SciForum
Mol2Net

Solvent Accessible Surface Area Hot-Spot Detection Method

Cristian R. Munteanu[1,*], António Pimenta[2], Carlos Fernandez-Lozano[1], André Melo[3], Maria Cordeiro[3], Irina S. Moreira[2,*]

[1] Information and Communication Technologies Department, Computer Science Faculty, University of A Coruna, Campus de Elviña s/n, 15071, A Coruña, Spain; E-mail: crm.publish@gmail.com (CR.M.); carlos.fernandez@udc.es (C.FL.)

[2] CNC - Center for Neuroscience and Cell Biology; Rua Larga, FMUC, Polo I, 1ºandar, Universidade de Coimbra, 3004-517; Coimbra, Portugal; E-mail: caesar.m4d@gmail.com (A.P.): irina.moreira@cnc.uc.pt (I.S.M.)

[3] REQUIMTE/Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal; E-mail: asmelo@fc.up.pt (A.M.); ncordeir@fc.up.pt (M.C.)

*Author to whom correspondence should be addressed; E-Mail: irina.moreira@cnc.uc.pt or crm.publish@gmail.com.

Tel.: +351-239-820-190 (ext. 123); Fax: +351-239-822-776.

Published: 4 December 2015

Abstract: The natural tendency of proteins to bind to each other, as well as to many different molecules, forming stable and specific complexes is fundamental to all biological processes. The structural and functional description of protein-protein and protein-ligand complexes and their comprehension is a key concept, not only to increase the scientific knowledge in basic terms but also for the application to the biomedical and pharmaceutical industry. In this work we have look for more accurate ways of predicting the crucial residues for complex binding (Hot-spots) that can be used to model protein structure, dynamics and function. We developed an algorithm based in innovative series of descriptors, which have not been used in hot-spot determination and that can be applied to both protein-protein and protein-nucleic acid interfaces HS detection. A web-server for public use of the new methodological approaches was built and can be accessed at http://bio-aims.udc.es/MolStructPred.php

Keywords: Hot-spots; conservation; solvent accessible surface area; machine-learning, protein-protein interfaces; protein-nucleic acid interfaces.

1. Introduction

Protein-protein interactions (PPIs) are fundamental for all life processes and it is vital to understand their dynamics, structural and energetic characteristics in order to find new improved ways to influence these molecular machineries[1]. Traditional mutagenesis approaches, including the use of hybrid receptors and alanine scanning mutagenesis techniques, have led to important insights into the structural basis underlying PPIs. However, experimental mutagenesis scanning of a complete interface is highly costly from a financial and time point of view[1-3]. To overcome this problem it was needed an efficient and fast computational technique that allows the detection of the major binding determinants at a protein-protein interface: the Hot-Spots (HS). HS tend to be conserved residues tightly clustered in the central part of protein-protein interfaces forming a network of specific interactions that are optimized and cooperative[4]. Figure 1 illustrates an example of a protein-protein complex in which HS are highlighted in a vdW red representation and non-HS (called Null-Spots NS) in a yellow one. HS tend to be surrounded by a region of supposedly "less important" residues, largely hydrophobic, that leads to solvent occlusion and results in a lower local dielectric constant environment and enhancement of specific electrostatic and hydrogen bond interactions (Figure 1)[5]. So, according to this theory (O-ring theory), HS regions have a low number of interfacial waters, implying that water entropy effects provide one of the driving forces to complex formation[6] and that occlusion of bulk solvent slows down dissociation. Having these knowledge gathered through the years about HS[1-4,7], we decided to look for a method based on genomic conservation scores and 12 different Solvent Accessible Surface Areas features (described at reference[8]).

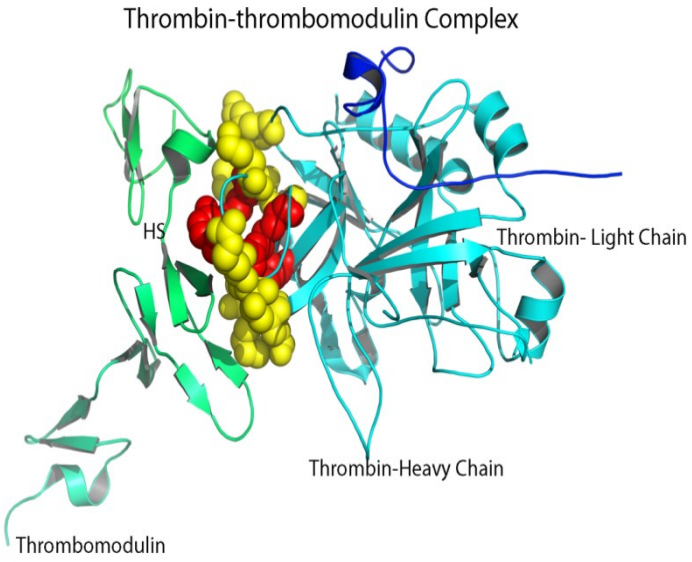

Figure 1. Structural representation of a protein-protein complex (PDBid: 1DX5[9]) in which the HS and NS are highlighted in a red and yellow vdW representation, respectively.

2. Results and Discussion

The performance in ML is usually measured using predictive accuracy, which could be problematic if the data is unbalanced[10]. Dataset S1 comprised 71 HS/406 NS, dataset S2 35 HS/56 NS, dataset S3 60 HS/162 NS and dataset S4 20 HS/80 NS, which demonstrates that our datasets (described at reference[8]) are highly unbalanced (classes are not equally represented as HS are less represented in Nature). This way, we evaluated the performance of each model by taking into account Recall (TPR), Precision, Specificity and FPR as well as F1-score and AUROC. We showed that simple Bayes Networks were able to classify HS for protein-protein interactions but only complex methods such as GA-SVM-Full could be used to classify HS for protein-nucleic acid interactions.

The best classifier for protein-protein case uses four features: CONSURF score, $\Delta SASA_i$, $_{rel/res}SASA_i$ and $_{rel/ave}SASA_i$. (TPR=0.79, FPR=0.21, Precision=0.87, F1-score=0.83 and AUROC=0.85). Our algorithm was assessed against some of the state-of-the-art methods available by web-servers and proven to more accurately predict HS at protein-protein interfaces.

For protein-Nucleic Acid the best classifier uses two features: ConSurf score, $_{del/res}SASA_i$ (TPR=0.82, FPR=0.30, Precision=0.82, F1-score=0.85 and AUROC=0.83).

3. Materials and Methods

Three different datasets were used for the protein-protein interfaces: ASEdb,[11] BID[12] and SKEMPI[13] (comprising a total of 790 residues from 58 complexes) and one for protein-nucleic Acid: Pronit[14-16] (a total of 117 residues from 28 complexes). The datasets were constituted by protein complexes for which simultaneously exists experimental alanine scanning mutagenesis data, genetic conservation scores and tridimensional crystallographic structures of the bounded complex. These ones were filtered to ensure that a maximum of 35% sequence identity could be found for at least one protein in each interface[8]. Various machine-learning (ML) techniques were employed for this particular problem and in order to improve the performance and to reduce the number of features in the input space we also performed a Feature Selection (FS) approach as the number and relevance of the input variables can affect the performance of the model. Several statistics analyzes were performed to ensure the achievement of the high accuracy method.

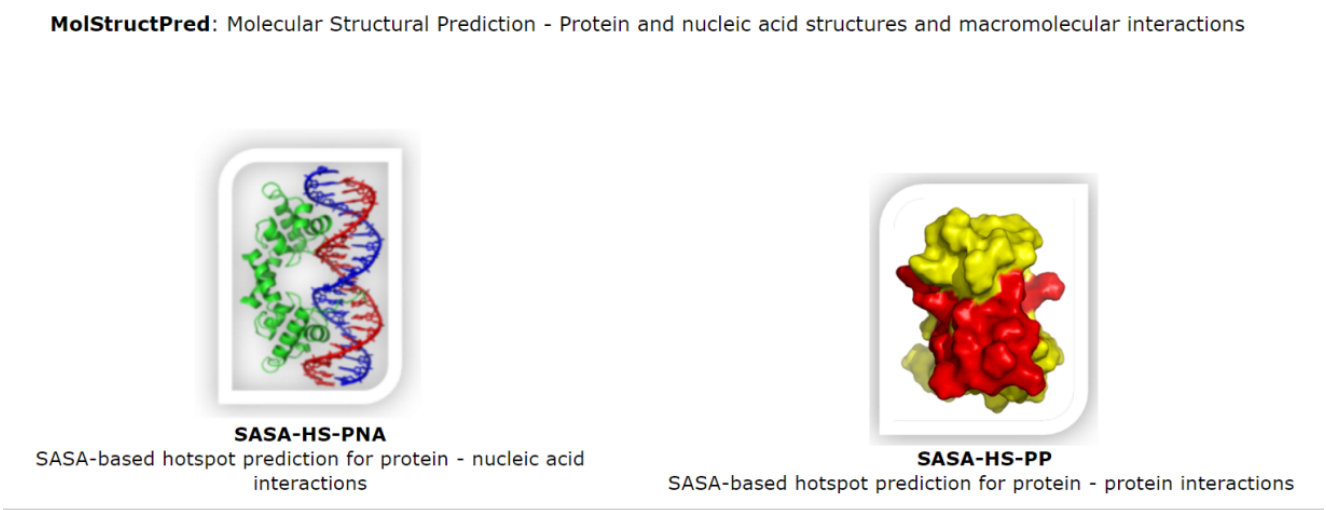

Figure 2. Web-server for HS detection.

4. Conclusions

Our methods are accurate and time efficient. Moreover, our method can be applied not only to protein-protein but as well, and for the first time, to protein-nucleic acid complexes[8]. Web-servers were also constructed and made available for the scientific community at BioAIMS portal (http://bio-aims.udc.es/MolStructPred.php). The code of the Web tools is available as pySBHD repository (https://github.com/muntisa/pySBHD).

Acknowledgments

This work was funded by FEDER funds through the Operational Programme for Competitiveness Factors (COMPETE) and by National Funds through FCT (Foundation for Science and Technology) under the projects PEST-C/EQB/LA0006/2013 and COMPETE FCOMP-01-0124-FEDER-37285. This work was further supported by "Collaborative Project on Medical Informatics (CIMED)" PI13/00280 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013-2016 and the European Regional Development Funds (FEDER). The authors acknowledge the support by the Galician Network of Drugs R+D REGID (Xunta de Galicia R2014/025). ISM was supported by FCT Ciência 2008 and FCT Investigator programmes - IF/00578/2014 (co-financed by European Social Fund and Programa Operacional Potencial Humano).

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

1. Moreira, I.S.; Fernandes, P.A.; Ramos, M.J. Hot spots—a review of the protein–protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics* **2007**, *68*, 803-812.
2. Moreira, I.S.; Martins, J.M.; Ramos, R.M.; Fernandes, P.A.; Ramos, M.J. Understanding the importance of the aromatic amino-acid residues as hot-spots. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2013**, *1834*, 404-414.
3. Moreira, I.S.; Ramos, R.M.; Martins, J.M.; Fernandes, P.A.; Ramos, M.J. Are hot-spots occluded from water? *Journal of Biomolecular Structure and Dynamics* **2013**, *32*, 186-197.
4. Moreira, I.S. The role of water occlusion for the definition of a protein binding hot-spot *Curr Top Med Chem* **2015**, *15*, 2068-2079.
5. Bogan, A.A.; Thorn, K.S. Anatomy of hot spots in protein interfaces. *J Mol Biol* **1998**, *280*, 1-9.
6. Oshima, H.; Yasuda, S.; Yoshidome, T.; Ikeguchi, M.; Kinoshita, M. Crucial importance of the water-entropy effect in predicting hot spots in protein-protein complexes. *Physical Chemistry Chemical Physics* **2011**, *13*, 16236-16246.
7. Martins, J.M.; Ramos, R.M.; Pimenta, A.C.; Moreira, I.S. Solvent-accessible surface area: How well can be applied to hot-spot detection? *Proteins: Structure, Function, and Bioinformatics* **2014**, *82*, 479-490.
8. Munteanu, C.; Pimenta, A.C.; Fernandez-Lozano, C.; Melo, A.; Dias Soeiro Cordeiro, M.N.; Moreira, I.S. Sasa-based hot-spot detection 2 (sbhd2) methods for protein-protein and protein-nucleic acid interfaces. *Journal of Chemical Information and Modeling* **2015**.
9. Fuentes-Prior, P.; Iwanaga, Y.; Huber, R.; Pagila, R.; Rumennik, G.; Seto, M.; Morser, J.; Light, D.R.; Bode, W. Structural basis for the anticoagulant activity of the thrombin-thrombomodulin complex. *Nature* **2000**, *404*, 518-525.
10. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* **2002**, *16*, 321-357.
11. Thorn, K.S.; Bogan, A.A. Asedb: A database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* **2001**, *17*, 284-285.
12. Fischer, T.B.; Arunachalam, K.V.; Bailey, D.; Mangual, V.; Bakhru, S.; Russo, R.; Huang, D.; Paczkowski, M.; Lalchandani, V.; Ramachandra, C.*, et al.* The binding interface database (bid): A compilation of amino acid hot spots in protein interfaces. *Bioinformatics* **2003**, *19*, 1453-1454.
13. Moal, I.H.; Fernández-Recio, J. Skempi: A structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics* **2012**, *28*, 2600-2607.
14. Kumar, M.D.S.; Bava, K.A.; Gromiha, M.M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A. Protherm and pronit: Thermodynamic databases for proteins and protein–nucleic acid interactions. *Nucleic Acids Research* **2006**, *34*, D204-D206.
15. Prabakaran, P.; An, J.; Gromiha, M.M.; Selvaraj, S.; Uedaira, H.; Kono, H.; Sarai, A. Thermodynamic database for protein-nucleic acid interactions (pronit). *Bioinformatics* **2001**, *17*, 1027-1034.

16. Sarai, A.; Gromiha, M.M.; An, J.; Prabakaran, P.; Selvaraj, S.; Kono, H.; Oobatake, M.; Uedaira, H. Thermodynamic databases for proteins and protein-nucleic acid interactions. *Biopolymers* **2001**, *61*, 121-126.