on multiple alignment for detecting binding sites in protein-protein complexes. It recognizes the spatially conserved physico-chemical interactions, which often involve energetically important hot-spot residues that are crucial for protein-protein associations. Ertekin et al. [19] proposed a method based on the fluctuation behavior of residues to predict the putative protein binding sites. Further, machine learning techniques have been widely used to identify the binding sites in protein-protein complexes [20-23].

In most of these studies, binding sites have been defined with a criteria based on the contacts between amino acid residues in two partners of protein-protein complexes. The atomic contacts between C_α atoms, C_β atoms, any atoms in a residue as well as the distances of 5-7Å have been used to assign the contacts [24-26]. These criteria include the repulsive interactions in which two residues are close to each other. In addition, the residue pairs with different distances have been treated in a same manner. Other methods employed shape complimentarity, homologs, conservation and amino acid properties. In this work, we have developed a new approach based on interaction energy for defining the binding sites. We observed that the binding sites are dominated with aromatic and charged residues indicating the importance of electrostatic, aromatic and cation-π interactions. Further, we have analyzed the preference of interacting partners, variation of surrounding hydrophobicity and influence of medium and long-range contacts in binding and non-binding sites as well as the preference of residues in binding segments of different lengths. The salient features of the results will be discussed.

MMG, KY and KF are with the Computational Biology Research Center of the National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, Japan (phone: +81-3-3599-8046; fax: +81-3-3599-8081; e-mail: michael-gromiha@aist.go.jp).

II. Interaction Energy in Protein-Protein Complexes

A. Dataset

We have developed a non-redundant dataset of 153 heterodimer protein-protein complexes from Protein Data Bank [27] that have the sequence identity of less than 25% [28]. In each protein-protein complex, proteins with high and low molecular weights are termed as receptors and ligands, respectively.

B. Computation of Interaction Energy

The interaction free-energy between atoms in protein-protein complexes was calculated using AMBER potential [29], which is widely used to analyze and understand the recognition mechanism in protein complexes [30]. It is given by:

$$E_{inter} = \Sigma [(A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6}) + q_i q_j/\varepsilon r_{ij}] \quad (1)$$

where $A_{ij} = \varepsilon_{ij}^*(R_{ij}^*)^{12}$ and $B_{ij} = 2\ \varepsilon_{ij}^*(R_{ij}^*)^6$; $R_{ij}^* = (R_i^* + R_j^*)$ and $\varepsilon_{ij}^* = (\varepsilon_i^*\ \varepsilon_j^*)1/2$; R^* and ε^* are, respectively, the van der Waals radius and well depth and these parameters are obtained from Cornell et al. [29]; q_i and q_j are, respectively, the charges for the atoms i and j, and r_{ij} is the distance between them.

C. Distribution of Interaction Energy

In a protein-protein complex, we have computed the interaction energy (Eqn. 1) of each residue in receptor with all residues in ligand. We have repeated the calculations for all the complexes and analyzed the interaction energies of all the residues in intervals of 0.1 from -15 to 5 kcal/mol. The frequency of occurrence of residues in receptors at different intervals of interaction free energies (from -2 to 1 kcal/mol) are displayed in Fig. 1.

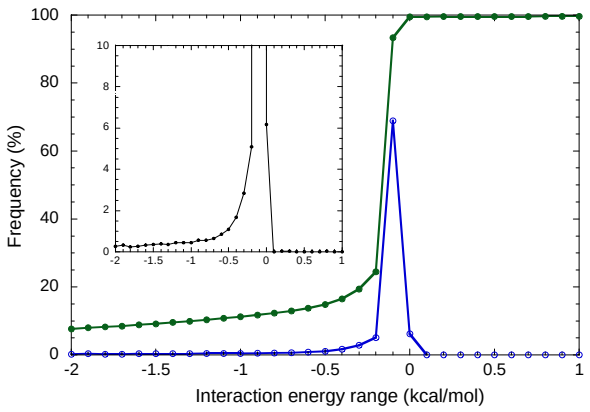

Fig. 1 Occurrence of amino acid residues in different ranges of interaction energies in receptors. The open and closed circles show the fraction and total percentage of residues. The expanded data for the percentage fraction of residues in different ranges of interaction energies is shown in the inset

In this figure, we present the results for both the fraction of residues and total percentage of residues at each interval. We observed that 7.7% of the residues have strong interactions with ligands and the interaction free energy is less than -2 kcal/mol. On the other hand, 6.2% of residues have repulsive energies and 77% of the residues have the interaction energy in the range of -0.3 to 0 kcal/mol, which might be due to the presence of residues that are far away in 3D structures [31]. Among 48,657 residues 5255 of them have the interaction free-energy less than -1 kcal/mol. Interestingly, we observed similar number of residues (4957) in ligands that are interacting with receptors. As the total number of residues in ligands are almost half of that in receptors the percentage of interacting residues are twice to that in receptors. Similar characteristics are observed for the binding site residues obtained with the contacts between residues in receptors and ligands in protein-protein complexes.

D. Comparison with Distance Based Methods

We have compared the results obtained with the energy criteria used in this work and the criteria with different cutoff distances for defining binding site residues [24-26] in Table I.

TABLE I
Number and Percentage of Binding Site Residues Using Different Methods

Criterion	Cutoff	N_{bind}	$\%_{bind}$
Energy	<1 kcal/mol	5255	10.8
Energy	<0.5 kcal/mol	6718	13.8
C_α distance	6Å	1972	4.0
C_β distance	6Å	3449	7.1
Any heavy atoms	5Å	6644	13.6

We noticed that the number and percentage of binding site residues obtained with energy based approach is similar to the one defined with the distance of 5Å between any heavy atoms in receptors and ligands. However, the analysis of binding site residues obtained in these approaches showed significant differences between them. Only 1459 residues are common to each other and this result indicates the importance of considering the energy between different atoms to define the binding residues. In addition, 4% of the residues have strong repulsive energies and all these residues have been identified as binding residues in distance based criteria, which are not probable binding residues in protein-protein complexes.

III. Preference of Residues in Binding Sites

A. Computation of Binding Propensity

The binding propensity for the 20 amino acid residues in both receptors and ligands in protein-protein complexes has been developed as follows: we have computed the frequency of occurrence of amino acid residues in binding sites (f_b) and in the receptor (ligand) as a whole (f_t). The binding propensity (P_{bind}) is calculated using the equation:

$$P_{bind}(i) = f_b(i)/f_t(i) \quad (2)$$

where, i represents each of the 20 amino acid residues.

B. Binding Propensity for the 20 Amino Acid Residues

We have computed the binding propensity in both receptors and ligands using Eqn.2 and the results for receptors are presented in Table II.

TABLE II
BINDING PROPENSITY OF AMINO ACID RESIDUES IN RECEPTORS

Residue	N_{bind}	N_{tot}	Propensity
ALA	264	3673	7.19
ASP	261	2831	9.22
CYS	82	800	10.25
GLU	323	3233	9.99
PHE	292	2057	14.20
GLY	270	3441	7.85
HIS	140	1066	13.13
ILE	304	2735	11.12
LYS	271	2840	9.54
LEU	482	4510	10.69
MET	132	1084	12.18
ASN	240	2161	11.11
PRO	262	2317	11.31
GLN	244	1917	12.73
ARG	366	2414	15.16
SER	272	2968	9.16
THR	287	2653	10.82
VAL	354	3505	10.10
TRP	132	764	17.28
TYR	277	1807	15.33

We observed that the aromatic as well as positively charged residues contribute significantly to interact between receptors and ligands. Interestingly, the behavior is similar in both receptors and ligands [31]. This result indicates the importance of cation-π, aromatic and electrostatic interactions for the recognition of protein-protein complexes. The highly favorable pairs of interacting residues also reveal the presence of several pairs formed by aromatic and charged residues.

IV. COMPARISON WITH EXPERIMENTS

A. Protein-protein Interaction Database

We have compared the results obtained in this work with experimental binding energies of protein-protein complexes. This has been done with the data on changes in binding free energy change upon amino acid substitutions. We have developed a protein-protein interaction thermodynamic database (PINT), which has the data for more than 140 complexes and 1700 interactions [32]. The database is available at http://www.bioinfodatabase.com/pint/index.html and the features are shown in Fig. 2.

Fig. 2 Snapshot of PINT database showing the available features

The search on PINT showed the presence of 217 interactions, which have the difference in binding free energy of >2 kcal/mol. Generally residues that can cause the binding free energy of >2 kcal/mol are identified as hotspots. Further analysis on 217 interactions revealed that 68 of them are unique. We have analyzed all the unique interactions and we observed that 38 residues are charged, 32 of them are positive charged and aromatic. On the other hand only seven residues are hydrophobic. This result demonstrates the importance of electrostatic, cation-π and aromatic interactions for the recognition of protein-protein complexes. Our computational analysis revealed the importance of these interactions, showing the good agreement with experiments.

B. E6AP-UbcH7 Complex

Eletr and Kuhlman [33] measured the binding free energies of 49 mutants in E6AP-UbcH7 complex and 15 of them are identified as hotspots. We have analyzed all the hotspot residues and observed that 10 residues are positively charged/ aromatic, 9 are positively/negatively charged and 3 are hydrophobic. Further, the replacement of F63A altered the binding free energy of 3 kcal/mol. We have analyzed the energetic contribution of F63 in 1C4Z and found that F63 in UbcH7 makes a strong aromatic interaction with Y694 in E6AP and the interaction free energy is -1.2 kcal/mol. Fig. 3a shows the aromatic interactions between the residues F63 and Y694 in E6AP-UbcH7 complex.

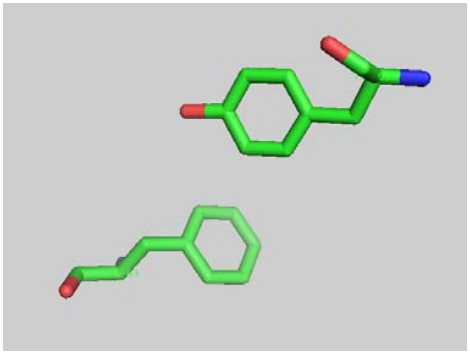

Fig. 3a Interactions between receptors and ligands in E6AP-UbcH7 complex

C. Interleukin4 Receptor Binding Protein Complex

Zhang et al. [34] carried out binding experiments on interleukin4 receptor binding protein complex (1IAR) and reported binding free energies for 29 mutants. The analysis on ΔΔG values shows the presence of 11 hotspots. Interestingly, six of them are cation-π interaction forming residues, 3 are charged and 2 are hydrophobic residues. We have analyzed the contribution due to different residues and the interaction between Y127 in interleukin4 and R85 in binding protein is shown in Figure 3b. We observed the presence of a cation-π interaction and the substitution of Y127A changed the binding free energy to 2.2 kcal/mol.

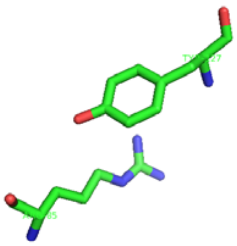

Fig. 3b Interactions between receptors and ligands in Interleukin4 receptor binding protein complex

D. Ras-Rap Complex

Kiel et al. [35] studied the thermodynamic behavior of binding in Ras-Rap complex and measured the binding free energy of 27 mutants in both Ras and Rap protein. We noticed the presence of six hotspots and four of them involved charged residues, Lys, Arg and Asp. Five residues have the capability of cation-π interactions and there is no residue with hydrophobic behavior. In this experiment Kiel et al. [35] reported that the replacement of K32A in Rap with wild type Ras altered the free energy of 2.5 kcal/mol whereas the substitution of D238A in Ras with wild type Rap contributed to the free energy of 3.9 kcal/mol. We have analyzed the interaction between K32 and D238 and the contribution towards electrostatic interaction is shown in Fig. 3c. This analysis verifies the importance of electrostatic interactions obtained in this work with experimental observations.

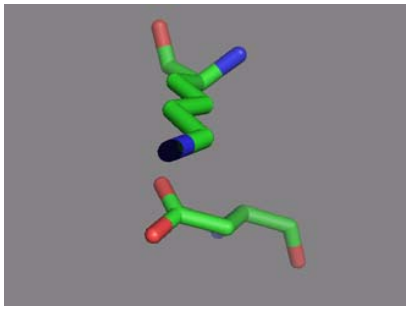

Fig. 3c Interactions between receptors and ligands in Ras-Rap complex

V. Structural Analysis of Binding and Non-Binding Sites in Protein-Protein Complexes

A. Surrounding Hydrophobicity

We have computed the surrounding hydrophobicity of residues using the following procedure: The amino acid residues in a protein molecule are represented by their α-carbon atoms. The surrounding hydrophobicity (H_p) of a given residue is defined as the sum of hydrophobic indices of various residues, which appear within 8 Å radius limit from it [36,37].

$$H_p(i) = \Sigma\, n_{ij}\, h_j \qquad (3)$$

where, n_{ij} is the total number of surrounding residues of type j around ith residue of the protein and h_j is the experimental hydrophobic index of residue type j in kcal/mol [38,39]

We have computed the surrounding hydrophobicity of all the residues in binding and non-binding sites using Eqn.3 and the results obtained for various ranges of hydrophobicity are presented in Fig. 4.

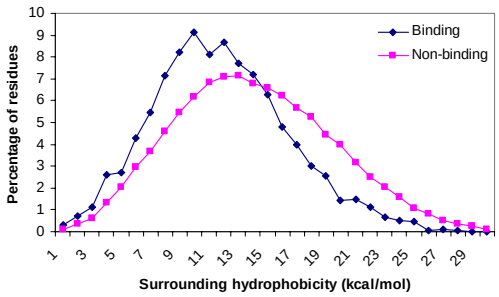

Fig. 4 Frequency of occurrence of binding and non-binding residues at various ranges of surrounding hydrophobicity

We observed that more binding site residues prefer to have the surrounding hydrophobicity of <16 kcal/mol than non-binding residues. On the other hand, more number of non-binding residues accommodates high hydrophobic regions compared with binding site residues. This analysis shows that the presence of polar residues and the location of such residues in surface influence binding in protein-protein complexes.

B. Medium and Long-Range Contacts

The residues in a protein molecule are represented by their α-carbon atoms. Using the C_α coordinates, a sphere of radius 8Å is fixed around each residue and the residues occurring in this volume are identified. The composition of surrounding residues is analyzed in terms of the location at the sequence level and the contributions from < ±3 residues are treated as short range contacts, ±3 or ±4 residues as medium range contacts and > ±4 residues are treated as long-range contacts [40]. Fig. 5 shows the illustration of short, medium and long-range contacts.

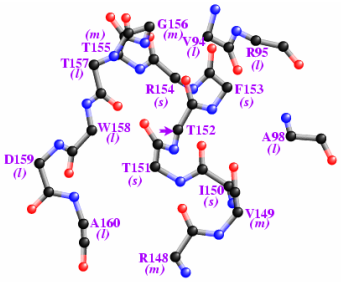

Fig. 5 Representation of short, medium and long-range contacts in protein structures. A typical example for the contacting residues of Thr152 in T4 lysozyme within 8Å is shown: s: short-range contacts, m: medium-range contacts and l: long-range contacts

The percentage of residues with different numbers of long-range contacts is shown in Fig. 6.

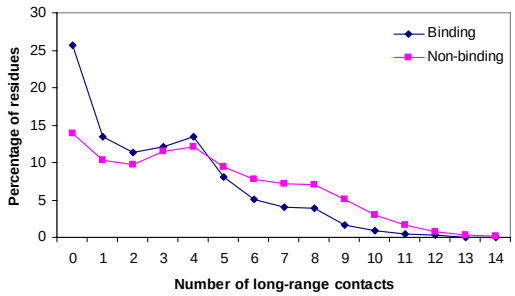

Fig. 6 Percentage of binding and non-binding residues with different long-range contacts

The data shown in Fig. 6 reveals a transition at four long-range contacts and the more binding residues prefer to have less than 4 long-range contacts than non-binding residues. An opposite trend is observed for the binding and non-binding residues with more than 4 long-range contacts. This result suggests that the non-binding residues tend to form long-range contacts with several residues in protein structures where as the binding site residues have less number of long-range contacts and these residues interact with the residues in partner protein.

ACKNOWLEDGMENT

This research was supported by Strategic International Cooperative Program, Japan Science and Technology Agency (JST).

REFERENCES

[1] Ofran Y, Rost B, Analysing six types of protein-protein interfaces. J Mol Biol 2003, 325: 377–387.
[2] Chakrabarti P, Janin J Dissecting protein-protein recognition sites. Proteins 2002, 47: 334–343.
[3] Guharoy M, Chakrabarti Conservation and relative importance of residues across protein-protein interfaces. P Proc Natl Acad Sci U S A 2005, 102: 15447–15452.
[4] Ma B, Elkayam T, Wolfson H, Nussinov R Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci U S A 2003, 100: 5772–5777.
[5] Sheinerman FB, Honig B On the role of electrostatic interactions in the design of protein-protein interfaces. J Mol Biol 2002, 318: 161–177.
[6] Kortemme T, Baker D A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci U S A 2002, 99: 14116–14121.
[7] Kundrotas PJ, Alexov E Electrostatic properties of protein-protein complexes Biophys J 2006, 91: 1724–1736.
[8] Ofran Y, Rost B. Protein-protein interaction hotspots carved into sequences. PLoS Comput Biol. 2007; 3: e119.
[9] Aung Z, Tan SH, Ng SK, Tan KL. PPiClust: efficient clustering of 3D protein-protein interaction interfaces. J Bioinform Comput Biol. 2008; 6: 415-33.
[10] Urakubo Y, Ikura T, Ito N. Crystal structural analysis of protein-protein interactions drastically destabilized by a single mutation Protein Sci. 2008; 17: 1055-65.
[11] Lensink MF, Méndez R. Recognition-induced conformational changes in protein-protein docking Curr Pharm Biotechnol. 2008; 9: 77-86.
[12] Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part I. Experimental techniques and databases PLoS Comput Biol. 2007; 3: e42
[13] Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners PLoS Comput Biol. 2007; 3: e43.
[14] 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.
[15] Keskin O, Ma B, Rogale K, Gunasekaran K, Nussinov R. Protein-protein interactions: organization, cooperativity and mapping in a bottom-up systems biology approach. Phys. Biol. 2005, 2, S24-35
[16] Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis J Mol Biol 1997, 272: 133–143.
[17] Sikić M, Tomić S, Vlahovicek K. Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. PLoS Comput Biol. 2009; 5: e1000278.
[18] Shulman-Peleg A, Shatsky M, Nussinov R, Wolfson HJ. Nucleic Acids Res. 2008; 36: W260-264.
[19] Ertekin A, Nussinov R, Haliloglu T. Association of putative concave protein-binding sites with the fluctuation behavior of residues.Protein Sci. 2006; 15: 2265-2277
[20] Fariselli P, Pazos F, Valencia A, Casadio R Prediction of protein--protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002, 269: 1356–1361.
[21] Koike A, Takagi T Prediction of protein-protein interaction sites using support vector machines.Protein Eng Des Sel 2004, 17: 165–173.
[22] Res I, Mihalek I, Lichtarge O. An evolution based classifier for prediction of protein interfaces without using protein structures Bioinformatics. 2005; 21: 2496-501
[23] Ofran Y, Rost B. ISIS: interaction sites identified from sequence. Bioinformatics. 2007;23(2):e13-6
[24] Li W, Keeble AH, Giffard C, James R, Moore GR, Kleanthous C. Highly discriminating protein-protein interaction specificities in the context of a conserved binding energy hotspot J Mol Biol. 2004; 337: 743-59
[25] Keskin O, Tsai CJ, Wolfson H, Nussinov R. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications Protein Sci. 2004; 13:1043-55.
[26] Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces Proteins. 2001; 43: 89-102.
[27] Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data Nucleic Acids Res. 2007; 35: D301-303.
[28] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970; 48: 443-453.
[29] Cornell,W.D.; Cieplak, P.; Bayly,C. I.; Gould, I. R.;Merz, K.M., Jr.; Ferguson,D.M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J.W.; Kollman, P. A. 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995, 117, 5179–5197.

[30] Pichierri F, Aida M, Gromiha MM, Sarai A. Free-energy maps of base-amino acid interactions in protein-DNA recognition. J. Amer. Chem. Soc. 1999, 121, 6152-6157.
[31] Gromiha MM, Yokota K, Fukui K. Energy based approach for understanding the recognition mechanism in protein-protein complexes. Mol. Biosystems, DOI: 10.1039/B904161N.
[32] Kumar MD, Gromiha MM. PINT: protein-protein interactions thermodynamic database. Nucleic Acids Res. 2006; 34: D195-8.
[33] Eletr ZM, Kuhlman B. Sequence determinants of E2-E6AP binding affinity and specificity J Mol Biol. 2007; 369: 419-28
[34] Zhang JL, Simeonowa I, Wang Y, Sebald W. The high-affinity interaction of human IL-4 and the receptor alpha chain is constituted by two independent binding clusters. J Mol Biol. 2002; 315: 399-407
[35] Kiel C, Serrano L, Herrmann C. A detailed thermodynamic analysis of ras/effector complex interfaces. J Mol Biol. 2004; 340: 1039-58.
[36] Manavalan P, Ponnuswamy PK. Hydrophobic character of amino acid residues in globular proteins. Nature. 1978;275(5681):673-4.
[37] Ponnuswamy PK. Hydrophobic characteristics of folded proteins. Prog Biophys Mol Biol. 1993;59(1):57-103.
[38] Nozaki Y, Tanford C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. J Biol Chem. 1971;246(7):2211-7.
[39] Jones DD. Amino acid properties and side-chain orientation in proteins: a cross correlation approach. J Theor Biol 1975;50:167-183.
[40] Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. Prog Biophys Mol Biol. 2004;86(2):235-77.