SciForum
MOL2NET

Coevolution importance on binding Hot-Spot prediction methods

José G. Almeida [1,*], António J. Preto [1], Rita Melo[1,2], Zeynep H. Gümüş[3], Joaquim Costa[4], Alexandre M.J.J. Bonvin[5] and Irina S. Moreira[1,5,*]

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

[2] Centro de Ciências e Tecnologias Nucleares, Instituto Superior Técnico, Universidade de Lisboa, Estrada Nacional 10 (ao km 139,7), 2695-066 Bobadela LRS, Portugal; E-mail: ritamelo@ctn.ist.utl.pt (R.M.)

[3] Department of Genetics and Genomics and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA; E-mail: zeynep.gumus@gmail.com (Z.H.G.)

[4] Department of Mathematics, Faculdade de Ciencias, Universidade do Porto, Portugal; E-mail: jpcosta@fc.up.pt (J. C.)

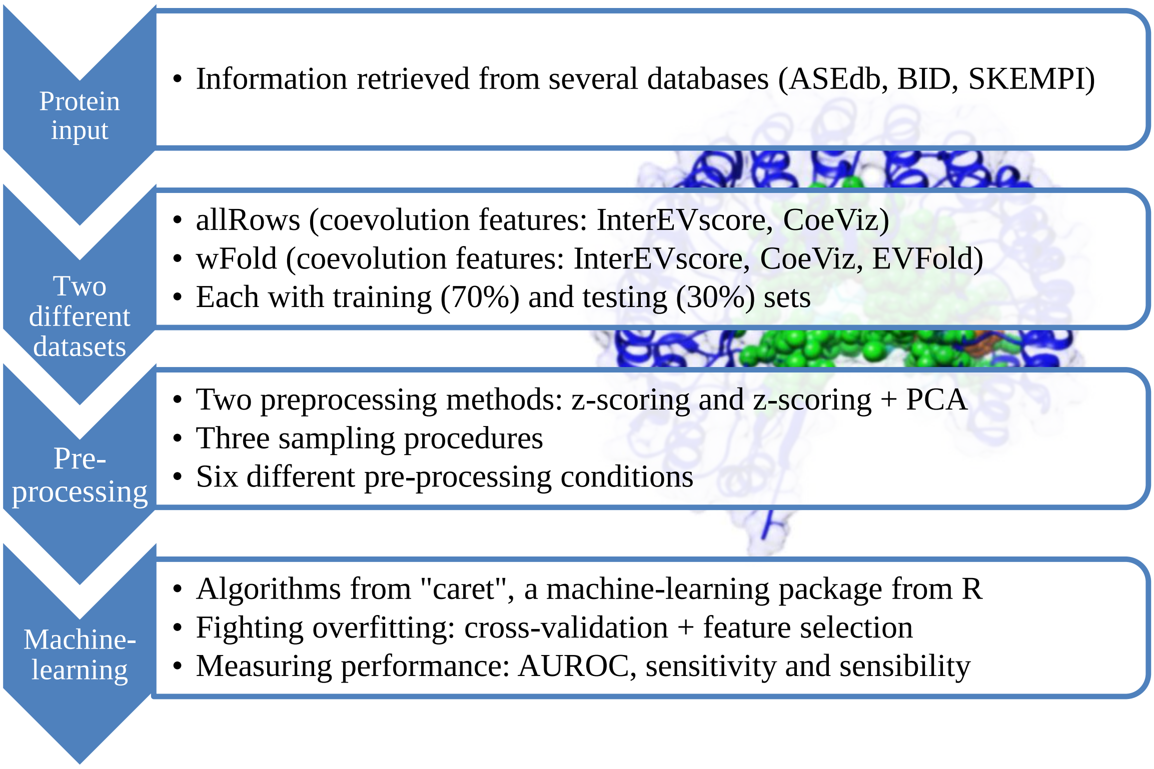
[5] Bijvoet Center for Biomolecular Research, Faculty of Science—Chemistry, Utrecht University, Utrecht 3584CH, The Netherlands; E-mail: a.m.j.j.bonvin@uu.nl (A.M.J.J.B.).

*****Author to whom correspondence should be addressed; E-Mail: irina.moreira@cnc.uc.pt; Tel.: +351-239-820-190 (ext. 123); Fax: +351-239-822-776.

Abstract: Protein-protein interactions (PPIs) have proven necessary for the majority of biological processes, making their understanding vital for the development of new therapies and techniques in life sciences research. Among the residues that constitute a typical protein-protein interface, Hot-Spots (HS) are the most important ones due to their highly stabilizing nature. However, HS experimental detection has proven to be a burden as it is time consuming and expensive, which prompted the need to develop new computational approaches that ensure both speed and precision. Evolution plays a major role in protein structure and PPIs refinement, and therefore the incorporation of such data into a predictive model may lead to better performance. With this in mind and taking into account the data already available from alanine scanning mutagenesis studies and protein structures, we incorporated several structure- (i.e. solvent accessible surface area-related values, sequence- (i.e. position-specific scoring matrix), and evolutionary-based (i.e. InterEVScore and CoeViz) features into a predictive machine-learning classification model. We considered six different pre-processing conditions such as Principal Component Analysis (PCA) and z-scoring (scaling) with normal, up- and down-sampling of minor and major classes. Our results point towards overall better scores when using more evolutionary features, in particular EVFold scores.

Keywords: *hot spots, machine-learning, protein-protein interaction*

Graphical Abstract:


Protein input
• Information retrieved from several databases (ASEdb, BID, SKEMPI)
Two different datasets
• allRows (coevolution features: InterEVscore, CoeViz)
• wFold (coevolution features: InterEVscore, CoeViz, EVFold)
• Each with training (70%) and testing (30%) sets
Pre-processing
• Two preprocessing methods: z-scoring and z-scoring + PCA
• Three sampling procedures
• Six different pre-processing conditions
Machine-learning
• Algorithms from "caret", a machine-learning package from R
• Fighting overfitting: cross-validation + feature selection
• Measuring performance: AUROC, sensitivity and sensibility

Introduction: Almost all biological processes require specific Protein-Protein Interactions (PPI) with high complexity [1]. This fact contributes to diverse vital functions in cellular communication, gene regulation, and immune response [2]. Despite this complexity in function, structural differences are mainly in surface complementarity of PPI. Hot-Spots (HS) represent the main residues involved in PPI with major contributions to binding free energy [3-5], and were experimentally defined upon alanine mutagenesis experiments. However, experimental scanning of a complete interface is both very expensive and time consuming [1, 6, 7]. Evolution-optimized cooperativity and specific interactions have been considered crucial in characterizing HS [8], which provides an important direction when studying both HS [9] and protein-protein interfaces [10, 11]. Adding features as Solvent Accessible Surface Area (SASA) [12] as well as sequence-derived features such as instance Position Specific Scoring Matrices (PSSM) [13] enrich predictive models and promising results can be attained. Coevolution aims at assessing evolutionary conservation of protein sequences and functions [14]. Computationally, they are expressed as coevolution-scores of inter-residue interactions for individual proteins. These data can serve also as input features when using Machine-Learning (ML) techniques for HS prediction.

Materials and Methods: Three different databases were used to construct our final HS dataset: ASEdb [15], BID [16] and SKEMPI [17]. It comprises 533 residues across 53 complexes, all of which have known crystallography-determined structures and known alanine scanning mutagenesis data [18]. Two different datasets were then created – allRows, featuring InterEVscore [19] and CoeViz [20] scores as evolutionary features plus the ones described at Melo *et al.* [21] (533 observations), and wFold, which features the same evolutionary features plus some EVFold [22] scores (264 observations). Due to the lack of a sufficient amount of sequences, we were not able to calculate EVFold scores for all residues at the dataset. Both these datasets were split into training (70%) and testing (30%) sets. Pre-processing of the datasets was performed with two different methods: centering and scaling of all variables (z-scoring) and z-scoring followed by a Principal Component Analysis (PCA). Three sampling methods were also employed: regular sampling, up-sampling of the minor class and down-sampling of the major class for both sets, totaling 6 different "pre-processing conditions" – PCA, PCAUp, PCADown, Scaled, ScaledUp and ScaledDown. Several algorithms for ML from the R package "caret" were used and, in order to improve the performance and to handle the high number of features, training methods such as cross-validation and feature selection were performed. Standard performance ML metrics were used, such as the Area Under the Receiver Operating Curve (AUROC), sensitivity (true positive rate) and specificity (true negative rate).

Results and Discussion: Figure 1 depicts the average AUROC, sensitivity and specificity measured across the 31 ML algorithms tested (for the 2 datasets in every condition). It shows that EVFold scores are quite valuable when predicting HS, as the best results across all pre-processing conditions were achieved for the wFold dataset.

One of the most interesting aspects of this study was the importance of the pre-processing conditions for the outcome of the model. While PCA provided the best results for the wFold testing set, the ScaledUp pre-processing provided the highest AUROC for the wFold training set, which might be an evidence towards overfitting in the ScaledUp pre-processing condition. Furthermore, regarding still the AUROC values, while these scores where lower for the wFold training set as compared with the allRows training set, the scores for the testing set were higher (+0.05), with the testing set AUROC values being higher than the corresponding values for the training set. However, the AUROC should be considered an estimate of the scalability of the model [23] and the testing AUROC vs. training AUROC unorthodox differences might be a consequence of this. Nonetheless, considering that the improvement in the wFold dataset predictive power is observed in all pre-processing conditions, this is regarded not as a consequence of estimation but as evidence towards the importance of evolutionary features in this model. As for the values observed for sensitivity and specificity, these show an apparent opposite tendency, with the best results for the sensitivity achieved in the ScaledDown pre-processing condition and the best results for the specificity achieved in the PCA pre-processing condition. As it is known, sensitivity is the ability of a model to correctly predict a positive case (a HS, in our case), while the specificity is the ability to do so with a negative case (a non-HS). As such, a compromise must be reached between the pre-processing conditions in not risking the ability of the model to correctly assign a residue as HS vs. not overfitting. However, considering that no model showed the best possible results for all three considered metrics, we calculated the mean value for the AUROC, sensitivity and specificity. From this, ScaledUp is highlighted as the best pre-processing condition. It presents a mean value for all three metrics of 0.74, with an AUROC of 0.72, a sensitivity of 0.70 and specificity of 0.80 in the test set.

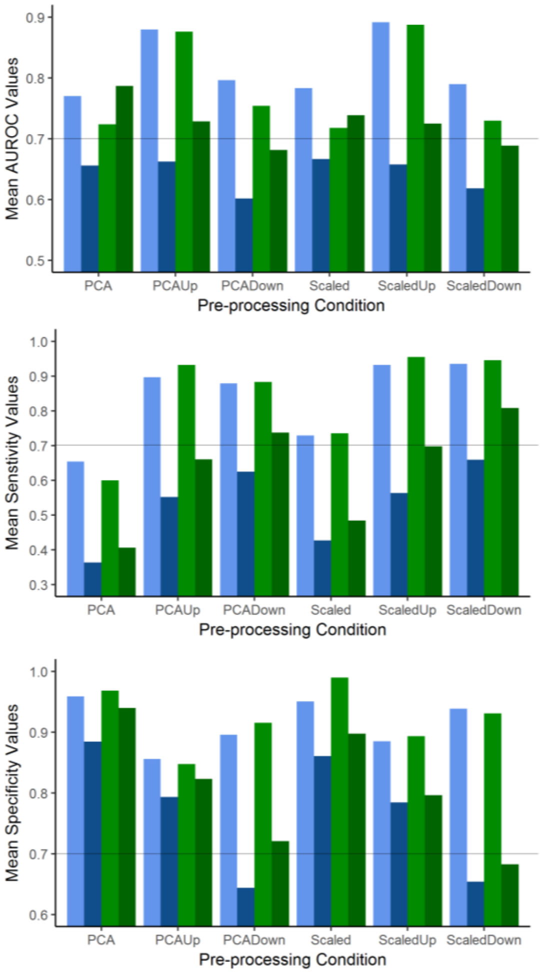


Figure 1 *– Results for the average AUROC, sensitivity and specificity for each one of the pre-processing conditions in the training and testing sets (light blue – allRows training set; dark blue – allRows testing set; light green – wFold training set; dark green – wFold testing set. A threshold value of 0.70 (black horizontal line) was considered in order to classify a model as good (above 0.7) or bad (below 0.7).*

Conclusions: We observed that by increasing the number of evolutionary features in our model we were able to improve ML performance, stressing the underlying role for coevolution as a driving force of HS formation across different protein-protein interfaces. As such, one should take into account as much as possible evolutionary-based features when creating new protein structure/interface prediction methods. Unfortunately, the number of available sequence may still not be sufficient to systematically introduce coevolution in larger systems and/or large datasets.

Acknowledgements: Rita Melo acknowledges support from the Fundação para a Ciência e a Tecnologia (FCT—SFRH/BPD/97650/2013; UID/Multi/04349/2013 project). Irina S. Moreira acknowledges support by the FCT Investigator program—IF/00578/2014 (co-financed by European Social Fund and Programa Operacional Potencial Humano) and by a Marie Skłodowska-Curie Individual Fellowship MSCA-IF-2015 (MEMBRANEPROT 659826). Irina S. Moreira also acknowledges FEDER (Programa Operacional Factores de Competitividade—COMPETE 2020) and FCT–project UID/NEU/04539/2013. Zeynep H. Gümüş acknowledges support from the Center for Basic and Translational Research on Disorders of the Digestive System, Rockefeller University, through the generosity of the Leona M. and Harry B. Helmsley Charitable Trust and from the start-up funds of the Icahn School of Medicine at Mount Sinai. The authors would like to thank R. Fieldhouse for help with the existing EVFold software.

References:

1. Moreira, I.S., P.A. Fernandes, and M.J. Ramos, *Hot spots—A review of the protein–protein interface determinant amino-acid residues.* Proteins: Structure, Function, and Bioinformatics, 2007. **68**(4): p. 803-812.
2. Chothia, C. and J. Janin, *Principles of protein-protein recognition.* Nature, 1975. **256**(5520): p. 705-708.
3. Zerbe, B.S., et al., *Relationship between Hot Spot Residues and Ligand Binding Hot Spots in Protein-Protein Interfaces.* Journal of chemical information and modeling, 2012. **52**(8): p. 2236-2244.
4. Blomenrohr, M., H.F. Vischer, and J. Bogerd, *Receptor mutagenesis strategies for examination of structure-function relationships.* Methods Mol Biol, 2004. **259**: p. 307-22.
5. Cunningham, B.C. and J.A. Wells, *High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis.* Science, 1989. **244**(4908): p. 1081-5.
6. Moreira, I.S., et al., *Understanding the importance of the aromatic amino-acid residues as hot-spots.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2013. **1834**(1): p. 404-414.
7. Moreira, I.S., et al., *Are hot-spots occluded from water?* Journal of Biomolecular Structure and Dynamics, 2013. **32**(2): p. 186-197.
8. Moreira, I.S., *The Role of Water Occlusion for the Definition of a Protein Binding Hot-Spot* Curr Top Med Chem, 2015. **15**(20): p. 2068-2079.
9. Hu, Z., et al., *Conservation of polar residues as hot spots at protein interfaces.* Proteins, 2000. **39**(4): p. 331-42.
10. Ofran, Y. and B. Rost, *Protein-protein interaction hotspots carved into sequences.* PLoS Comput Biol, 2007. **3**(7): p. e119.
11. Guharoy, M. and P. Chakrabarti, *Conservation and relative importance of residues across protein-protein interfaces.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15447-52.
12. Xia, J.-F., et al., *APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility.* BMC Bioinformatics, 2010. **11**: p. 174-174.
13. Dehzangi, A., et al., *Proposing a highly accurate protein structural class predictor using segmentation-based features.* BMC Genomics, 2014. **15**(Suppl 1): p. S2.
14. Tuller, T., Y. Felder, and M. Kupiec, *Discovering local patterns of co - evolution: computational aspects and biological examples.* BMC Bioinformatics, 2010. **11**: p. 43-43.
15. Thorn, K.S. and A.A. Bogan, *ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions.* Bioinformatics, 2001. **17**(3): p. 284-5.

16. Fischer, T.B., et al., *The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces.* Bioinformatics, 2003. **19**(11): p. 1453-4.
17. Moal, I.H. and J. Fernández-Recio, *SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models.* Bioinformatics, 2012. **28**(20): p. 2600-2607.
18. Munteanu, C., et al., *SASA-Based Hot-spot Detection 2 (SBHD2) Methods for Protein-Protein and Protein-Nucleic Acid Interfaces.* Journal of Chemical Information and Modeling, 2015.
19. Andreani, J., G. Faure, and R. Guerois, *InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution.* Bioinformatics, 2013. **29**(14): p. 1742-9.
20. Baker, F.N. and A. Porollo, *CoeViz: a web-based tool for coevolution analysis of protein residues.* BMC Bioinformatics, 2016. **17**: p. 119.
21. Melo, R., et al., *A Machine-Learning Approach for Hot-Spot Detection at Protein-Protein Interfaces.* IJMS, 2016: p. 1-16.
22. Braun, T., J. Koehler Leman, and O.F. Lange, *Combining Evolutionary Information and an Iterative Sampling Strategy for Accurate Protein Structure Prediction.* PLoS Comput Biol, 2015. **11**(12): p. e1004661.
23. Calders, T. and S. Jaroszewicz, *Efficient AUC Optimization for Classification*, in *Knowledge Discovery in Databases: PKDD 2007: 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007. Proceedings*, J.N. Kok, et al., Editors. 2007, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 42-53.