Novel consensus approach for protein active sites detection

Goran Piskachev, Georgina Mirceva
Ss. Cyril and Methodius University in Skopje, Skopje, Macedonia,
goran.piskachev@gmail.com, georginamirceva@gmail.com

Danco Davcev
University for Information Science and Technology "St. Paul The Apostole", Ohrid, Macedonia,
dancodavcev@gmail.com

Abstract: *The analysis of protein structure has an enourmous impact on the development of medicine and pharmacy. We focus our research on detecting the active sites of the protein structure, which are the amino acid residues where the interaction with other protienis would occur. The knowledge about the protein active sites could be later used for discovering the function of the protein structure. In this paper we compare several existing methods for protein active sites detection. Also, we propose a novel consensus approach that combines the predictions obtained from several existing methods. This consensus approach is less biased towards the active sites class. The knowledge stored in the BIND database is considered as standard of truth.*

Keywords: *protein structure, protein active sites, protein function, BIND.*

1. INTRODUCTION

The protein structures are stored in the world-wide repository Protein Data Bank (PDB) [3], which is the primary repository for experimentally determined protein structures. With the technology innovation the number of protein structures increases every day. The determination of protein functions is more complex than discovering new proteins, so this is the reason for developing computer methods that can be used besides the experimental analyses that are very expensive.

The protein's function can be analysed on different level; from the atomic characteristics to describing the function in some biological process or the influence in some illnesses. Most of the proteins have individual functions, but while interacting with other proteins they can gain new functions. The information for the interactions is stored in different databases like MIPS [15], IntAct [9], DIP [18], BIND [2] etc. The tertiary structure that describes the conformation of the molecule in 3D space is very important for the protein function determination because it has much information for the interaction possibilities of the molecule. There are different types of computer based methods for discovering protein functions like methods based on analysis of the protein interaction-networks [11], [17], method based on analysis of the conservation of the protein sequence and/or structure [7] and methods based on protein active sites detection [2]. Since the methods for protein function determination based on active sites detection are the most similar with the manual way in which molecular biologist annotates the protein structure, we focus on this type of methods.

In this paper we analyse several existing methods for detecting protein active sites. These methods consider different features of the amino acids. Also, we propose a novel consensus approach that combines the predictions obtained from several existing methods. In the evaluation of the methods, the knowledge stored in the BIND database [2] is considered as standard of truth.

2. METHODS FOR PROTEIN ACTIVE SITES DETECTION

In order to detect the protein active sites, first we should determine which amino acids are on the protein surface. For that purpose, we need to extract the accessible surface area (ASA) of the amino acids. ASA is first described by Lee and Richards in 1971. It is typically calculated using the 'rolling ball' algorithm [16] developed by Shrake and Rupley. In this algorithm a sphere with a specific radius *R* is rolled on the surface of the molecule. ASA of an amino acid is defined as the area of the amino acid that can get in contact with the sphere that is rolling. The radius *R* is given by the sum of the van der Waals radius of the atom and the chosen radius of the solvent molecule (probe). Typically, the solvent molecule has the same radius (1.4 Å) as the water molecule. The choice of the radius of the sphere is important because smaller radius will calculate for greater ASA values.

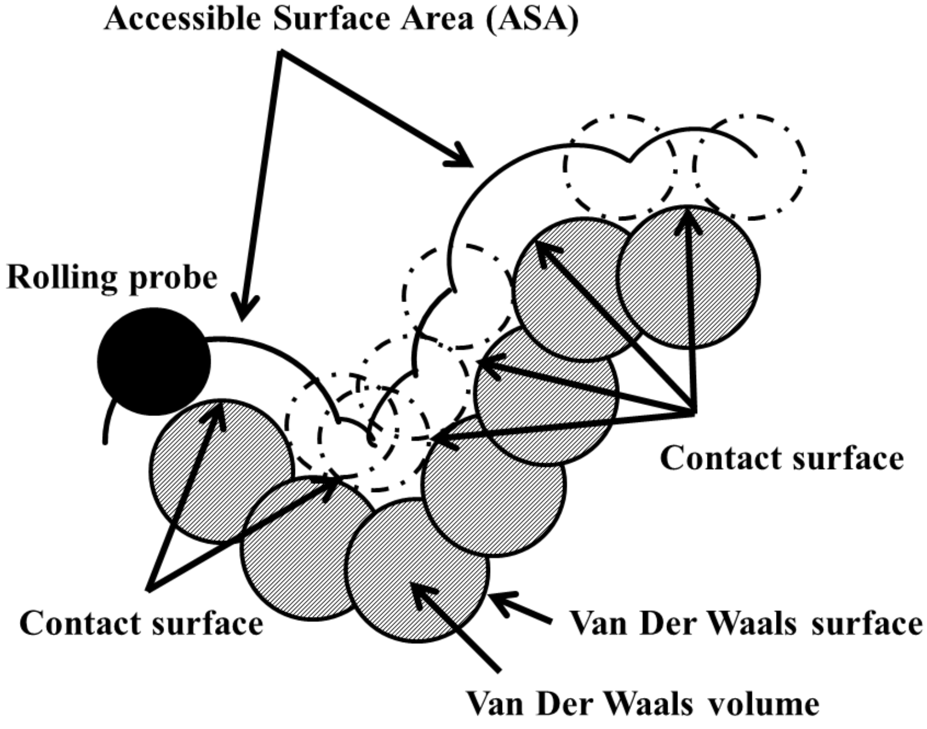

Fig. 1: Extraction of the Accessible Surface Area (ASA).

The value of ASA is expressed in Å^2 (square Ångstroms). Ångstrom is a unit for length and it is 10,000,000,000 part of one meter, and it is yet not included in SI (International System of Units), but it is very often used for inter-molecular distances. In [8], an analytical method for calculating the ASA value is proposed. However, the analytic calculation of the ASA value is difficult and it needs advanced knowledge of geometry. Therefore, most of the methods for calculation of the ASA value are numerical.

The 'rolling ball' algorithm [16] makes successive thin slices - z-slices through the surface to calculate the accessible surface of the individual atoms. The intersections of the solvent sphere with a given z-slice appear as arcs. The ASA is the sum of these arcs' lengths over all z-slices. The overlapping arcs representing the atoms of the same molecule are eliminated. The drawing in any slice becomes the trace of the envelope of the Van Der Waals surface of the molecule. The accessible surface area per slice is approximated by (1), where L_i is the length of the arc drawn in slice i, Z_i is the perpendicular distance from the centre of the sphere to the slice i, ΔZ is the spacing between the slices, while $\Delta' Z$ is $\Delta Z/2$ or R-Z_i (whichever is smaller). The ASA is calculated by summing the arcs drown by all z-slices. This algorithm for calculation of the ASA is not fast because it requires nested loops, but it is one of the easiest for realization and understanding.

(1)
$$ASA = \frac{R}{\sqrt{R^2 - Z_i^2}} * D * L_i$$

(2)
$$D = \frac{\Delta Z}{2} + \Delta' Z$$

Based on the ASA value, we estimate whether a given amino acid is at the protein surface or it is buried in the protein interior. Namely, according to [6] we consider a given amino acid to be at the surface if at least 5% of its area could be solved by the rolling probe. For that reason we use the values for the total area of each amino acid given in [6]. In the evaluation we consider only the amino acids that are located at the protein surface, since amino acids buried in the interior could not be active sites.

In this research we analyze five methods for protein active site detection. The first method is introduced by Ofran and Rost [14] and is based on the Atom Nucleus Distance. It is the simplest one and it classifies two residues as interacting if there is at least one pair of non-hydrogen atoms from each residue, which are closer than some threshold distance, usually 4.5–6 Å.

The second method is introduced by Aytuna [1] and is based on the Atom Van Der Waals Radii Distance. For this method the distance of the two interacting atoms should be smaller than the sum of their Van Der Waals Radii plus a given threshold, usually 0.5–1.5 Å. In this method the Van Der Waals Radii is considered, which is introduced by Johannes Diderik van der Waals, who received Nobel Prize for Physics.

The third method is introduced by Jones and Thornton [10] and is based on the ASA change on complexation. The method first calculates the ASA value in unbound state and then in bound state of the structure. The residue is classified as an interacting one if the difference of these two ASA values is greater than a given threshold.

Protein Interaction Atom Distance Algorithm (PIADA) [12] is the fourth method considered in this research. It includes the hydrophobic and electrostatic characteristics of the amino acids. This method classifies the interactions in four types: ionic, Van Der Waals, hydrophobic and polar. Hydrophobic interaction is defined between any two non-polar amino acid residues Alanine, Isoleucine, Leucine, Methionine, Phenylalanine, Proline and Valine, if the distance between any two residues is smaller than 4.7 Å.

The fifth method, PRISM (Protein Interactions by Structural Matching) [13], uses a database with templates of interfaces. In order to find every possible binary interaction between pairs of structures, the surfaces of the target proteins are extracted and successive structural alignments between these surfaces and the partner chains of interfaces in the template interface dataset is performed in all-against-all manner. In this way the 'structural similarity' is measured. Further, the presence of hotspots on the target structure is checked. The hotspot match ratio is used for the calculation of the 'evolutionary similarity score', whereas the structural match ratio is used for a 'structural similarity score'. A combination of these scores is used to calculate the overall prediction score.

In this paper we propose a novel consensus approach that combines the predictions of the three methods with best prediction power. These methods are the method based on the Atom Nucleus Distance, the method based on the Atom Van Der Waals Radii Distance and PIADA. If at least two methods predict that the amino acid is an active site, than the proposed method would classify that amino acid as an active site.

3. EVALUATION

For evaluation of the methods considered in this study we used a part of the BIND database [2], [4] that contains experimentally obtained knowledge. We used a representative non-redundant test dataset so that each pair of protein chains has less than 20% sequence similarity, using the selection criterion in [5]. Since this selection criterion considers protein chains with a low sequence similarity, the most representative protein chains are considered as a test data. In this way we obtain 1858 test protein chains. The total number of amino acids in the test chains is 608434 from which 484637 are on the surface. From these 484637 surface amino acids, only 47501 are classified as active sites in the BIND database. Since the non-active class is dominant, we have to use appropriate evaluation measure that would not favour the dominant class.

For evaluation, we used several standard measures that are widely used in information retrieval, i.e. True Positive Rate, True Negative Rate and Area under ROC curve (AUC-ROC). In order to calculate these measures, we have to calculate TP (true positives), FP (false positives), TN (true negatives) and FN (false negatives), where positive examples are those examples that are classified as active sites and negative examples are those examples that are classified in the non-active class. TP is the number of the correctly classified positive examples, FP is the number of negative examples classified as positive, TN is the number of correctly classified negative examples and FN is the number of positive examples classified as negative. Based on these measures, the True Positive Rate (TPR), the True Negative Rate (TNR) and the Area under ROC curve (AUC-ROC) could be calculated (3). The AUC-ROC is the most appropriate measure, especially when the classes are of very different sizes, like it is in our case. AUC-ROC is a value between 0 and 1, where 1 represents a perfect prediction, while 0 an inverse prediction.

$$\begin{gathered} \mathrm{TPR} = \mathrm{TP} / (\mathrm{TP}+\mathrm{FN}) \\ \mathrm{TNR} = \mathrm{TN} / (\mathrm{TN}+\mathrm{FP}) \\ \mathrm{AUC\text{-}ROC} = \mathrm{TPR}*\mathrm{TNR} + \mathrm{TPR}*(1-\mathrm{TNR}) / 2 + \mathrm{TNR}*(1-\mathrm{TPR}) / 2 = (\mathrm{TPR}+\mathrm{TNR}) / 2 \end{gathered} \tag{3}$$

The results of the comparison are given in Tab. 1. From the existing methods, the method based on the Atom Van Der Waals Radii Distance is the best with AUC-ROC = 0.832, then the method based on the Atom Nucleus Distance and PIADA follow with AUC-ROC of 0.824 and 0.823 respectively. Next, the method based on ASA change follows with AUC-ROC = 0.810, while the PRISM method has significantly lower AUC-ROC of 0.674 and it significantly favors the non-active class (TNR = 0.949 and TPR = 0.399).

It can be seen that the proposed consensus approach obtain lower value for AUC-ROC (0.825) than the method based on the Atom Van Der Waals Radii Distance. However, the proposed approach is less biased towards the active class than most of the methods. We expect that if we combine the predictions of more methods using appropriate weighting the AUC-ROC would be increased.

Tab. 1: Evaluation of the methods.

Method	TPR	TNR	AUC-ROC
ASA change	0.875	0.744	0.810
Atom Nucleus Distance	0.861	0.788	0.824
Van der Waals distance	0.840	0.824	0.832
PIADA	0.864	0.783	0.823
PRISM	0.399	0.949	0.674
The proposed consensus approach	0.860	0.791	0.825

4. CONCLUSION

In this paper we analyzed several existing methods for protein active sites detection, i.e. the method based on ASA change, the method based on the Atom Nucleus Distance, the method based on the Atom Van Der Waals Radii Distance, PIADA and PRISM. Also, we proposed a novel consensus method that combines the predictions of the three methods with highest AUC-ROC.

The analysis showed that the method based on the Atom Van Der Waals Radii Distance has best prediction power, then the other methods except PRISM follows with comparable AUC-ROC, while the PRISM method has significantly lower AUC-ROC. The proposed consensus approach has lower AUC-ROC than the method based on the Atom Van Der Waals Radii Distance. However, it is less biased towards the active class than most of the methods. We expect that if we combine the predictions of more methods using appropriate weight function the AUC-ROC could be increased.

As a future work we plan to analyze other methods for protein active sites detection, and to investigate more accurate consensus approach where using corresponding weight function we would give different significances to the methods' predictions.

5. REFERENCES

[1] Aytuna, A.S., Gursoy, A., Keskin, O. (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* 21, 12, 2850-2855.

[2] Bader, G.D., Betel, D., Hogue, C.W.V. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research* 31, 1, 248-250.

[3] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research* 28, 235-242.

[4] BIND database, http://bond.unleashedinformatics.com.

[5] Chandonia, J.M., Hon, G., Walker, N.S., Conte, L.L., Koehl, P., Levitt, M., Brenner, S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Research* 32, D189-192.

[6] Chothia, C. (1976) The Nature of the Accessible and Buried Surfaces in Proteins. *Journal of Molecular Biology* 105, 1, 1-12.

[7] Hannenhalli, S.S., Russell, R.B. (2000) Analysis and prediction of functional subtypes from protein sequence alignments. *Journal of Molecular Biology* 303, 1, 61-76.

[8] Hayryan, S., Hu, C.K., Skrivánek, J., Hayryane, E., Pokorný, I. (2005) A New Analytical Method for Computing Solvent-Accessible Surface Area of Macromolecules and its Gradients. *Journal of Computational Chemistry* 26, 4, 334-343.

[9] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Research* 32, D452-D455.

[10] Jones, S., Thornton, J.M. (1997) Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology* 272, 1, 121-132.

[11] Letovsky, S., Kasif, S. (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* 19, 197-204.

[12] Mihel, J., Šikić, M., Tomić, S., Jeren, B., Vlahoviček, K. (2008) PSAIA – Protein Structure and Interaction Analyzer. *BMC Structural Biology* 8, 21.

[13] Ogmen, U., Keskin, O., Aytuna, A.S., Nussinov, R., Gursoy, A. (2005) PRISM: protein interactions by structural matching. *Nucleic Acids Research* 33, W331-W336.

[14] Ofran, Y., Rost, B. (2003) Predicted protein-protein interaction sites from local sequence information. *FEBS Lett.* 544, 236-239.

[15] Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H.W., Ruepp, A., Frishman, D. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics* 21, 6, 832-834.

[16] Rupley, J.A., Shrake, A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology* 79, 2, 351-371.

[17] Schwikowski, B., Uetz, P., Fields, S. (2000) A network of protein-protein interactions in yeast. *Nature Biotechnology* 18, 12, 1257-1261.

[18] Xenarios, I., Salwínski, L., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* 30, 1, 303-305.