BMC Structural Biology

Software

A tool for calculating binding-site residues on proteins from PDB structures

Jing Hu[1,2] and Changhui Yan*[1]

Address: [1]Department of Computer Science, Utah State University, Logan, UT, USA and [2]Department of Mathematics & Computer Science, Franklin & Marshall College, Lancaster, PA, USA

Email: Jing Hu - jing.hu@fandm.edu; Changhui Yan* - charles.yan@usu.edu

* Corresponding author

Published: 3 August 2009

BMC Structural Biology 2009, **9**:52 doi:10.1186/1472-6807-9-52

Received: 1 March 2009
Accepted: 3 August 2009

This article is available from: http://www.biomedcentral.com/1472-6807/9/52

Abstract

Background: In the research on protein functional sites, researchers often need to identify binding-site residues on a protein. A commonly used strategy is to find a complex structure from the Protein Data Bank (PDB) that consists of the protein of interest and its interacting partner(s) and calculate binding-site residues based on the complex structure. However, since a protein may participate in multiple interactions, the binding-site residues calculated based on one complex structure usually do not reveal all binding sites on a protein. Thus, this requires researchers to find all PDB complexes that contain the protein of interest and combine the binding-site information gleaned from them. This process is very time-consuming. Especially, combing binding-site information obtained from different PDB structures requires tedious work to align protein sequences. The process becomes overwhelmingly difficult when researchers have a large set of proteins to analyze, which is usually the case in practice.

Results: In this study, we have developed a tool for calculating binding-site residues on proteins, TCBRP http://yanbioinformatics.cs.usu.edu:8080/ppbindingsubmit. For an input protein, TCBRP can quickly find all binding-site residues on the protein by automatically combining the information obtained from all PDB structures that consist of the protein of interest. Additionally, TCBRP presents the binding-site residues in different categories according to the interaction type. TCBRP also allows researchers to set the definition of binding-site residues.

Conclusion: The developed tool is very useful for the research on protein binding site analysis and prediction.

Background

Proteins perform various functions through interactions with other molecules, such as DNA, RNA, proteins, carbohydrates, and ligands. To understand the mechanisms of these interactions, many studies have been performed to analyze the properties of binding sites on proteins, such as residue composition, secondary structure, geometric shape, electrostatic potentials, etc [1-10]. To perform such an analysis, researchers must first identify the amino acid residues (referred to as *binding-site residues*) that are involved in the interactions. In other studies where the goal is to build computational predictors for predicting functional sites on proteins (e.g. DNA-binding sites, RNA-binding sites, protein-protein binding sites), researchers also need to identify binding-site residues on the training and test sets to train and evaluate their predictors [11-17].

In most, if not all, of the above-mentioned studies, the binding-site residues are calculated from the 3-dimensional (3D) structures deposited in Protein Data Bank (PDB) [18]. Usually, researchers collected a non-redundant set of PDB structures, and then calculated binding-sites based on the PDB structures. However, one serious problem with this approach is that a protein may have multiple binding sites that interact with different interacting partners, but one PDB structure usually does not show all of these interactions. For example, T7 RNA polymerase appears in both PDB 1ARO and 1QLN. 1ARO reveals the binding-site residues on T7 RNA polymerase that are involved in the protein-protein interaction (red color in Figure 1A), while 1QLN reveals the binding-site residues on T7 RNA polymerase that are involved in DNA binding (magentas color in Figure 1B) and RNA binding (brown color in Figure 1B). Even when two PDB structures reveal the same type of interaction on the same protein, the binding-sites can still be different depending on the interacting partner. For example, both 1UON and 1N1H are a complex of retrovirus polymerase lambda-3 with RNA, but 1UON shows that the RNA-binding site on lambda-3 consists of 59 residues (red color in Figure 2A), while 1N1H shows that the RNA-binding site of lambda-3 contains only 27 residues (red color in Figure 2B).

Thus, for the same protein, different sets of binding-site residues might be obtained depending on the PDB structure that is considered, and a residue of a protein may be defined as binding-site residue in one PDB structure but as non-binding-site residue in another. This inconsistency can cause serious problems in research. Thus, for a given protein, researchers need to identify all PDB structures that contain the protein, and calculate binding-site residues on the protein using all of them.

After users have found all the PDB structures that contain a given protein, the protein sequences shown in different PDB structures must be aligned properly to combine the binding-site information obtained from different structures. This step is not as simple as it may first appear. It cannot be done by matching the sequence indexes of residues in the PDB structures, because the same protein chain may have different sequence indexing in different PDB structures. For example, 1qqi_A and 1gxp_A are the same protein chain in different PDB structures. In PDB 1gxp, the first residue in chain A is ALA with sequence index of 127. However, the same residue in PDB 1qqi has an index of 2. It can neither be done by performing a simple one-to-one mapping between the two sequences from head to tail, because residue missing occurs frequently in PDB structures. Thus, researchers need a tool that can efficiently combine binding-site information from different PDB structures.

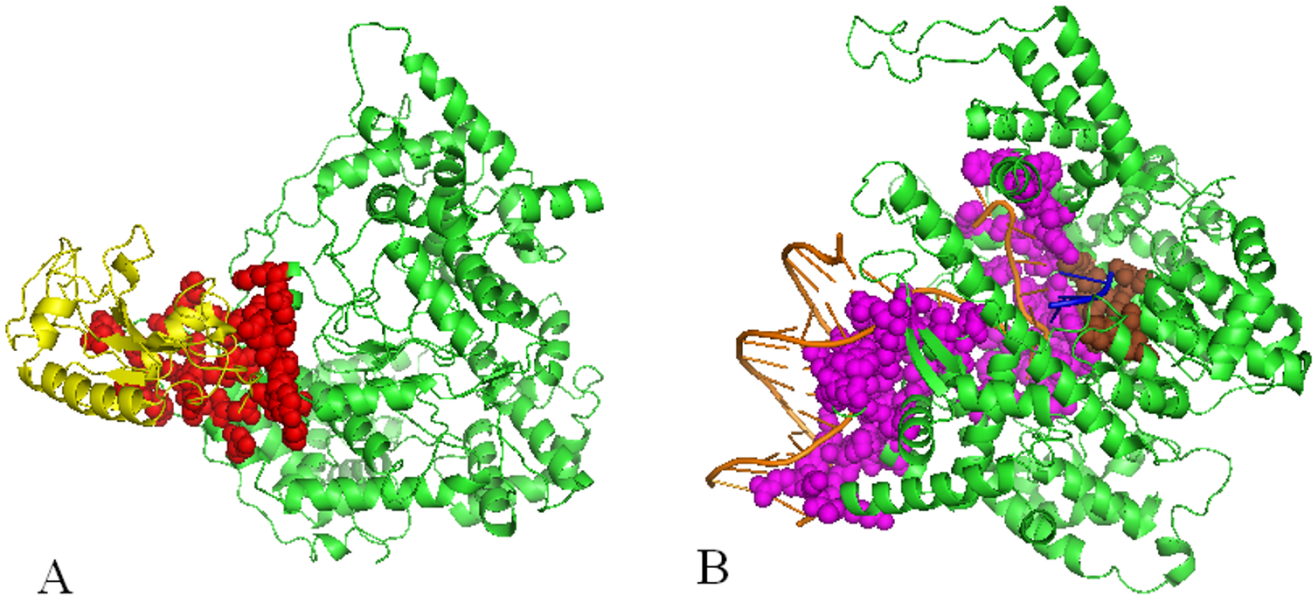

Figure 1
Binding-site residues on the T7 RNA polymerase shown by different PDB structures. **A**: PDB id 1ARO: A complex of T7 RNA polymerase with T7 Lysozyme. Green: T7 RNA polymerase; Yellow: T7 Lysozyme; Red: protein-binding residues on T7 RNA polymerase; **B**: PDB id 1QLN: A complex of T7 RNA polymerase with DNA and RNA. Green: T7 RNA polymerase; Orange: DNA; Blue: RNA; Magentas: DNA-binding residues; Brown: RNA-binding residues.

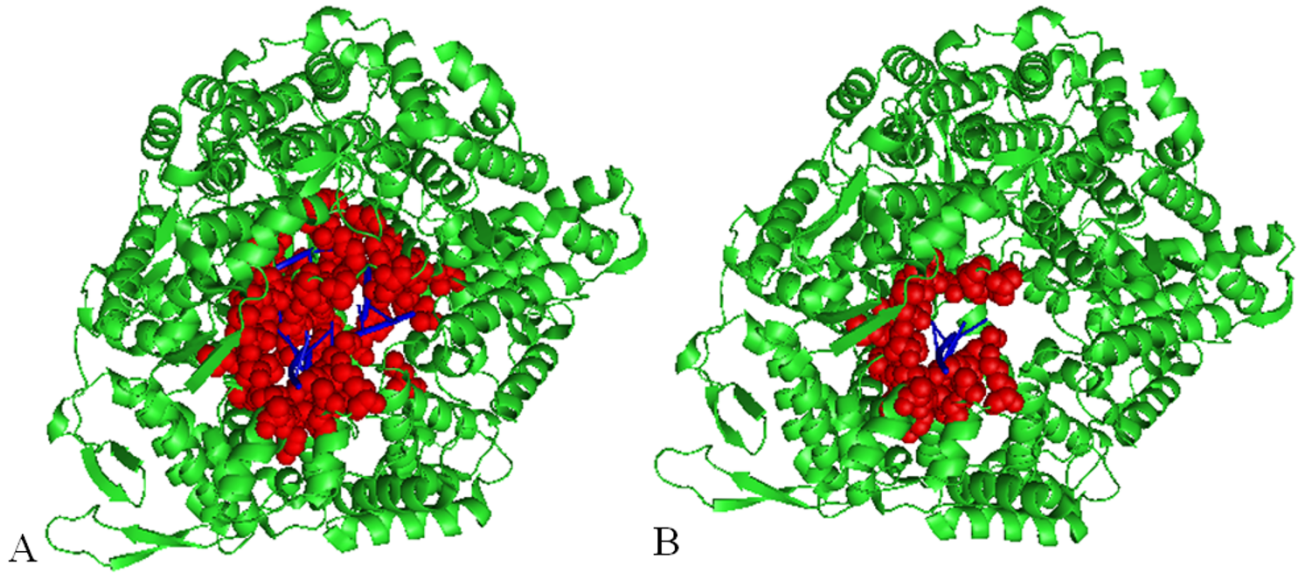

Figure 2
Different PDB structures show different RNA-binding residues on the retrovirus polymerase lambda-3. **A**: PDB id 1UON; **B** PDB id 1N1H. Blue: RNA; Green: retrovirus polymerase lambda-3; Red: RNA-binding residues on lambda-3.

The abovementioned needs become overwhelmingly impressive when users have a large set of proteins to analyze. Against these needs, we have developed TCBRP, a tool for calculating binding-site residues on proteins. For an input protein, TCBRP can quickly find all binding-site residues on it by integrating binding-site information obtained from all PDB structures that contain the protein of interest. Additionally, the TCBRP presents the binding-site residues by categories based on the type of the molecule that they contact, e.g. DNA, RNA, protein, carbohydrates, and ligands. An extra benefit of TCBRP is that it allows users to choose the definition of binding-site residues.

Implementation

Figure 3 shows the schema of TCBRP. First, users input a protein of interest and choose a definition of binding-site residues. There are two types of definition for binding-site residues. One is based on the reduction of solvent accessible surface upon the formation of complex [7]. A residue is defined to be a binding-site residue if its solvent accessible surface area (ASA) is reduced by at least a certain amount (default threshold is 1 $Å^2$) during the formation of the complex. The second definition is based on the atom distance [5]. A residue is defined as a binding-site residue if its distance to the interacting partner is less than a certain distance (default threshold is 5 Å). For both definitions, users can set the threshold (See figure 4).

Upon the input, TCBRP searches the entire PDB database to find all the complex structures that contain a protein that shares at least 95% sequence similarity with the input protein. Then, the biological units derived from these structures are used to calculate binding-site residues on the protein of interest. We use biological units instead of the raw PDB structure because the biological units show the functional state of the protein in life systems. Additionally, using biological units can avoid the artificial interactions that due to artificial packing in the raw PDB structures. TCBRP focuses on the binding sites involved in inter-molecule interactions, because they correspond to functional sites on proteins. Intra-molecule interactions

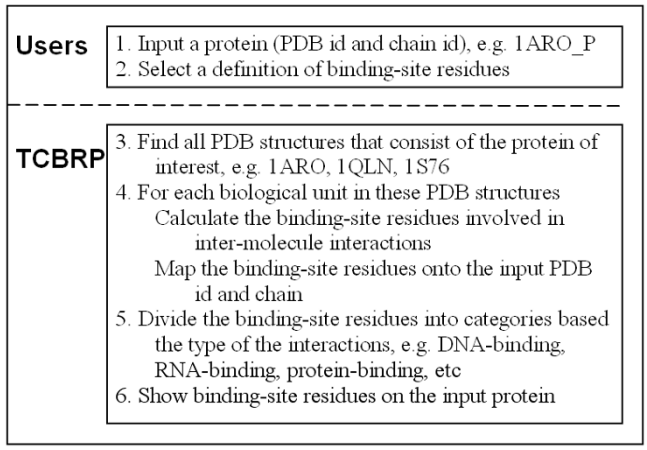

Figure 3
The schema of TCBRP.

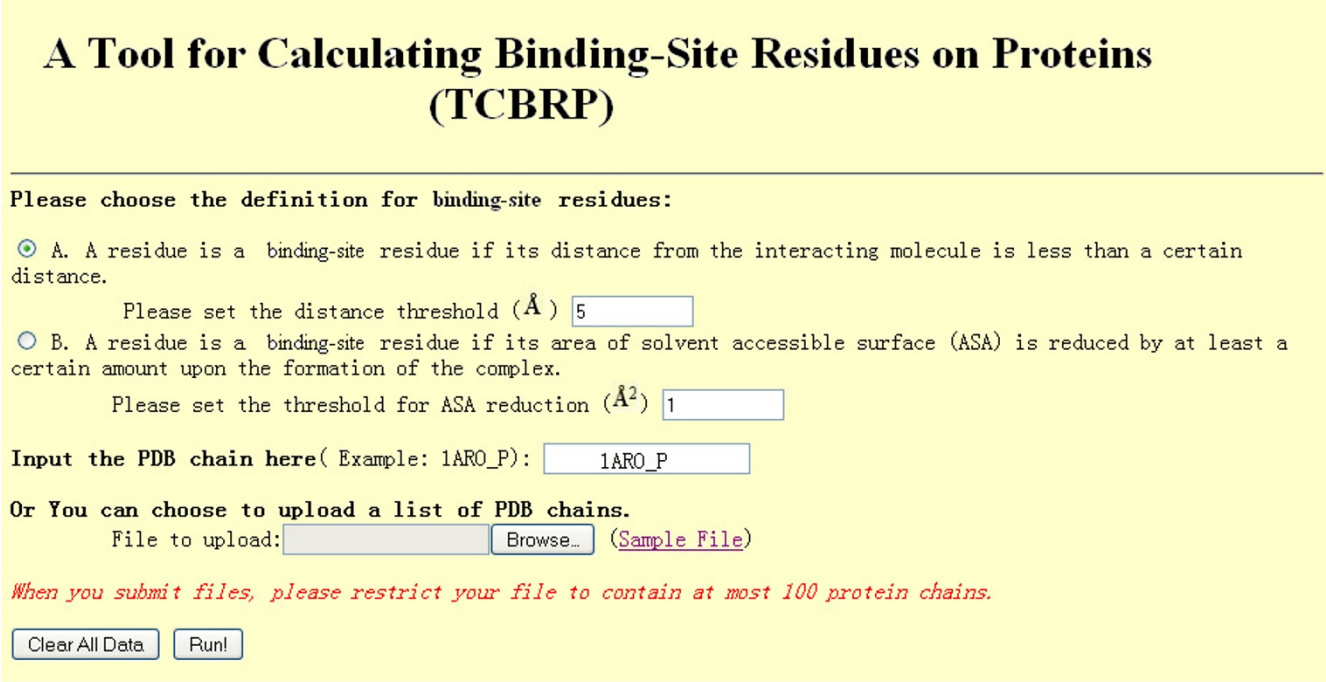

Figure 4
Input form of TCBRP.

that involve residues from the same chain or from different chains of the same molecule are discarded.

To reveal all binding sites on the input protein, the binding-site residues obtained from different biological units must be mapped to the input protein. To do this, the sequences of the different copies of the protein in different PDB structures must be aligned properly. In TCBRP, this step is achieved by aligning the protein sequences in PDB structures with the protein sequence found in the Uniprot [19] using global alignment.

Proteins are involved in various functions. Depending on the interacting partner, the binding sites on proteins can be divided into different categories, such as DNA-binding sites, RNA-binding sites, protein-protein binding sites, carbohydrate-binding sites, and ligand-binding sites. In many studies, researchers like to distinguish different types of protein binding sites. In response to this need, when a protein is involved in different types of interactions, the TCBRP show the binding-site residues for every type of interaction separately (Figure 5).

Using TCBRP, users can input one protein chain in a time, or input a file of protein chains in a batch. For a protein that consists of multiple chains, users can submit a file consisting of all the chains, then the TCBRP will show the binding-site residues on each chain that are involved in inter-molecule interactions.

Results and discussion

Figures 4 and 5 show an example of input and output for TCBRP. Assume that 1ARO is one of the PDB structures from which a user wants to find binding sites. Note that 1ARO is a complex of T7 RNA Polymerase (chain P) with T7 Lysozyme (Chain I). Without TCBRP, the user may use 1ARO to calculate binding-site residues on T7 RNA polymerase and only find 26 binding-site residues that correspond to the interaction between T7 RNA Polymerase and T7 Lysozyme. However, RNA polymerase interacts with multiple molecules including RNA, DNA, and proteins. In the research on functional site prediction and analysis, the user will need to find all the functional sites on the T7 RNA polymerase. To obtain these results without TCBRP, the user would need to go through a long and painful process of finding all complexes that contain T7 RNA polymerase, calculating binding-site residues using each of the complexes, and combining the information given by different structures. Using TCBRP, the user can obtain the results easily. Figure 4 shows the input page. Here, the input is the P chain of 1ARO, which is T7 RNA polymerase. Upon the input, the TCBRP automatically finds all PDB structures that contain T7 RNA polymerase, i.e. 1ARO, 1QLN, 2PI4, 1S76, 1MSW, 1CEZ, 1S77, and

```
PDB Chain submitted: 1ARO_P
Distance threshold: 5
Binding-site residues calculated based on the input PDB structure:
>1ARO_P
    Protein binding: 303P, 307P, 309P, 310P, 311P, 312P, 717P, 720P, 721P, 723P, 724P, 726P, 728P, 729P,
    Ligand binding:  125P, 136P, 137P, 140P, 296P, 308P, 313P, 316P, 317P, 397P, 398P, 400P, 401P, 431P,
Binding-site residues calculated based on other PDB structures that contain the same protein chain:
>1QLN_A
    DNA binding:     93P, 94P, 95P, 96P, 97P, 98P, 99P, 100P, 101P, 102P, 104P, 131P, 133P, 135P, 136P,
    RNA binding:     425P, 435P, 436P, 441P, 784P, 810P, 811P, 812P
    Ligand binding:  129P, 394P, 437P
>2PI4_A
    DNA binding:     93P, 94P, 95P, 96P, 97P, 98P, 99P, 100P, 101P, 102P, 104P, 131P, 133P, 135P, 136P,
    Ligand binding:  394P, 425P, 435P, 436P, 437P, 471P, 472P, 537P, 538P, 539P, 541P, 571P, 627P, 631P,
>1S76_D
    DNA binding:     50P, 53P, 54P, 57P, 158P, 160P, 163P, 164P, 182P, 267P, 298P, 385P, 392P, 396P, 421
    RNA binding:     300P, 386P, 389P, 390P, 393P, 394P, 397P, 425P, 435P, 436P, 437P, 441P, 742P, 746P,
    Ligand binding:  537P, 538P, 539P, 540P, 541P, 542P, 571P, 627P, 631P, 635P, 639P, 784P, 787P, 812P,
>1MSW_D
    DNA binding:     50P, 57P, 164P, 267P, 298P, 300P, 385P, 386P, 388P, 391P, 392P, 396P, 421P, 422P, 4
    RNA binding:     301P, 389P, 390P, 393P, 394P, 425P, 435P, 436P, 437P, 441P, 537P, 639P, 746P, 748P,
>1CEZ_A
    DNA binding:     93P, 94P, 95P, 96P, 97P, 98P, 99P, 100P, 101P, 102P, 131P, 133P, 135P, 136P, 139P,
>1S77_D
    DNA binding:     50P, 53P, 54P, 57P, 58P, 160P, 163P, 164P, 182P, 267P, 298P, 300P, 385P, 389P, 392P
    RNA binding:     386P, 390P, 393P, 394P, 397P, 435P, 436P, 437P, 441P, 740P, 741P, 742P, 746P, 748P,
    Ligand binding:  425P, 472P, 537P, 538P, 539P, 540P, 541P, 542P, 569P, 571P, 627P, 631P, 635P, 639P,
>2PI5_A
    DNA binding:     93P, 94P, 95P, 96P, 97P, 98P, 99P, 100P, 101P, 102P, 131P, 133P, 135P, 136P, 139P,
```

```
Binding-site residues based on all structures:
    Protein binding: 303P, 307P, 309P, 310P, 311P, 312P, 717P, 720P, 721P, 723P, 724P, 726P, 728P, 729P,
    DNA binding:     50P, 53P, 54P, 57P, 58P, 93P, 94P, 95P, 96P, 97P, 98P, 99P, 100P, 101P, 102P, 104P,
    RNA binding:     300P, 301P, 386P, 389P, 390P, 393P, 394P, 397P, 425P, 435P, 436P, 437P, 441P, 537P,
    Ligand binding:  125P, 129P, 136P, 137P, 140P, 296P, 308P, 313P, 316P, 317P, 394P, 397P, 398P, 400P,
```

Figure 5
The return form of TCBRP. The upper box shows the binding-site residues mapped on the input protein, 1ARO_P, when different PDB structures are used to calculate binding sites. The lower box shows all the binding-site residues on the input protein by combing all the results.

2PI5. Binding-sites on T7 RNA polymerase are then calculated based on each of these structures and mapped to the sequence of the input chain 1ARO_P (upper box in figure 5). In the end, the TCBRP combines all the results and shows the binding-site residues involved in each type of interaction separately (lower box in figure 5), which include 26 residues in protein-protein binding sites, 112 in DNA-binding, 28 RNA-binding sites, and 54 ligand-binding residues.

Conclusion

Many studies have been conducted on protein functional site prediction and analysis. Calculating binding-site residues on proteins based on the PDB structures has been a necessary and yet painful and time-costly step for these studies. TCBRP has been developed to address this problem with ease. Using TCBRP, users will be able to collect all binding-site residues on proteins of interest very quickly. The developed web server will be useful for the studies on protein interaction and protein functional sites.

Availability and requirements

- **Project name**: A tool for calculating binding-site residues on proteins from PDB structures

- **Project home page**: http://yanbioinformatics.cs.usu.edu:8080/ppbindingsubmit

- **Operating system(s)**: Platform independent

Authors' contributions

CY conceived of the project, designed the architecture, supervised the implementation, and drafted and revised the manuscript. JH performed the coding and attended the discussions. All authors read and approved the final manuscript.

References

1. Reichmann D, Rahat O, Albeck S, Meged R, Dym O, Schreiber G: **From The Cover: The modular architecture of protein-protein binding interfaces.** *Proc Natl Acad Sci USA* 2005, **102(1):**57-62.
2. Zhang Z, Palzkill T: **Dissecting the protein-protein interface between beta-lactamase inhibitory protein and class A beta-lactamases.** *J Biol Chem* 2004, **279(41):**42860-42866.
3. Caffrey D, Somaroo S, Hughes J, Mintseris J, Huang ES: **Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?** *Protein Sci* 2004, **13(1):**190-202.
4. Halperin I, Wolfson H, Nussinov R: **Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking.** *Structure* 2004, **12(6):**1027-1038.
5. Ofran Y, Rost B: **Analysing six types of protein-protein interfaces.** *J Mol Biol* 2003, **325(2):**377-387.
6. Lo Conte L, Chothia C, Janin J: **The atomic structure of protein-protein recognition sites.** *J Mol Biol* 1999, **285:**2177-2198.
7. Jones S, Thornton JM: **Analysis of protein-protein interaction sites using surface patches.** *J Mol Biol* 1997, **272(1):**121-132.
8. Chothia C, Janin J: **Principles of protein-protein recognition.** *Nature* 1975, **256(5520):**705-708.
9. Nooren IM, Thornton JM: **Structural characterisation and functional significance of transient protein-protein interactions.** *J Mol Biol* 2003, **325(5):**991-1016.
10. Keskin O, Mab B, Nussinov R: **Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues.** *J Mol Biol* 2005, **345(5):**1281-1294.
11. Yan C, Dobbs D, Honavar V: **A two-stage classifier for identification of protein-protein interface residues.** *Bioinformatics* 2004, **20(Suppl 1):**i371-i378.
12. Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D, Honavar V: **Identifying amino acid residues involved in protein-DNA interactions from sequence.** *BMC Bioinformatics* 2006, **7:**262.
13. Jones S, Thornton JM: **Prediction of protein-protein interaction sites using patch analysis.** *J Mol Biol* 1997, **272(1):**133-143.
14. Wang L, Brown SJ: **BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences.** *Nucl Acids Res* 2006, **34:**W243-W248.
15. Hwang S, Gou Z, Kuznetsov I: **DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins.** *Bioinformatics* 2007, **23(5):**634-636.
16. Tjong H, Zhou H-X: **DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces.** *Nucl Acids Res* 2007, **35(5):**1465-1477.
17. Ofran Y, Rost B: **ISIS: interaction sites identified from sequence.** *Bioinformatics* 2007, **23(2):**e13-16.
18. Berman HM, Henrick K, Nakamura H: **Announcing the worldwide Protein Data Bank.** *Nat Struct Bio* 2003, **10(12):**980.
19. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, *et al.*: **The Universal Protein Resource (UniProt): an expanding universe of protein information.** *Nucl Acids Res* 2006, **34(suppl_1):**D187-191.