Protein−Protein Binding Site Prediction by Local Structural Alignment

Nejc Carl,[†] Janez Konc,[†] Blaž Vehar,[†] and Dušanka Janežič*[,†,‡]

National Institute of Chemistry, Hajdrihova 19, SI-1000, Ljubljana, Slovenia, and Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Glagoljaška 8, SI-6000 Koper, Slovenia

Received July 9, 2010

Generalization of an earlier algorithm has led to the development of new local structural alignment algorithms for prediction of protein−protein binding sites. The algorithms use maximum cliques on protein graphs to define structurally similar protein regions. The search for structural neighbors in the new algorithms has been extended to all the proteins in the PDB and the query protein is compared to more than 60 000 proteins or over 300 000 single-chain structures. The resulting structural similarities are combined and used to predict the protein binding sites. This study shows that the location of protein binding sites can be predicted by comparing only local structural similarities irrespective of general protein folds.

1. INTRODUCTION

Protein binding sites are specific parts of protein structure involved in interactions with other proteins, DNA, or small ligands, and a knowledge of protein interactions is essential for a complete understanding of protein behavior. Elucidation of the details of protein interactions,[1] a goal of computational chemistry, will shed light on the functioning of proteins. Experimental methods for determining protein interactions exist, but they typically fail to identify the specific parts of the protein surface where these interactions occur and they are replete with false positives. Continuing efforts in genomics provide an increasing number of protein sequences from a large variety of organisms; at the same time, structural biology initiatives contribute structures to the Protein Data Bank (PDB),[2] and many computational methods have been developed for automated prediction of binding sites on protein surfaces. A knowledge of specific areas on the protein surface where protein−protein interactions occur is however vital to an understanding of interaction mechanisms and to rational drug design. We have sought this information by computational means.

Most computational methods for determining protein−protein binding sites are based upon one of two approaches.[3] One approach calls for the design of a physically relevant parametric function that characterizes, linearly or nonlinearly, the properties of proteins.[4−8] The other approach optimally combines different properties from a large number of parameters by means of a machine learning algorithms such as neural networks,[9−11] support vector machines,[12,13] and Bayesian networks.[14,15] The properties that appear to promise some predictive power with respect to protein−protein interfaces fall into three distinct groups. The first of these deals with the type and properties of amino acid residues, such as desolvation, interface propensity, hydrogen-bonding potential, and hydrophobicity.[16,17] The second group is based on the degree of evolutionary conservation, which can be assessed by comparing a given sequence to those of homologous proteins.[18,19] Even proteins with similar sequences but different functions can evolve from a common precursor through evolution of promiscuous protein functions.[20] The third group is comprised of protein properties such as surface accessibility, secondary structure, tertiary structure (i.e., structural neighbors as proteins with similar overall folds), and properties that describe the chemical composition and the shape of the protein, derivable from the atomic coordinates of the structure.[13,21−24] Algorithms must combine these three distinct groups of properties and formulate a prediction of a protein−protein interface. The output of protein−protein binding site detecting algorithms consists of surface patches of amino acid residues that can form a contiguous interface surface or a combination of interfaces through which binding to other proteins can occur.

We have previously described[21] an algorithm that detects binding sites by comparing physicochemical properties of structurally similar protein surfaces. The algorithm searched for structural similarities in a small number of structurally similar proteins obtained with a vector alignment search tool (VAST).[25] VAST first identified elementary fragment-pair similarities consisting of pairs of aligned helices and β strands, where the unit of structural similarity consisted of a pair of secondary structure elements whose type and relative orientation were similar within a specified tolerance. Proteins that were found to be sufficiently similar, termed structural neighbors, were compared to the query protein, and the structural similarities with the query protein were combined to generate predictions. However, when no structural neighbors are found, prediction of protein binding sites cannot be performed with this approach. A recently developed algorithm[26,27] for protein binding sites prediction that also uses a maximum clique algorithm adopts a different approach from the one presented here to combine local structural similarities into a prediction. It compares the query protein against a curated database of proteins, in which highly

* Author to whom the correspondence should be addressed. E-mail: dusa@cmm.ki.si.
[†] National Institute of Chemistry.
[‡] University of Primorska.

10.1021/ci100265x
Published on Web 10/04/2010

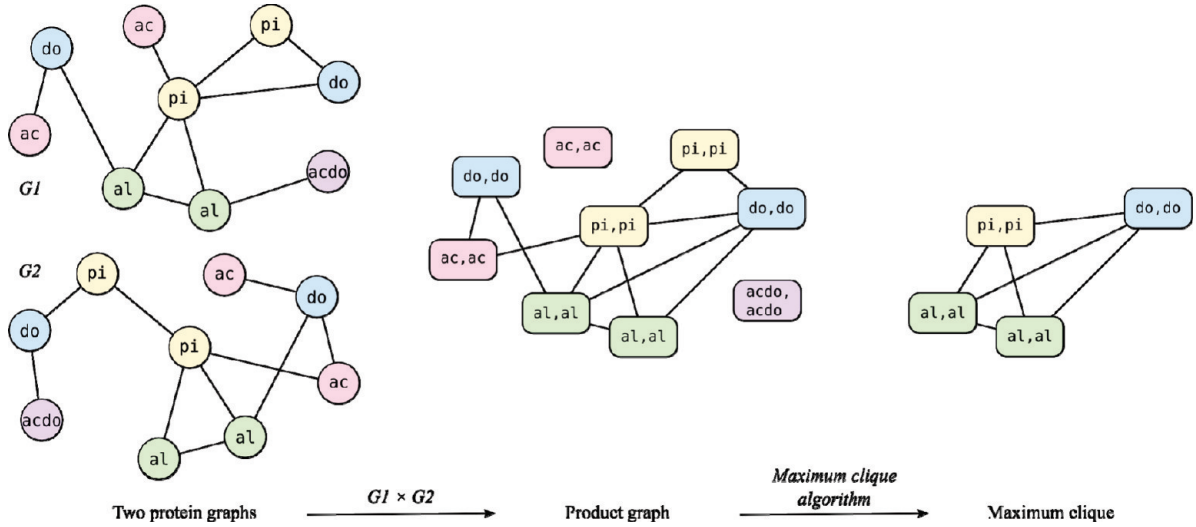

Figure 1. The process of finding a maximum clique in a product graph that was constructed from two protein subgraphs, G1 and G2. Notations for different functional groups are ac, do, acdo, al, and pi for hydrogen bond acceptor, hydrogen bond donor, mixed acceptor/donor, aliphatic, and aromatic functional group, respectively. The clique with the largest number of vertices in the product graph is detected. This corresponds to the largest common substructure in the two protein subgraphs.

sequence-identical chains have been removed, only a single representative of each sequence identity being retained.

In this paper, we present a generalization of our earlier algorithm[21] for protein−protein binding site predictions. A newly developed local structure alignment (LSA) algorithm detects locally similar surface patches of proteins independently of the different folds present in the protein and combines them to predict protein−protein binding sites. The query protein is compared to more than 60 000 protein entries from the Protein Data Bank (PDB), which was downloaded on March 6, 2010. The LSA algorithm does not require the removal of proteins with similar sequences; the query structure is compared to many proteins with identical or similar sequences. Such structures may, despite their virtually identical sequences, adopt different 3D orientations of the functional groups of their surface amino acid residues, and the local comparison to such proteins can lead to different local alignments. Since the query structure may adopt a conformation not common in the PDB database, such an overcomparison is an advantage in these cases. The LSA algorithm is able to detect structural similarity within proteins with similar sequences and utilize these similarities to detect protein−protein binding sites. It is fold-independent and allows for the detection of binding sites on proteins independently of their structural neighbors, even in cases where no structural neighbors can be found. It exploits a new approach combining the detected structural similarities into a binding sites prediction and utilizing different conformational states of sequence identical comparison proteins to predict protein−protein binding sites.

2. METHODS

Our LSA algorithm, an extension of a previous algorithm for prediction of protein−protein binding sites,[21] compares individual protein surfaces using a graph theoretical approach,[28,29] by constructing protein graphs. The algorithm first extracts the solvent-accessible atoms defined by rolling a sphere over the atoms of a protein surface.[30] The physicochemical properties of the functional groups of surface amino acid residues are then encoded with one of five labels: hydrogen-bond donor, hydrogen-bond acceptor, mixed acceptor/donor, aromatic, and aliphatic group. Each functional group is then substituted with one labeled point that represents potential interactions of this particular functional group with other molecules (e.g., proteins, ligands, or solvent).[31] Once a physicochemical representation of the protein surface is established in this way, different surfaces are compared by constructing protein product graphs, and this is followed by a search for maximum cliques.

A graph G is a pair of sets (V, E) such that $E \subseteq [V]^2$; thus, the elements of E are two-element subsets of V. The elements of V are the vertices of the graph G and the elements of E are its edges. The product graph is a compact representation of similarities between two graphs. The vertices of the product graph are pairs of points, one from the first protein graph and the other from the second protein graph, where the two vertices in a pair must have the same label and must be surrounded by similar vertices at similar distances. An edge is drawn between those two product graph vertices, where the distance between the first two points is equal, with some tolerance, to the distance between the second two points. A product graph is constructed for each pair of labeled points with similar surroundings, and a maximum clique is found in each resulting product graph. A clique is a fully connected subgraph, such that $V^c \subseteq V$ and $E^c \subseteq E$ and that $\forall[V^c]^2 \in E^c$, where V^c and E^c are the vertex and the edge set of the clique (Figure 1). A maximum clique in a product graph corresponds to a similar patch in the two proteins compared. Cliques, rather than some less complete clusters of vertices, are used because each element of a clique is connected to every other element and this leads to a greater similarity in space orientation of functional groups.[32]

The maximum cliques in product graphs correspond to protein surface similarities. The smallest cliques considered are of size 4 and their size can range to over 20. The process of transforming two protein graphs into a maximum clique is shown in Figure 1.

A decision was necessary concerning which proteins to use in maximum clique searches. In our previous algorithm,

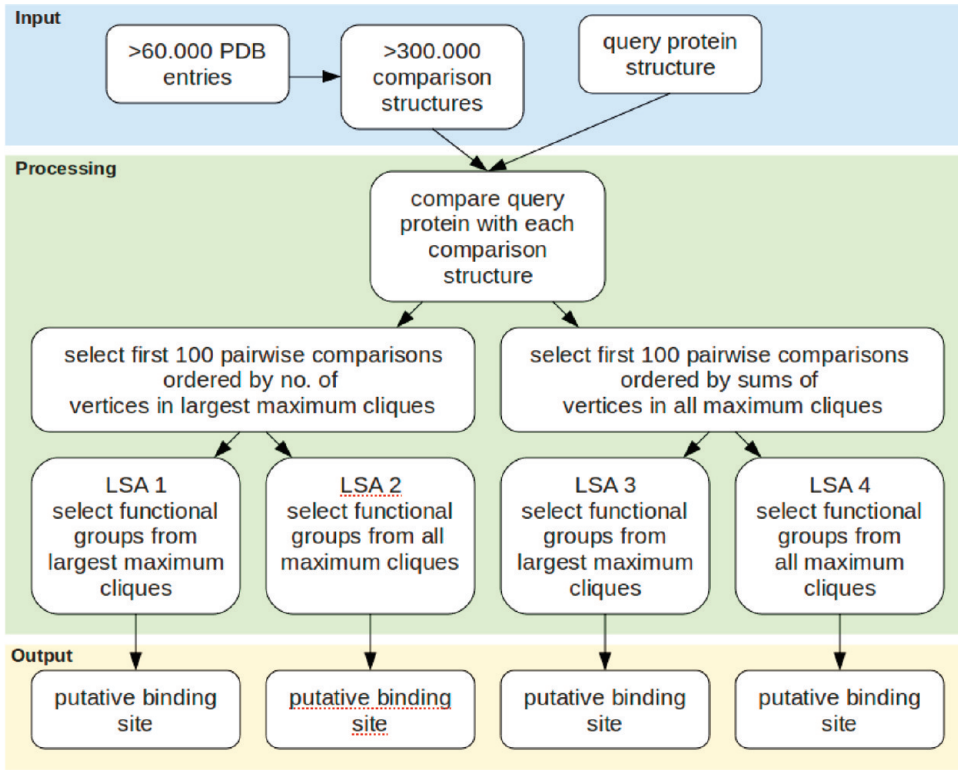

Figure 2. The process of binding site prediction on the query protein. Four variants of the local structural alignment algorithm are presented. The preferred approach is underlined in red.

the comparison proteins were chosen with the VAST server,[25] which provided proteins with similarly aligned secondary structure elements, i.e., proteins with similar folds. These proteins were used to search for local structural similarities, each of them was compared to the query protein, and local structure similarities were mapped to the surface of the query protein. In the new LSA approach, the number of structural neighbors is not limited to those obtained by the VAST routine. The search for structural neighbors is extended to all the proteins in the PDB and the query protein is compared to all entries in a database of protein structures. The protein similarities that are found are combined to develop a prediction of the binding site.

At the advanced search interface of the PDB Web site (http://www.pdb.org/pdb/search/advSearch.do) all structures with macromolecular type "protein" were chosen, while the presence of nucleic acids was ignored. Since our algorithm is structurally sensitive and is able to detect different local similarities for different conformations of the same protein, even structures that differ only in local arrangements of amino acid functional groups can provide some additional information about the way that protein structure has to be arranged locally in order to facilitate binding. As a consequence, structures with similar sequences were not removed, and the resulting search provided more than 60 000 PDB entries, which were downloaded. The set of 60 000 protein structures was split into different model structures; NMR, for example, typically provides many conformations of the same protein. The model structures were further split into individual protein chains, and in this way, >300 000 single-chain comparison structures were extracted.

To execute a local similarity search, a query protein chain is selected and a pairwise comparison performed with each of the 300 000 comparison structures. Each of these pairwise comparisons may result in a number of different maximum cliques, i.e., local structural similarities between the two proteins being compared. A maximum clique is a set of pairs of vertices, where one vertex in each pair is from the query protein and the other from the comparison structure; the cluster of vertices that originates from the query and the one from the comparison structure are in a similar 3D arrangement and match in their physicochemical properties. For each pairwise comparison, a new file is created and all maximum cliques that resulted from this pairwise comparison are saved into this file. The vertices of a maximum clique, represented by their physicochemical labels and their positions in 3D space, as well as the names and the numbers of their parent amino acid residues, are recorded. Two additional values are calculated and saved in each file: the size of the largest maximum clique and the sum of sizes of all different maximum cliques found by a pairwise comparison, where the size of a maximum clique is defined as the number of its vertices.

Once all pairwise comparisons are completed, they are sorted so that those richer in information, those containing the larger structural similarities, will be put forward. For each pairwise comparison, the list of maximum cliques obtained is sorted in decreasing order of their size. The goal is to find an ordering of the pairwise comparisons and an ordering of the maximum cliques that will produce the best binding site predictions after combining the information they contain. Four variants of LSA algorithms were designed, each being a combination of two consecutive binary decisions (see Figure 2): (1) to select either the pairwise comparisons that yield the largest maximum cliques (LSA 1, LSA 2) or the ones that yield the largest sums of all maximum clique sizes (LSA 3, LSA 4) and (2) to make binding site predictions containing the functional groups (vertices) from either only

the largest maximum cliques (LSA 1, LSA 3) or the ones from all maximum cliques (LSA 2, LSA 4).

It was decided arbitrarily that N top ranked pairwise comparisons ($1 \leq N \leq 100$) would be used in the prediction of binding sites. In each of the four LSA algorithms, the top N pairwise comparisons are selected according to the first binary decision, i.e., either by the sizes of their largest maximum cliques or by the sums of sizes of all maximum cliques. Within results obtained from these top N selected pairwise comparisons, the occurrences of surface residues of the query protein are counted according to the second binary decision, i.e., either in only the largest maximum cliques or in all maximum cliques that were generated. The most frequently occurring residues were then flagged as being a part of the predicted binding site.

LSA 1. In this approach, pairwise comparisons are selected by the size of the largest maximum cliques, and surface residues are counted in only the largest maximum cliques. As a comparison of two proteins often yields many maximum cliques of different sizes that represent different similar surface patches of residues, this approach is best suited to detect binding sites that are large, contiguous, and rigid. Since only the largest maximum clique from each comparison is considered, no other alternative smaller conserved sites can be detected with this approach.

LSA 2. In this approach, pairwise comparisons are selected by the size of the largest maximum cliques, and surface residues are counted in all maximum cliques. In this way each maximum clique can contribute to the similarity found, and alternative smaller similarities are also detected. This approach is best suited to detect a large contiguous patch of conserved residues that provides a structural scaffold to support the binding site. Additionally, smaller binding site patches or noncontiguous binding sites can be detected with this approach.

LSA 3. In this approach, pairwise comparisons are selected by sums of sizes of all maximum cliques, and surface residues are counted in only the largest maximum cliques. In this way, proteins that share many small similarities with the query protein can surface at the top of the list and contribute to the predicted binding site. This approach is able to detect noncontiguous binding sites, where the two proteins forming a complex interact through two or more separate interfaces.

LSA 4. In this approach, pairwise comparisons are selected by sums of sizes of all maximum cliques, and surface residues are counted in all maximum cliques. This approach can detect binding sites that are structurally flexible, as both the process of selecting proteins for comparisons and the process of combining amino acids into a prediction include searching for the largest number of small similarities that may be scattered across the protein surface.

3. RESULTS

The binding site predictions were tested with the set of 24 biologically relevant protein−protein complexes that had been used previously.[21] The complexes were split into single protein chains, and one chain from each complex was examined in order to detect the protein−protein binding site through which it interacts with the other chain(s). The true binding sites for each of the chosen 24 complexes were

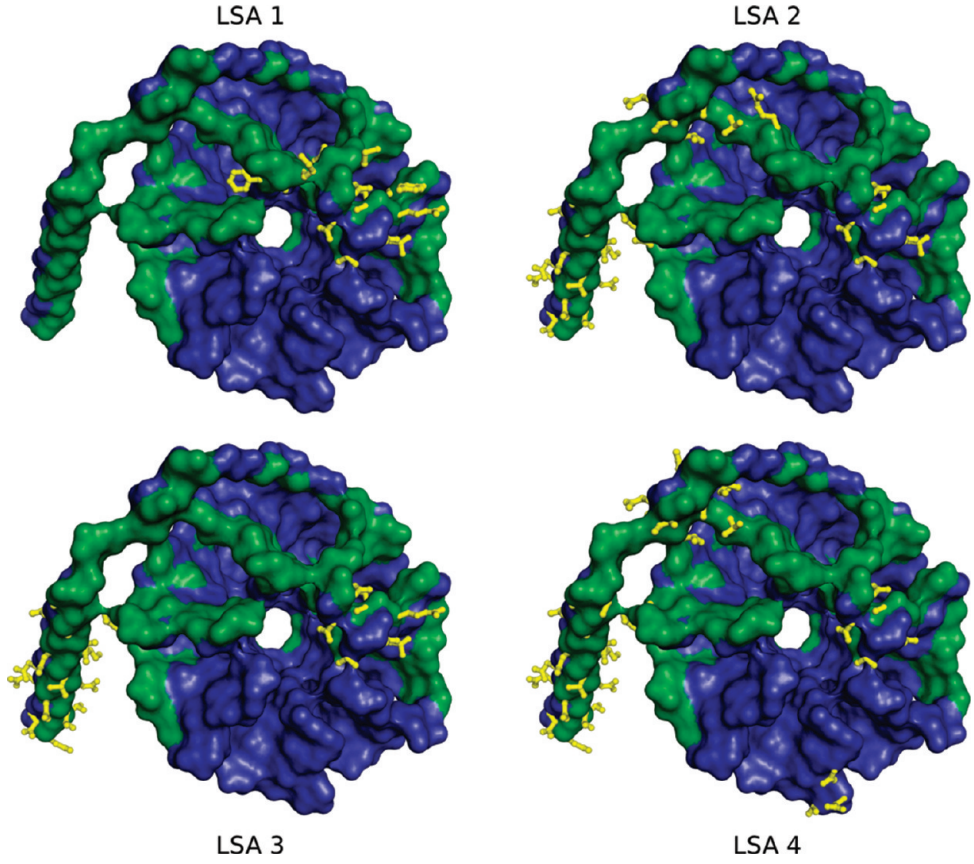

Figure 3. Differences in binding site predictions of four LSA approaches, demonstrated on the β_1-subunit of the signal-transducing G protein heterotrimer (PDB code 1got, chain B). The protein subunit is shown as the surface of its backbone (in blue and green). The actual binding site is colored green. Residues from predicted binding sites are depicted as yellow sticks (side chains only).

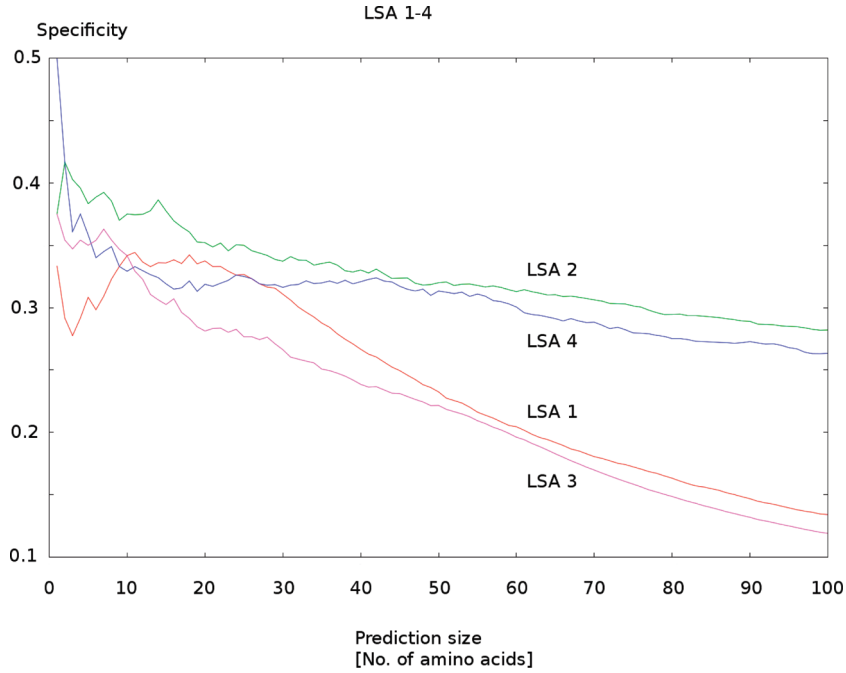

Figure 4. Average specificities for 24 test set proteins when the top 15 pairwise comparisons are considered. Four different LSA approaches are shown, with LSA 1 in red, LSA 2 in green, LSA 3 in magenta, and LSA 4 in blue.

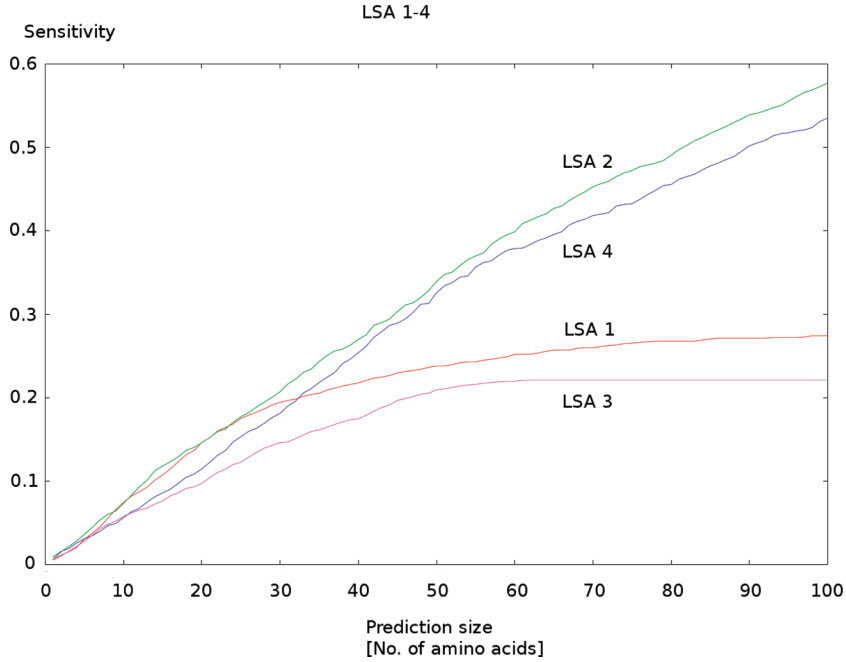

Figure 5. Average sensitivities for 24 test set proteins when the top 15 pairwise comparisons are considered. Four different LSA approaches are shown, with LSA 1 in red, LSA 2 in green, LSA 3 in magenta, and LSA 4 in blue.

identified from the data in the PDB using the arbitrary rule that a residue on the query protein chain is considered to be a part of the actual binding site if any of its atoms is <3 Å from any atom of the interacting protein chain(s).[21]

Each of the 24 single-chain query structures was used to search the database of >300 000 single-chain structures, and the binding site predictive capability of each of the four LSA approaches was examined. The LSA 1 approach detects the largest contiguous patches of conserved amino acid residues on the query protein chain, located at a relatively rigid part of its surface. No smaller similarities can contribute to this prediction. The LSA 2 and LSA 4 approaches are both designed to detect smaller conserved patches of surface residues along with the large ones, but they differ in the ordering of the pairwise comparisons (for details see Figure 2). The LSA 4 approach can detect both rigid and flexible conserved residues. The LSA 3 predictions showed the worst correlation with the actual binding sites of all the tested approaches on average. An example of the binding site predictions is given in Figure 3, generated with PyMol,[33] which shows the binding sites predicted by each LSA approach for the β_1-subunit of the signal-transducing G protein heterotrimer (PDB code 1got, chain B). In this figure, the surface of the protein subunit is blue and the crystallographically defined binding site within it is colored green.

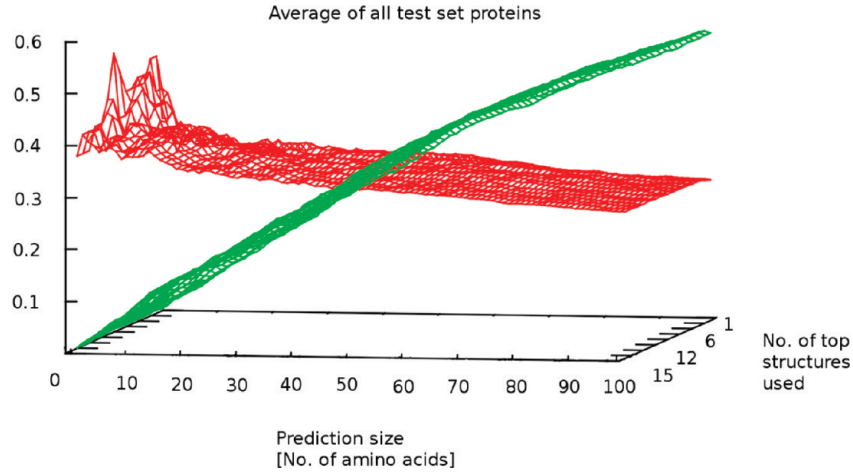

Figure 6. Plots of specificity (red) and sensitivity (green) on the *z*-axis with prediction size expressed as the absolute number of amino acid residues on the *x*-axis. The number of top pairwise comparisons (*N*) used to make a prediction is shown on the *y*-axis.

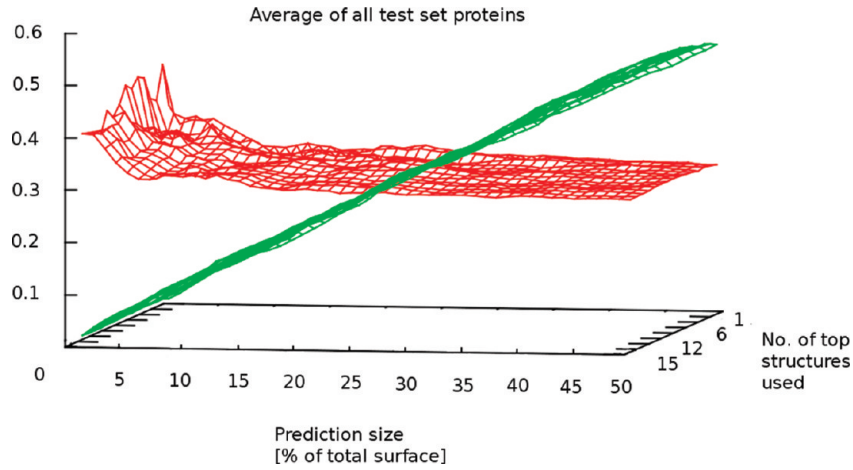

Figure 7. Plots of specificity (red) and sensitivity (green) on the *z*-axis with prediction size expressed as the percentage of all surface residues on the *x*-axis. The number of top pairwise comparisons (*N*) used to make a prediction is shown on the *y*-axis.

The side chains of residues in the predicted binding sites are depicted as yellow stick bonds.

The binding sites predicted by the algorithm were assessed in terms of their specificity and their sensitivity.[13,21] The specificity is the proportion of the binding site residues correctly predicted to be in the binding site, and the sensitivity is the proportion of the actual binding site that was predicted. Predicted binding sites with higher sensitivities are more likely to include residues that are not part of the actual protein−protein binding site, and consequently, increasing sensitivity will decrease the specificity of the predictions. For optimal results, a compromise must be found between these two parameters.

For each approach and each query protein, two 3D plots were constructed, one with specificity and the other with sensitivity plotted against the size of the predicted binding site and, on the third axis, the number *N* of top pairwise comparisons used (*y*-axis). The plots were averaged over the 24 query proteins for each approach, and the 2D graphs derived from these plots at $N = 15$ are shown in Figures 4 and 5.

LSA 2 was found to produce the highest specificity and highest sensitivity among the tested approaches for all prediction sizes (i.e., the number of considered amino acid residues). Figures 4 and 5 show that LSA 2 and LSA 4 produce significantly better specificities and sensitivities than LSA 1 and LSA 3. Since LSA 2 and LSA 4 both select residues from all maximum cliques in their predictions, the second binary decision (see Figure 2) seems to have a bigger influence on specificity and sensitivity than the first one that concerns the ordering of pairwise comparisons.

For the LSA 2 approach, the balance between specificity and sensitivity of predictions was optimized through two parameters, the number of top pairwise comparisons and the prediction size. Two alternate methods were applied to accomplish this optimization: (1) an absolute number of residues was chosen to define the predicted binding site, which implies that for two proteins to bind a limited surface area suffices regardless of the protein size, and (2) a number of conserved residues relative to protein surface size was considered as the predicted binding site, so that bulkier proteins have larger interfaces.

For each of these methods, the specificity and sensitivity plots were constructed for each of the 24 test set proteins, and from these plots, two average plots shown in Figures 6 and 7 were calculated, revealing the balance between sensitivity and specificity. The data for individual proteins can be found in the Supporting Information. The first method (see Figure 6) was chosen as preferable, since it produces a better average specificity at an equal or better average sensitivity.

Table 1. Specificities and Sensitivities for Our Previous Algorithm,[21] Which Used VAST[25] for Finding Proteins To Make Comparisons with, and for Our New Algorithm, Which Uses all PDB Structures To Make Predictions

	specificities		sensitivities	
PDB code and chain ID	previous algorithm	LSA 2 algorithm	previous algorithm	LSA 2 algorithm
1apm E	14.5	8.0	26.8	9.7
1efu A	28.6	38.0	26.5	22.8
1efu B	75.0	36.0	33.7	20.2
1g3n A	25.9	52.0	21.5	40.0
1g3n B	24.5	12.0	46.4	21.4
1g3n C	45.8	26.0	28.9	34.2
1got A	21.1	10.0	33.3	13.8
1got B	60.5	64.0	18.7	26.0
1k9o E	28.7	20.0	65.8	26.3
1k9o I	8.3	4.0	14.3	9.5
1rrp A	47.2	40.0	23.6	27.7
1rrp B	7.7	32.0	1.5	23.5
1ugh E	0.0	12.0	0.0	17.1
1ugh I	0.0	42.0	0.0	61.7
1ytf A	24.2	12.0	44.4	33.3
1ytf D	83.7	68.0	55.4	45.9
1all A	37.5	42.0	69.8	48.8
1aze A	37.0	46.0	37.0	85.1
1bnc A	16.8	26.0	45.2	30.9
1daa A	30.7	46.0	36.5	36.5
1luc A	52.1	20.0	58.5	20.0
1hcg A	13.5	26.0	45.4	30.3
1lw6 I	32.4	32.0	66.7	88.8
1tco B	41.3	54.0	47.0	40.9
average values	31.5	32.0	35.3	34.0

The number of top pairwise comparisons needed to maximize specificity without compromising sensitivity of the prediction was found to be 15 through analysis of a plot similar to Figure 6, but with the number of pairwise comparisons (N) plotted between 1 and 100. No distinction between protein classes was made. When >15 top pairwise comparisons were included, specificity and sensitivity of the predicted binding sites did not significantly improve, while a smaller value decreased the average results. The prediction size was set to 50 amino acid residues, which is close to the established size of a typical protein−protein binding site.[26]

The results of protein−protein binding site prediction with the optimized LSA 2 approach for all protein chains in our test set and the comparison with our previous method are presented in Table 1. The average specificity for the optimized LSA 2 approach is 32%, which is roughly equivalent to 31.5% from our previous algorithm.[21] Thus, we have shown that searching for similarities in fold-independent comparison protein sets is as efficient as in sets of proteins with similar folds.

4. DISCUSSION

The aim of this work was, first, to develop a method for structure similarity searching that will not depend on proteins with similar folds and will enable binding site predictions for proteins with few or no structural neighbors; second, to show that only local structural similarities suffice to predict protein−protein interactions; and, third, to simplify the method of combining data from maximum cliques into binding site predictions without loss of specificity or sensitivity.

In our previous algorithm[21] maximum cliques that represented adjacent local similarities in 3D space were combined into clusters that were then sorted by rmsd between the query and the comparison protein. This procedure called for proteins that were similar in the orientation of secondary structure elements and were less suitable for finding non-contiguous binding sites. To make a fold-independent comparison, an alternative way of combining local structural similarities into a binding site prediction was developed. Whereas in the previous algorithm[21] we sought local similarities that can form larger clusters, in the LSA algorithm only the frequency of a particular functional group in different local structural similarities guides the prediction. The average specificities and sensitivities for both algorithms are roughly equal, showing that both approaches are equally successful on average.

However, the case by case divergence between the current and the previous method[21] is caused by different approaches to combining local similarities into binding sites predictions. For some protein structures, any of the two methods will emphasize structural similarities that are a part of an interface from the data set, while for other protein structures, similarities that stem from binding interfaces not present in the data set or from nonbinding protein building blocks will be found. The differences in specificities and sensitivities for individual proteins originate from the differences between the two methods in detecting the most important local structural similarities.

5. CONCLUSIONS

We have described a local structure alignment (LSA) algorithm for prediction of protein−protein binding sites. This algorithm detects locally similar protein structures from a protein database and combines their similarities into a predicted binding site. It has been tested in four different variants, from which LSA 2 was chosen as superior. This local surface-oriented approach, which utilizes all currently available protein structural information to predict protein−protein binding sites, could be particularly useful for detection of binding sites on protein structures with novel folds. Furthermore, our LSA 2 algorithm was shown to be able to filter a large database to detect locally similar proteins and successfully predict binding sites without loss in efficiency by searching for local similarities only (i.e., fold independently).

ACKNOWLEDGMENT

The financial support through grant P1-0002 of the Ministry of Higher Education, Science, and Technology of Slovenia and the Slovenian Research Agency is acknowledged.

Supporting Information Available: Sensitivities and specificities, as a function of the number of top structures selected, and prediction size are documented. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Matthiesen, R. Methods, Algorithms and Tools in Computational Proteomics: A Practical Point of View. *Proteomics* **2007**, *7*, 2815–2832.

(2) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(3) de Vries, S. J.; Bonvin, A. M. J. J. How Proteins Get in Touch: Interface Prediction in the Study of Biomolecular Complexes. *Curr. Protein Pept. Sci.* **2008**, *9*, 394–406.
(4) Liang, S. D.; Zhang, C.; Liu, S.; Zhou, Y. Q. Protein Binding Site Prediction Using an Empirical Scoring Function. *Nucleic Acids Res.* **2006**, *34*, 3698–3707.
(5) Kufareva, I.; Budagyan, L.; Raush, E.; Totrov, M.; Abagyan, R. PIER Protein Interface Recognition for Structural Proteomics. *Proteins* **2007**, *67*, 400–417.
(6) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH−A Hierarchic Classification of Protein Domain Structures. *Structure* **1997**, *5*, 1093–1108.
(7) Li, J. J.; Huang, D. S.; Wang, B.; Chen, P. Identifying Protein−Protein Interfacial Residues in Heterocomplexes Using Residue Conservation Scores. *Int. J. Biol. Macromol.* **2006**, *38*, 241–247.
(8) Zhong, S. J.; MacKerell, A. D. Binding Response: A Descriptor for Selecting Ligand Binding Site on Protein Surfaces. *J. Chem. Inf. Model.* **2007**, *47*, 2303–2315.
(9) Porollo, A.; Meller, J. Prediction-Based Fingerprints of Protein−Protein Interactions. *Proteins* **2007**, *66*, 630–645.
(10) Ofran, Y.; Rost, B. Analysing Six Types of Protein−Protein Interfaces. *J. Mol. Biol.* **2003**, *325*, 377–387.
(11) Chen, H. L.; Zhou, H. X. Prediction of Interface Residues in Protein−Protein Complexes by a Consensus Neural Network Method: Test Against NMR Data. *Proteins* **2005**, *61*, 21–35.
(12) Res, I.; Mihalek, I.; Lichtarge, O. An Evolution Based Classifier for Prediction of Protein Interfaces Without Using Protein Structures. *Bioinformatics* **2005**, *21*, 2496–2501.
(13) Bradford, J. R.; Westhead, D. R. Improved Prediction of Protein-Protein Binding Sites Using a Support Vector Machines Approach. *Bioinformatics* **2005**, *21*, 1487–1494.
(14) Neuvirth, H.; Raz, R.; Schreiber, G. ProMate: A Structure Based Prediction Program to Identify the Location of Protein−Protein Binding Sites. *J. Mol. Biol.* **2004**, *338*, 181–199.
(15) Bradford, J. R.; Needham, C. J.; Bulpitt, A. J.; Westhead, D. R. Insights into Protein−Protein Interfaces Using a Bayesian Network Prediction Method. *J. Mol. Biol.* **2006**, *362*, 365–386.
(16) Burgoyne, N. J.; Jackson, R. M. Predicting Protein Interaction Sites: Binding Hot-Spots in Protein−Protein and Protein−Ligand Interfaces. *Bioinformatics* **2006**, *22*, 1335–1342.
(17) Schmidt am Busch, M.; Lopes, A.; Amara, N.; Bathelt, C.; Simonson, T. Testing the Coulomb/Accessible Surface Area Solvent Model for Protein Stability, Ligand Binding, and Protein Design. *BMC Bioinf.* **2008**, *9*, 148.
(18) Wang, B.; Chen, P.; Huang, D. S.; Li, J. J.; Lok, T. M.; Lyu, M. R. Predicting Protein Interaction Sites from Residue Spatial Sequence Profile and Evolution Rate. *Febs Lett.* **2006**, *580*, 380–384.
(19) Shoemaker, B. A.; Zhang, D. C.; Thangudu, R. R.; Tyagi, M.; Fong, J. H.; Marchler-Bauer, A.; Bryant, S. H.; Madej, T.; Panchenko, A. R. Inferred Biomolecular Interaction Server−A Web Server to Analyze and Predict Protein Interacting Partners and Binding Sites. *Nucleic Acids Res.* **2010**, *38*, D518–D524.
(20) Aharoni, A.; Gaidukov, L.; Khersonsky, O.; Gould, S. M.; Roodveldt, C.; Tawfik, D. S. The 'Evolvability' of Promiscuous Protein Functions. *Nat. Genet.* **2005**, *37*, 73–76.
(21) Carl, N.; Konc, J.; Janezic, D. Protein Surface Conservation in Binding Sites. *J. Chem. Inf. Model.* **2008**, *48*, 1279–1286.
(22) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of Functional Sites in Protein Structures. *J. Mol. Biol.* **2004**, *339*, 607–633.
(23) Shulman-Peleg, A.; Shatsky, M.; Nussinov, R.; Wolfson, H. J. Spatial Chemical Conservation of Hot Spot Interactions in Protein−Protein Complexes. *BMC Biol.* **2007**, *5*, 43.
(24) Block, P.; Paern, J.; Hullermeier, E.; Sanschagrin, P.; Sotriffer, C. A.; Klebe, G. Physicochemical Descriptors To Discriminate Protein−Protein Interactions in Permanent and Transient Complexes Selected by Means of Machine Learning Algorithms. *Proteins* **2006**, *65*, 607–622.
(25) Gibrat, J. F.; Madej, T.; Bryant, S. H. Surprising Similarities in Structure Comparison. *Curr. Opin. Struc. Biol.* **1996**, *6*, 377–385.
(26) Konc, J.; Janezic, D. ProBiS Algorithm for Detection of Structurally Similar Protein Binding Sites by Local Structural Alignment. *Bioinformatics* **2010**, *26*, 1160–1168.
(27) Konc, J.; Janezic, D. ProBiS: A Web Server for Detection of Structurally Similar Protein Binding Sites. *Nucleic Acids Res.* **2010**, *38*, W436–W440.
(28) Weskamp, N.; Kuhn, D.; Hullermeier, E.; Klebe, G. Efficient Similarity Search in Protein Structure Databases by *k*-Clique Hashing. *Bioinformatics* **2004**, *20*, 1522–1526.
(29) Konc, J.; Janezic, D. Protein−Protein Binding-Sites Prediction by Protein Surface Structure Conservation. *J. Chem. Inf. Model.* **2007**, *47*, 940–944.
(30) Konc, J.; Hodoscek, M.; Janezic, D. Molecular Surface Walk. *Croat. Chem. Acta* **2006**, *79*, 237–241.
(31) Schmitt, S.; Kuhn, D.; Klebe, G. A New Method To Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.* **2002**, *323*, 387–406.
(32) Konc, J.; Janezic, D. An Improved Branch and Bound Algorithm for the Maximum Clique Problem. *MATCH−Commun. Math. Comput. Chem.* **2007**, *58*, 569–590.
(33) The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC. http://www.pymol.org/.

CI100265X