Scientific paper

Binding-sites Prediction Assisting Protein-protein Docking

Janez Konc, Joanna Trykowska Konc, Matej Penca and Dušanka Janežič*

National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

** Corresponding author: E-mail: dusa@cmm.ki.si*

Received: 14-03-2011

Dedicated to Professor Dušan Hadži on the occasion of his 90th birthday

Abstract

Most biological actions of proteins, including their ability to interact with one another, involve some specific parts of their three-dimensional structure, called binding sites. These have evolved for their ability to bind other molecules effectively and are often conserved in different proteins. Identifying protein-protein binding sites in a protein that is known to interact with other proteins can provide important clues to the function of the protein and can also be used in protein-protein docking studies to reduce the search space explored by docking algorithms. We have developed an algorithm for structural similarity search in a database of non-redundant protein structures to find conserved binding regions on proteins involved in protein-protein interactions. We have used this algorithm to find conserved regions on a protein surface. The structurally conserved residues found were labeled as a protein-protein binding site, which allowed us to tune the AutoDock docking algorithm to predict the native protein complex structure from unbound protein structures. The conservation of protein structures that correctly predicted protein-protein binding site was used in AutoDock program to improve protein-protein docking. A web application based on our method is available at http://probis.cmm.ki.si.

Keywords: Binding-sites, prediction, protein structure, docking, structural similarity, probis

1. Introduction

The number of proteins, that are known to interact, is growing fast, but the structures of protein complexes deposited in Protein Data Bank (PDB) are still relatively scarce.[1–3] The experimental methods for obtaining structures of protein complexes are inherently demanding, since interacting proteins may fail to form a complex under the conditions needed for the crystallization. This creates an opportunity for the development of computational approaches, which can both confirm and guide the experiments. Protein-protein docking, one of the most prominent computational approaches for inferring protein-protein interactions, aims to predict the three-dimensional structure of a multimeric protein complex from its constituent protein structures. Success of docking can be significantly improved by pre-knowledge of location of protein-protein binding sites.[4–8] Restricting a docking algorithm, so that it only searches relevant parts of phase space, facilitates in finding the native structure of a protein complex. Since experimental data about protein-protein binding sites are not always available it would be quite rewarding to have computational tools for predicting binding sites on proteins.

The paradigm in computational prediction of protein-protein binding sites is to analyze interfaces of a set of existing protein complexes and to determine parameters which differentiate binding sites from the rest of the protein surface.[9] It is known that certain residues appear more often in interfaces than in the rest of the protein and that certain residues, namely hotspots, which contribute the most to the binding free energy, are conserved.[10–13]

Recently, we developed the algorithm which predicts protein-protein binding sites using conserved protein surface structure and physical-chemical properties in a database of non-redundant protein structures.[14–16] We then implemented this algorithm inside ProBiS, a server for detection of protein binding sites.[17] The algorithm is based on the idea that the most conserved part of the protein surface in terms of the physical-chemical properties must be related either to the binding of small endogenous ligands or of other proteins. To find the conserved part of the protein surface, the algorithm compares the query protein with other proteins, and finds those which share local surface similarities with the input protein. The algorithm was tested in predicting protein-protein binding sites on different sets of known protein complexes, whose structures were obtained from the PDB. Our algorithm was found to detect both protein-protein interfaces, as well as alternative conserved sites on protein surfaces[14–16].

In this paper, we provide proof of concept of the usefulness of our approach for prediction of protein-protein binding sites based on protein surface conservation in docking of unbound protein structures. Using AutoDock docking program[18] combined with our algorithm for predicting protein-protein binding sites enables us to find the complex formed between two proteins from their unbound (monomer) structures. Also we have modified the force field used in AutoDock program, which was previously used mainly for docking of small ligands, by altering Lennard-Jones energy parameters for atoms in the predicted protein-protein binding site[19]. Our approach was then applied to binding site prediction and docking of two unbound protein structures, a transforming growth factor β (TβR-1) and immunophilin FKBP12, which are known to interact[20–21]. Our structural conservation algorithm accurately predicts the binding site on one of the unbound proteins, which is later used together with the modified AutoDock force field to restrain the docking of this unbounded protein. The docking is thus focused only to the residues which make up the protein-protein interface. Using our approach, native protein complex structures are found in 150 docking runs, whereas docking without previously predicted protein-protein binding sites provides no structures of the complex. The best docked protein conformation resulting from our newly developed approach is then compared with the known PDB structure of this protein complex and is found to be in a good agreement with the experimentally determined structure.

2. Experimental

2. 1. Similarity Ssearching Algorithm

In this section, we give a brief outline of our algorithm for prediction of protein-protein binding sites, a detailed description of which is given elsewhere[14–16]. To predict protein-protein binding sites, the algorithm requires that the structure of the query protein and at least one structurally similar protein, i.e., a structural neighbor, are known.

- The algorithm first extracts the solvent accessible surface atoms of these two protein structures. An atom is counted as solvent accessible if its surface is less than 1.1 Å from the surface of a sphere with radius 1.4 Å which is rolled over the atoms[22].
- The atoms of the surface residues for each of the two proteins are then replaced with labeled vertices, so that the physical-chemical properties of the functional groups are preserved. Five labels, hydrogen bond donor (DO), hydrogen bond acceptor (AC), mixed acceptor/donor (ACDO), aromatic (PI), and aliphatic (AL) are used to describe the potential interactions of the functional groups of surface residues[23].
- The distances between each vertex and its neighboring vertices are then calculated and stored as a distance matrix, to facilitate comparison of the two protein surfaces.
- The two proteins represented as labeled vertices are then compared, so that each vertex from protein 1 is compared with each vertex from protein 2. The number of positive matches could be enormous, i.e., for two proteins each of 1000 residues a maximum of million positive matches could be found. The number of positive matches is significantly reduced by comparing distance matrices instead of vertices alone.
- A graph theoretical algorithm for finding a maximum clique, which takes the set of matched vertices as an input, then determines the maximum similarity substructure between the two protein surfaces[24, 25].

2. 2. Protein Structures

The Database of Interacting Proteins was queried for proteins involved in protein-protein interactions[1]. We have chosen only interactions for which structures of both protein partners in an unbound form as well as the structure of the protein complex are available. We have selected two unbound interacting proteins, a transforming growth factor β (TβR-1) and immunophilin FKBP12, with their respective PDB codes *1ias* and *1d6o*. The corresponding structure of their protein complex is found under PDB code *1b6c*, where chain A represents bound form of FKBP12 and chain B bound form of TβR-1 protein. The proteins that were found to be structurally similar to the unbound FKBP12 protein were (their PDB codes): *1ix5, 1jvw, 1pbk, 1q6h, 1r9h, 1u79, 2awg, 2d9f, 2if4, 2ofn, 2pbc, 2uz5, and 3b7x,* and for unbound TβR-1 (their PDB codes): *1ckj, 1kob, 1m17, 1o6k, 1o9u, 1u59, 1wak, 1yhv, 1yvj, 2b7a, 2bfy, 2csn, 2f4j, 2ivt, 2izs, 2j0l, 2jbo, 2pzy, 2qkw, 2qlu, 2qr7, 2v7o, 3bkb,* and *3lck.*

2. 3. Prediction of Binding Sites

A non-redundant Protein Data Bank was queried with the polypeptide chain for which the prediction was performed using the ProBiS server[17]. We used strict parameters to avoid excessively dissimilar structures, use of which could lead to biased predictions. All of the protein structures in this list were structurally aligned and share local structure similarities with query protein. The process of comparing the query protein structure sequentially with the other proteins is shown in Figure 1. In each step of the algorithm, the detected conserved surface regions were mapped over the previous ones[14–16]. Each residue on the surface of a query protein was assigned a conservation score, which counts the number of times this residue is conserved in the predicted structurally similar proteins.

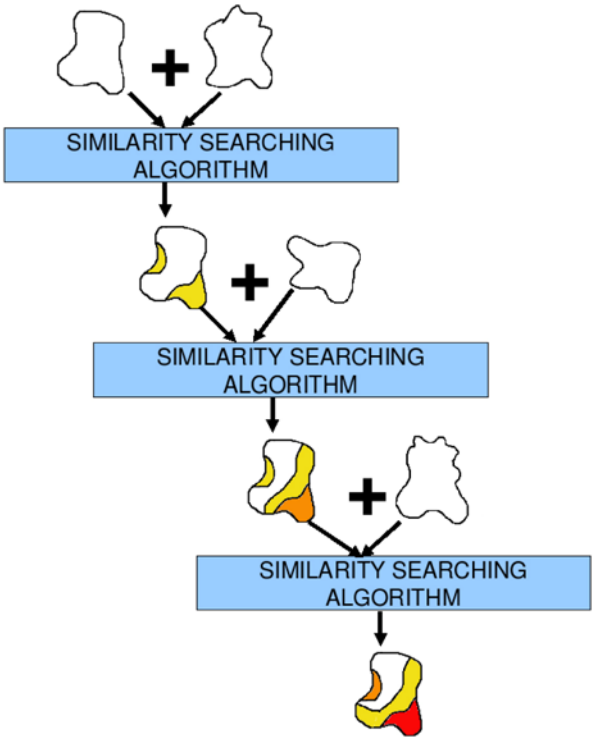

SIMILARITY SEARCHING ALGORITHM
SIMILARITY SEARCHING ALGORITHM
SIMILARITY SEARCHING ALGORITHM

Figure 1. Protein-protein binding sites prediction procedure. Structures of the non-redundant PDB proteins are sequentially compared with the query protein. The found similarities are mapped on the surface of the query protein.

2. 4. Docking Protocol

AutoDock 4.0 was used for the docking simulation. We employed the Lamarckian genetic algorithm (LGA) for ligand conformational searching, which is a hybrid of a genetic algorithm and a local search algorithm. This algorithm first builds a population of individuals (genes), each being a different random conformation of the docked molecule. Each individual is then mutated to acquire a slightly different translation and rotation and the local search algorithm then performs energy minimizations on a user-specified proportion of the population of individuals. The individuals with the low resulting energy are then transferred to the next generation and the process is repeated. The algorithm is called Lamarckian because every new generation of individuals is allowed to inherit the local search adaptations of their parents.

The preparation of the target protein TβR-1 (unbound target) with the AutoDockTools software involved adding all hydrogen atoms to the macromolecule, which is a step necessary for correct calculation of partial atomic charges. Gasteiger charges are calculated for each atom of the macromolecule in AutoDock 4.0 instead of Kollman charges which were used in the previous versions of this program. Three-dimensional affinity grids of size 277 × 277 × 277 Å with 0.6 Å spacing were centered on the geometric center of the target protein and were calculated for each of the following atom types: HD, C, A, N, OA, and SA, representing all possible atom types in a protein. Additionally, an electrostatic map and a desolvation map were calculated.

We set important docking parameters for the LGA as follows: population size of 150 individuals, 2.5 million energy evaluations, maximum of 27000 generations, number of top individuals to automatically survive to next generation of 1, mutation rate of 0.02, crossover rate of 0.8, 150 docking runs, and random initial positions and conformations. The probability of performing local search on an individual in the population was set to 0.06 and the maximum number of iterations per local search was set to 300. Unbound target TβR-1 and unbound ligand FKBP12 proteins were both treated as rigid. The docking jobs were distributed to the CROW Linux cluster[26], each producing 150 docked conformations and the calculations were completed in less than 2 days.

2. 5. Modified Force Field

We modified the force field used by the AutoDock docking program, so that the atoms belonging to the predicted binding site would feel stronger attraction to other atoms by enlarging the well depth parameters for those atoms[27]. The well depth of the energy function for each interaction between two atoms is defined by the Lennard-Jones parameter epsilon. The higher value of this parameter means lower well depth and stronger interaction. We take the modified values of this epsilon parameter for atom types found in proteins (HD, C, A, N, OA, and SA) to be 2, 5, and 10 times higher as the standard values.

2. 6. Docking Success

The success of docking was measured by comparing the locations of predicted binding site residues in docked conformations of ligand with the corresponding bounded ones from the crystal structure of the complex. We calculated the RMSD between the backbone atoms of the predicted binding site residues on unbound and bound structures, after superimposing the structure of known complex to the predicted complex.

To estimate the advantage of using the modified over the unmodified force field we determined the clustering of docked conformations. The premise was when a global minimum for the native conformation exists then the docked conformations would cluster towards this structure. A docking success can therefore be measured by the degree of clustering of docked conformations. We calculated the geometric centers of docked conformations, i.e., for each run 150 centers, and counted the number of neighboring centers in a 6 Å radius sphere around each geometric center. The number of neighboring geometric centers was then used as a measure for clustering of the docked conformations.

3. Results and Discussion

For protein-protein binding site prediction and docking, we selected two unbound protein structures, a transforming growth factor β (TβR-1) and immunophilin FKBP12. The corresponding protein complex structure (PDB code *1b6c*, chains A and B) was used for validation of our results. The structure of unbound FKBP12 protein in our similarity searching algorithm represents a query protein and in docking can be viewed as ligand. The structure of unbound TβR-1 is our target protein.

3. 1. Binding Sites Prediction

We predicted the binding site for the protein FKBP12 (unbound form) by searching for conservation in its structural neighbors. We found 12 structurally similar proteins for this query protein through the ProBiS web server[17]. All structural alignments were obtained by superimposing conserved patches of the protein surfaces (considering predicted binding site residues only) and not global structural alignments of proteins (considering all residues).

We checked for differences between the bound and unbound protein structures, to test if any conformational changes occur in proteins upon binding, since these differences affect docking. These differences were measured by the RMSD between the unbound target protein and its counterpart from the protein complex. The calculated value of the RMSD between the backbone atoms of the aligned target protein in bound and unbound form was 1.951 Å. We were also interested in the RMSD between binding site regions of these two proteins (25 residues) which was found to be 2.875 Å. This suggests that the interfacial residues undergo conformational changes during binding. The RMSD difference between the ligand in bound and unbound form was 0.349 Å suggesting only minor conformational changes during binding.

The conserved residues of unbound FKBP12 protein that were found by structural similarity search and residues belonging to the actual binding site from the bound FKBP12 are shown in Figure 2. We observe that the conserved residues strongly coincide with the actual protein-protein binding site. Two residues, Tyr 82 and Gly 83, are found to be conserved in 8 out of 13 structural neighbors (red in Figure 2, panel (a)) and are the most conserved part of the FKBP12 protein surface. They are located close to the center of the binding site region and when bound, Tyr 82 is completely buried in the interface.

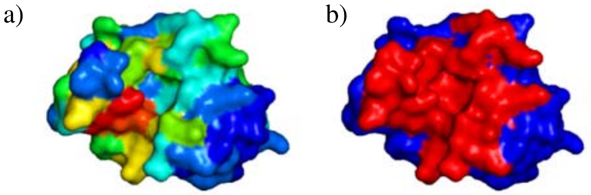
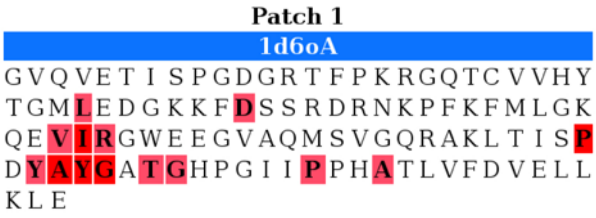


Figure 2. Surface maps on unbound FKBP12 protein: (a) predicted binding site (8 times conserved residues are in red; 7 times in orange; 6 times in light orange; 5 times in yellow; 4 times in light green; 3 times in green; 2 times in cyan; 1 times in light blue; not conserved are colored blue); (b) actual binding site (binding site residues are in red; the rest of the surface is in blue).

As a possible protein-protein binding site we considered residues with structural similarity scores > 0.7 (see Figure 2). These predicted binding site residues are shown in Figure 3. It can be seen that they are located at different parts of FKBP12 protein sequence, even though they are close together in the protein structure. It should be emphasized that these conserved residues are found by structural similarity search in protein structures and not by alignments of proteins sequences.



Patch 1

1d6oA

GVQVETISPGDGRTFPKRGQTCVVHY
TGM**L**EDGKKF**D**SSRDRNKPFKFMLGK
QE**VIR**GWEEGVAQMSVGQRAKLTIS**P**
D**YAYG**A**TG**HPGII**P**PH**A**TLVFDVELL
KLE

Figure 3. Sequence of FKBP12 protein and predicted binding site as displayed on our web page (http://probis.cmm.ki.si). Residues with high similarity scores are colored in different shades of red.

The unbound target protein (TβR-1) was also used to search for conserved regions of its surface. We find conserved large part of GS region, where the protein-protein binding site for FKBP12 is located, as well as the catalytic segment. However, the predictions were not used in docking, primarily because the predicted binding site surface was not a continuous surface patch. Our similarity searching algorithm could not distinguish between different conserved regions in this case.

3. 2. Docking

The AutoDock 4.0 docking program was employed for the docking of protein FKBP12 (unbounded ligand) to the protein TβR-1 (unbounded target). This docking program uses an all atom representation of protein structures and is therefore computationally demanding. The docking to the unbound target protein was performed with and without (blind docking) the previous knowledge of the location of the protein-protein binding site on unbounded ligand. To validate the modified force field effects on docking we observed position of the binding site residues relative to the target protein surface. The stronger force field affected the docking so that on average predicted binding site residues were docked closer to the target protein surface as shown in Table 1. It can be seen that five time larger modified force field parameters in docking program gave the highest number of best docked structures.

Table 1. Summary of docked conformations.

Force field	Orientation[a]	Best docked [b]	Clustering[c]
standard	2.5	2	0
modified (2x)	4.2	1	0
modified (5x)	3.1	9	4
modified (10x)	2.9	7	6

[a] Average number of predicted binding site residues < 10 Å from the surface of the target protein for all 150 docked conformations.
[b] Number of docked conformations with < 20 Å RMSD between predicted and known binding sites.
[c] Number of best docked conformations < 10 Å from the most populated region of their geometric centers.

The average number of predicted binding site residues less than 10 Å from the surface of the target protein was 2.5 for the unmodified and 3.1 for the modified force field. The docking performed with modified force field resulted in improved clustering of docked conformations, while in blind docking, docked conformations were almost evenly dispersed, which is shown in Figure 4.

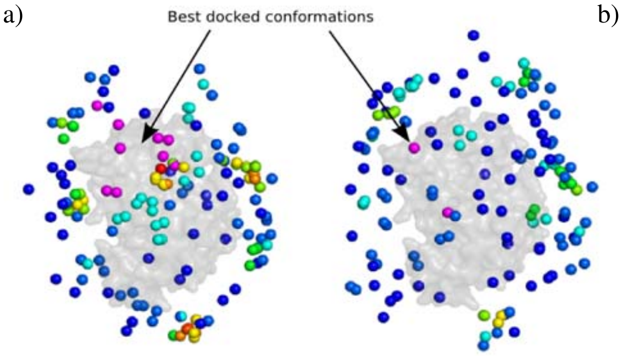


Figure 4. Clustering of docked conformations: (a) with the preknowledge of protein-protein binding site; (b) blind docking. Most populated clusters are colored red and least populated are colored blue. Best docked conformations on both panels are shown in magenta. Target protein in surface representation is colored grey.

The modified force field (5x) found 9 docked conformations with RMSD less than 20 Å between their predicted binding site residues and actual binding site residues (best docked conformations), while blind docking found only 2 best docked conformations. Additionally, docked structures found with modified force field were clustered near the most populated region of docked conformations (geometric centers); while in blind docking we observed no such clustering.

The clustering of best docked conformations near the most populated region of docked conformations is preferable. This suggests that the docking algorithm was able to converge towards the native structure of the protein complex. In Figure 5 are shown the best conformation of unbound ligand (red) docked to target protein (green) and the known crystal structure of the protein complex (chain A, blue; chain B, slate). The best RMSD between the docked unbound ligand and the bound ligand from the crystal structure is 11.1 Å.

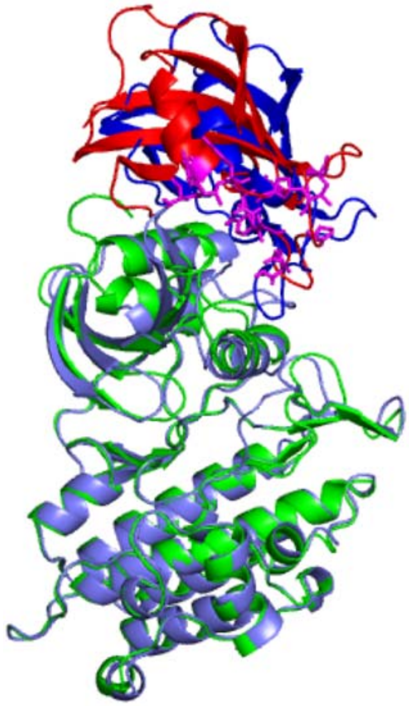


Figure 5. Docked structures: the best docked conformation of unbound ligand FKBP12 with modified force field parameters is colored red. The unbound target protein TβR-1 is colored green. The predicted binding site residues of docked ligand are colored magenta. The native conformations of FKBP12 and TβR-1 from crystal structure of the protein complex are colored blue and slate, respectively. All figures were rendered using PyMOL (http://pymol.sourceforge.net).

4. Conclusions

We describe a new approach for prediction of protein-protein binding sites and apply it to docking of unbound protein structures. Binding sites on a protein are found by searching the database of non-redundant protein structures for the most conserved surface patch of surface on the query protein. The AutoDock 4.0 docking program force field is modified to focus the docking on the predicted binding site. Comparison with blind docking (unmodified force field) reveals that being acquainted with the predicted protein-protein binding site significantly improves clustering of docked conformations around the native conformation.

Our approach for prediction of protein-protein binding sites using structural conservation of protein surfaces is proved to be useful in protein-protein docking, in particular if the interface residues are not experimentally determined.

5. Acknowledgement

The financial support through grants P1-0002 and Z1-3666 of the Ministry of Higher Education, Science, and Technology of Slovenia and the Slovenian Research Agency are acknowledged.

6. References

1. L. Salwinski., C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, D. Eisenberg, *Nucleic Acids Res.* **2004**, *32*, 449–451.
2. G. D. Bader, I. Donaldson, C. Wolting, B. F. F. Ouellette, T. Pawson, C. W. V. Hogue, *Nucleic Acids Res*. **2001**, *29*, 242–245.
3. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res*. **2000**, *28*, 235–242.
4. G. R. Smith, M. J. Sternberg, *Curr. Opin. Struct. Biol*. **2002**, *12*, 28–35.
5. A. Fahmy, G. Wagner, *J. Am. Chem. Soc*. **2002**, *124*, 1241–1250.
6. R. Mendez, R. Leplae, L. De Maria, S. J. Wodak, *Proteins* **2003**, *52*, 51–67.
7. S. J. de Vries, A. D. J. van Dijk, A. M. J. J. Bonvin, *Proteins* **2006**, *63*, 479–489.
8. D. Motiejunas, R. Gabdoulline, T. Wang, A. Feldman-Salit, T. Johann, P. J. Winn, R. C. Wade, *Proteins* **2008**, *71*, 1955–1969.
9. S. Jones, J. M. Thornton, *J. Mol. Biol*. **1997**, *272*, 133–143.
10. B. C. Cunningham, J. A. Wells, *Proc. Natl Acad. Sci. USA* **1991**, *88*, 3407–3411.
11. T. Clackson, J. A. Wells, *Science* **1995**, *267*, 383–386.
12. B. Ma, T. Elkayam, H. Wolfson, R. Nussinov, *Proc. Natl Acad. Sci. USA* **2003**, *100*, 5772–5777.
13. O. Keskin, B. Ma, R. Nussinov, *J. Mol. Biol*. **2005**, *345*, 1281–1294.
14. J. Konc, D. Janezic, *J. Chem. Info. Mod*. **2007**, *47*, 940–944.
15. N. Carl, J. Konc, D. Janezic, *J. Chem. Info. Mod*. **2008**, *48*, 1279–1286.
16. J. Konc, D. Janezic, *Bioinformatics* **2010**, *26*, 1160–1168.
17. J. Konc, D. Janezic, *Nucleic Acids Res.* **2010**, *38*, W436–W440.
18. G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson *J. Comp. Chem.* **1998**, *19*, 1639–1662.
19. R. Huey, G. M. Morris, A. J. Olson, D. S. Goodsell, *J. Comp. Chem.* **2007**, *28*, 1145–1152.
20. P. Burkhard, P. Taylor, M. D. Walkinshaw, *J.Mol.Biol*. **2000**, *295*, 953–962.
21. M. Huse, T. W. Muir, L. Xu, Y. G. Chen, J. Kuriyan, J. Massague, *Mol.Cell* **2001**, *8*, 671–682.
22. J. Konc, M. Hodoscek, D. Janezic, *Croat. Chem. Acta* **2006**, *79*, 237–241.
23. S. Schmitt, D. Kuhn, G. Klebe, *J. Mol. Biol*. **2002**, *323*, 387–406.
24. J. Konc, D. Janezic, *MATCH Commun. Math. Comput. Chem.* **2007**, *58,* 569–590.
25. J. Konc, D. Janezic, *Lect. Notes. Comp. Sci*. **2007**, *4432*, 399–406.
26. M. Hodoscek, U. Borstnik, D. Janezic, *Cell. Mol. Biol. Lett*. **2002**, *7*, 118–119.
27. A. M. Ferrari, B. Q. Wei, L. Costantino, B. K. Shoichet, *J. Med. Chem.* **2004**, *47*, 5067–5084.

Povzetek

Biološka funkcija proteinov, med drugim sposobnost tvorjenja medsebojnih interakcij, je odvisna predvsem od tridimenzionalne strukture njihovih vezavnih mest. Le-ta so se skozi evolucijo razvila tako, da učinkovito vežejo molekule, in so pogosto evolucijsko ohranjena. Identifikacija proteinskih vezavnih mest na proteinu za katerega vemo, da vstopa v interakcije z drugimi proteini, pomagajo predvsem pri odkrivanju njegove funkcije, lahko pa jih uporabimo tudi za izboljšanje proteinskega sidranja. Pri zadnjem pristopu poznavanje vezavnih mest zmanjša iskalni prostor algoritma za sidranje, s čimer se pospeši iskanje nativne konformacije proteinskega kompleksa. Razvili smo algoritem za iskanje lokalnih strukturnih podobnosti med proteinom in bazo neredundančnih proteinskih struktur, z namenom iskanja ohranjenih vezavnih mest na proteinih. Ta algoritem smo uporabili za odkrivanje ohranjenih mest na površini proteinskih struktur, ki smo jih označili kot proteinska vezavna mesta. Nato smo spremenili AutoDock program tako, da se pri sidranju osredotoča samo na napovedana proteinska vezavna mesta, s čimer smo pohitrili in izboljšali postopek proteinskega sidranja. Spletni program, ki to omogoča, je na voljo na http://probis.cmm.ki.si.