Chapter 14

BindML/BindML+: Detecting Protein-Protein Interaction Interface Propensity from Amino Acid Substitution Patterns

Qing Wei, David La, and Daisuke Kihara

Abstract

Prediction of protein-protein interaction sites in a protein structure provides important information for elucidating the mechanism of protein function and can also be useful in guiding a modeling or design procedures of protein complex structures. Since prediction methods essentially assess the propensity of amino acids that are likely to be part of a protein docking interface, they can help in designing protein-protein interactions. Here, we introduce BindML and BindML+ protein-protein interaction sites prediction methods. BindML predicts protein-protein interaction sites by identifying mutation patterns found in known protein-protein complexes using phylogenetic substitution models. BindML+ is an extension of BindML for distinguishing permanent and transient types of protein-protein interaction sites. We developed an interactive web-server that provides a convenient interface to assist in structural visualization of protein-protein interactions site predictions. The input data for the web-server are a tertiary structure of interest. BindML and BindML+ are available at http://kiharalab.org/bindml/ and http://kiharalab.org/bindml/plus/.

Key words Protein-protein interaction, Protein docking, Interface residues, Protein binding site prediction, Bioinformatics, Protein interaction design, Protein interaction propensity

1 Introduction

Protein-protein interactions (PPIs) are critical for mediating many biological functions in the cell. The plethora of knowledge divulged by the complexity of new PPI networks are continuing to be unraveled [1, 2] and tertiary structures of protein complexes are progressively determined and accumulated in databases [3]. However, the rapid accumulation and availability of sequence and structural data for individual proteins makes computational prediction of PPIs, including protein docking structure prediction [4] and prediction of PPI sites [5–7], invaluable for investigating a large number of interacting proteins that do not have solved structures of complexes. PPI site prediction is useful in guiding protein

Ilan Samish (ed.), *Computational Protein Design*, Methods in Molecular Biology, vol. 1529,
DOI 10.1007/978-1-4939-6637-0_14, © Springer Science+Business Media New York 2017

docking prediction [8] and for artificially designing of protein-protein interfaces [9].

We have previously developed BindML (Binding site prediction by Maximum Likelihood), a method to predict protein-protein interaction sites using phylogenetic substitution models [10]. BindML takes a protein structure and multiple sequence alignment (MSA) information to predict protein-protein interaction sites of a given protein surface. Protein-protein interaction site is predicted based on amino acid substitutions observed at a local region around a surface amino acid in question. Through a large performance benchmark, we demonstrated that BindML performed favorably against other existing methods.

Furthermore, we developed an extended framework named BindML+ [11], which utilizes mutation patterns specific for permanent and transient interaction sites to distinguish these two types. Proteins interact with each other with different affinities for specific functional reasons. Some protein pairs, for example oligomeric enzyme complex structures, interact tightly and permanently, while other proteins that are involved in signaling pathways have a mechanism for dissociation after binding, which helps to regulate protein activity at specific times (transient interaction). Distinguishing between the two interaction types provides clues for functions of interacting proteins and has important implications for furthering the understanding of the functional diversity exhibited in PPI networks. Being able to distinguish permanent and transient interaction will be the basis for controlling interaction affinity of designing protein interactions.

BindML and BindML+ are unique in that they use solely interaction site specific mutation patterns, i.e., evolutionary information, in comparison with existing methods that consider features of amino acids, including physicochemical properties [12–15], geometric features of surface shape [15, 16]. Our methods are also unique among methods that use a MSA of a query protein to identify structurally or functionally important regions, because most of such methods are based on the traditional principle that important regions of a protein are conserved in its MSA. BindML and BindML+ use mutation patterns observed in a MSA, i.e., regions in a MSA which do not exhibit apparent conservation and identify hidden structures of mutation events in protein sequences. In this sense, BindML and BindML+ are in common in their philosophy with correlated mutation analyses, which are used for predicting physically contacting residues [17–19] or functional residues [20] in proteins.

In this chapter, we present a web-based graphical user interface for BindML and BindML+ that assists in the prediction and structural visualization of protein-protein interactions sites. The web server provides convenient interactive tools to help identify protein-protein interaction site predictions and to intuitively locate

and associate top scoring predictions to an evaluated protein structure. BindML and BindML+ are freely available online as interactive web servers at http://kiharalab.org/bindml/ and http://kiharalab.org/bindml/plus/.

2 Algorithms

In this section, we briefly explain the essence of the BindML [10] and BindML+ [11] algorithms. For more details, please refer to the original papers.

2.1 BindML Algorithm

A structure of the target protein in the PDB format and corresponding MSA of its family including the target sequence are taken as the input for the BindML algorithm (Fig. 1). The main BindML algorithm starts with generating patches on the protein surface. For each surface residue, a patch is defined as neighboring residues within a 15.0 Å radius sphere. The β-carbon

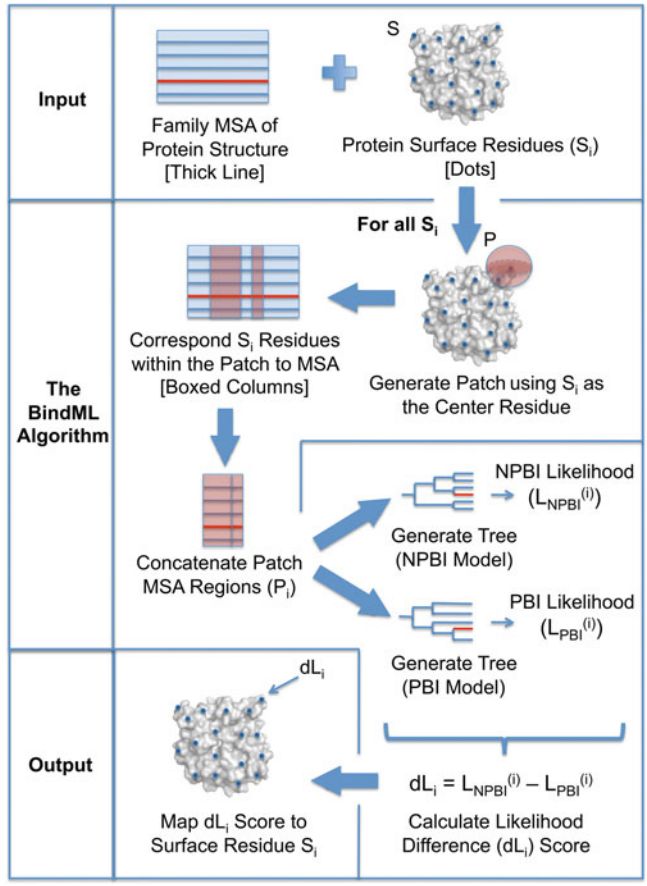


Fig. 1 The steps of the BindML algorithm

of a given amino acid (α-carbon is used for glycine) is selected as the representative point when computing the distance between amino acids. For a patch, all corresponding residues in MSA are concatenated together. The essence of the BindML algorithm is to concatenate surface amino acid residues in a surface patch into a "mini-"MSA (a patch MSA) and judge whether the patch MSA is more likely to occur at protein binding interface or not. A modified version of the PHYML (ver. 2.4.5) program [21] computes the likelihood that a patch MSA comes from protein binding interface (PBI) and non-protein binding interface (NPBI) by constructing phylogenetic trees using amino acid similarity matrices computed for residues at PBI and NPBI, respectively. More concretely, PHYML computes the likelihood of having the input patch MSA following the PBI amino acid similarity matrix (Eq. 1) or NPBI amino acid similarity matrix (Eq. 2) given the initial tree topology. Finally, the difference of the likelihood under PBI and NPBI models provides a score used to predict PPI sites (Eq. 3). For a patch MSA, P_i, which has residue i at the center,

$$L_{\mathrm{NPBI}} = \log\left\{\mathrm{Prob}\left(P_i,\ T_i^{\mathrm{NPBI}} \middle| M_{\mathrm{NPBI}}\right)\right\} \tag{1}$$

$$L_{\mathrm{PBI}} = \log\left\{\mathrm{Prob}\left(P_i,\ T_i^{\mathrm{PBI}} \middle| M_{\mathrm{PBI}}\right)\right\} \tag{2}$$

$$\mathrm{dL} = L_{\mathrm{NPBI}} - L_{\mathrm{PBI}} \tag{3}$$

where M_{NPBI} and M_{PBI} are the amino acid similarity matrices of NPBI and PBI, respectively, and T_i^{NPBI} and T_i^{PBI} are tree generated with M_{NPBI} and M_{PBI}, respectively, for the input patch MSA. The distance likelihood (dL) score is the difference between the log likelihood of the patch MSA being NPBI and PBI. Once all dL scores are calculated, these scores are recast into Z-scores and mapped to corresponding residues. Lower negative scores indicate higher likelihood of PBI mutation patterns, while higher scores show a smaller likelihood.

2.2 BindML+ Algorithm

BindML+ is an extension of BindML, which further predicts whether a predicted PBI site in a query protein performs permanent or transient interaction (Fig. 2). The first step of BindML+ is to predict PBI in the protein surface using BindML as described in the previous section. Then, in the subsequent step, the identified PBI site is classified into either permanent or transient interface. In the first step, once dL scores (Eq. 3) for all surface patches are calculated, these scores are recast into Z-scores and a threshold (0 is used for the website) is placed. A lower (negative) Z-score indicates larger likelihood of PBI mutation patterns. In BindML+, any center residue of a patch with a score that is equal to or smaller than the given threshold value is included in a PBI site for the subsequent step.

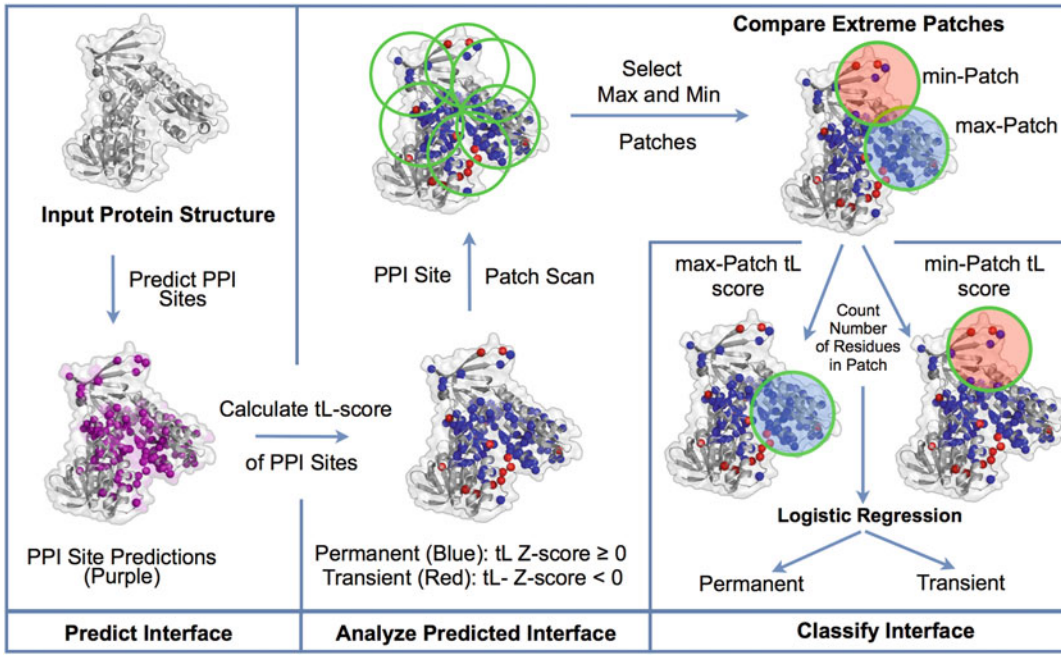


Fig. 2 The steps of BindML+ algorithm

Next, each predicted PBI residue is classified into either permanent or transient, using an amino acid substitution matrix computed from MSAs observed at permanent interaction sites (PERM) and another matrix computed from MSAs at transient interaction sites (TRAN). These two matrices capture characteristic amino acid substitution patterns at permanent and transient interaction sites, respectively. Using these two matrices, similar to Eq. 2, the likelihood that each patch-MSA centered at residue i in a predicted PBI site is from permanent or transient interface ($L_{\mathrm{PERM}}(i)$ and $L_{\mathrm{TRAN}}(i)$) is computed, respectively, and the difference between $L_{\mathrm{PERM}}(i)$ and $L_{\mathrm{TRAN}}(i)$ score, which is named the interface type likelihood (tL) score, is computed:

$$\mathrm{tL}(i) = L_{\mathrm{PERM}}(i) - L_{\mathrm{TRAN}}(i) \tag{4}$$

For a residue with a tL Z-score above zero it is more likely to be permanent, whereas a lower value below zero suggests that it is more likely to be transient.

Then, BindML+ will discriminate the interaction type of the query protein into either the permanent or the transient type using a logistic regression model (LRM). LRM is a binary classifier that tries to fit a set of features using a logit function. Features used in the LRM are based on the tL score and additional related scores of residues at the predicted PBI site. For the details, please refer to the original paper [11].

3 Input and Output of the Servers

3.1 Input Data

Both BindML and BindML+ need four input data to execute. Figure 3 shows the screen capture of the input windows at the top page of BindML.

1. User's email address: An email address is needed for receiving notifications when a submission is processed and completed. An email with the result page URL will be sent to this address.
2. A target PDB file: A query of BindML/BindML+ is a protein tertiary structure in the PDB format. The PDB file can contain chains that are not the target of binding site prediction because in the next step the chain ID of the target will be specified.
3. Specify a chain ID: users need to specify the chain ID in the PDB file. If there are no chain IDs in your PDB, put the underscore "_" instead.
4. Upload a MSA file: This is an optional input to use when users want to use their own MSA. A MSA file to upload must be in the FASTA format and the query PDB sequence is needed to be included in the MSA. Example input files are provided at the bottom of the submission page. If the MSA filed is left empty, the server will execute a search against the Pfam database [22]. Two Pfam databases will be searched. First, Pfam-A will be searched and Pfam-B will only be searched when Pfam-A does not return a match to the query protein sequence. If neither of them matches, an HMMER search [23] will be used to include weak matches from the Pfam database. Finally, the server automatically generates the MSA with the MUSCLE multiple sequence alignment program [24] using the full-length sequences of proteins included in the retrieved Pfam profile.

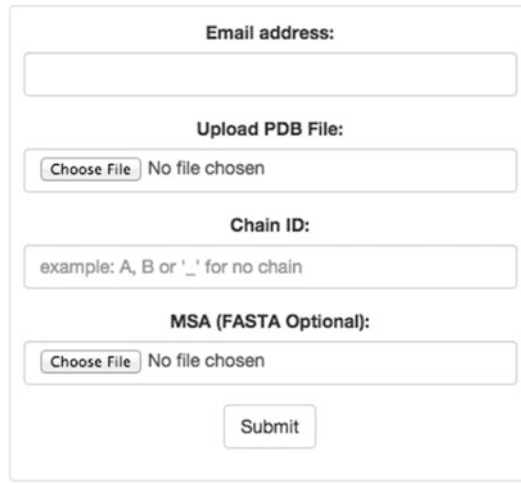


Fig. 3 Input data submission window for BindML

3.2 Output Page with Case Studies

After a submission, an email will be sent to the user when the computation is completed, which includes a link to the result page of the query. Computatation takes typically a few minutes but can take longer depending mainly on the size (length) of the query protein and the number of sequences in the MSA of the query. Below we explain how the results are presented.

3.2.1 BindML Output Page

The interactive BindML result page consists of an integrative structural-level view and a residue-level table with associated prediction scores (Fig. 4). The left panel shows the query protein structure with the JSmol structure viewer (http://wiki.jmol.org/index.php/Jmol_Javascript_Object), where residues are colored based on the *Z*-score of the dL score (Eq. 3). The color ranges from red to blue, where red indicates strong predictions of binding site residues while blue represents residues predicted to be at non-protein binding surface. To visualize prediction, the dL *Z*-score of each residue is written at the B-factor column in the PDB file of the query protein, and JSmol colors residues by reading the dL *Z*-score as B-factor values. The modified PDB file can be downloaded. The last line of the structure panel shows the MSA found for the query protein in the Pfam database. The structure can be rotated and zoomed in/out. All options of different visualization offered by JSmol are available by right click on the structure panel.

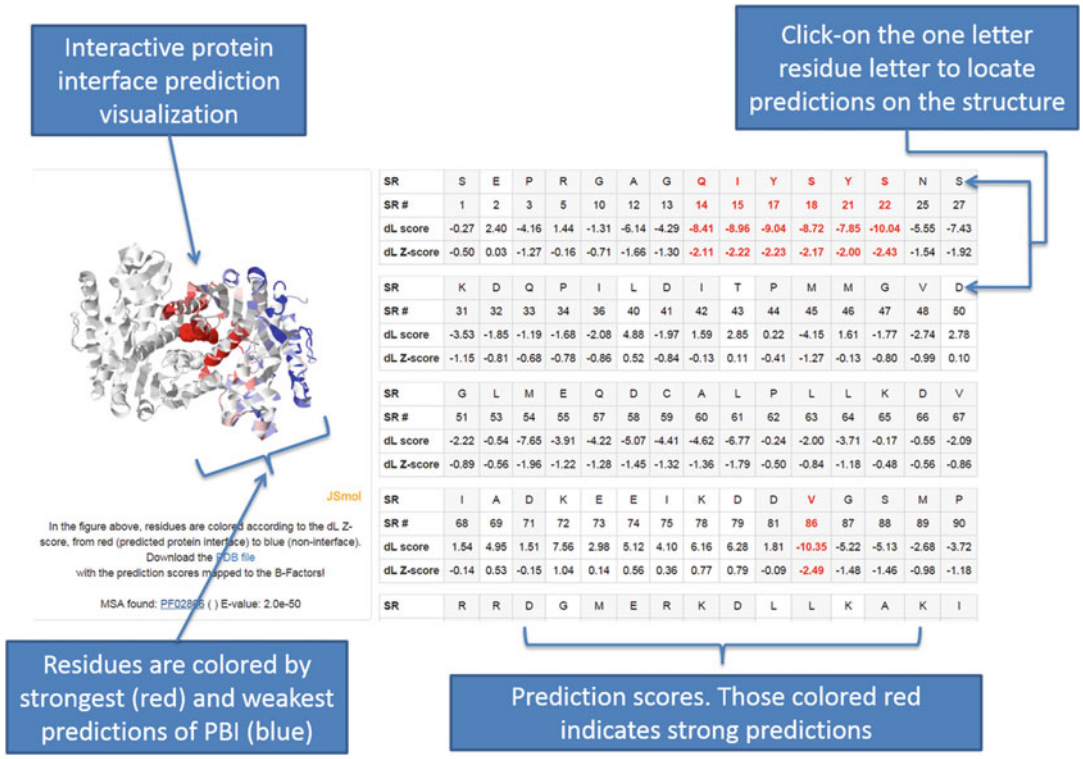


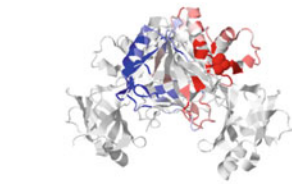
Fig. 4 Example of result page of BindML. PDB entry, 4MDH chain A was used

The right panel shows the detailed list of prediction scores for each residue. Only surface residues are listed. The third row, the dL score, corresponds to Eq. 3, and the final prediction is determined using the dL *Z*-score. Predicted protein binding site residues, i.e., residues that have negative dL *Z*-scores, are highlighted in gray, and those with a high confident score, i.e., dL *Z*-score < -2.0, are colored in red. Amino acid residues in the first row can be clicked and mapped on the left panel with a volume representation.

The example shown in Fig. 4 is the prediction computed for cytoplasmic malate dehydrogenase, A-chain (PDB code: 4MDH). The structure panel visualizes structures of a complex of chains A and B that are contained in the PDB file, but the prediction was computed only for chain A, the colored chain on the right-hand side, without considering the docking conformation with chain B. Apparently, the prediction captures protein binding interface residues of chain A very well with high confidence (red) and surface residues that are not involved in interaction are correctly captured (blue). The area under the curve (AUC) value of this prediction is 0.826. In the left panel, TYR17 is shown in a volume representation, which was invoked by clicking the residue in the table. The MSA used for this prediction is a Pfam entry, PF02866, as indicated at the bottom of the left panel. The match of the query to the Pfam entry is significant with a very low *E*-value of 2.0×10^{-50}. The entry ID is linked to its Pfam page where users can retrieve sequences in the MSA and related information. PF02866 is a MSA for lactate/malate dehydrogenase, alpha/beta C-terminal domain, which agrees with the name of the query protein.

3.2.2 *BindML+ Output Page*

The BindML+ result page essentially shares the same layout with the BindML result page (Fig. 5). The additional information



JSmol

In the figure above, residues are colored according to the tL Z-score, from red (predicted permanent interaction) to blue (predicted transient interaction). Download the PDB file with the predicted scores (dL Z-scores) mapped to the B-Factors and tL z-scores mapped to the occupancy!

MSA found: PF05543 (Peptidase_C47) E-value: 1.1e-85

SR	E	V	S	E	Q	D	P	N	C	A	T	F	P	N	Q
SR #	273	274	275	276	277	278	280	281	282	283	284	285	286	287	288
dL score	-0.07	-0.44	0.25	-0.65	-0.23	-0.49	-0.91	-1.31	-1.00	-2.41	-1.37	-1.86	-0.98	-0.93	-1.63
dL Z-score	-0.33	-0.60	-0.11	-0.75	-0.45	-0.63	-0.93	-1.21	-0.99	-2.00	-1.26	-1.60	-0.98	-0.94	-1.44
tL score	-0.78	0.04	-0.58	-0.13	-0.39	0.11	0.35	0.59	0.61	1.06	0.28	0.68	0.43	-0.51	1.30
tL Z-score	0.11	0.85	0.29	0.70	0.46	0.91	1.13	1.35	1.36	1.77	1.07	1.43	1.20	0.35	1.99

SR	I	E	K	S	Q	G	R	D	H	Y	Q	E	G	V	S
SR #	290	291	294	295	296	297	298	299	301	302	303	304	305	306	308
dL score	0.14	-0.63	-0.34	0.32	0.49	-0.06	0.35	0.34	1.61	0.51	-0.07	-0.74	-0.51	0.93	2.97
dL Z-score	-0.19	-0.73	-0.53	-0.06	0.06	-0.33	-0.04	-0.04	0.86	0.08	-0.33	-0.81	-0.65	0.38	1.82
tL score	-0.81	0.20	-0.19	-0.07	-0.62	0.19	-0.70	-0.65	-1.34	-1.47	-0.78	-0.65	-0.58	-2.75	-2.98
tL Z-score	0.08	0.99	0.64	0.75	0.25	0.98	0.18	0.23	-0.40	-0.51	0.11	0.23	0.29	-1.67	-1.88

SR	Y	N	Q	D	Q	L	K	D	N	V	Q	S	V	S	Q
SR #	309	310	311	313	314	315	317	318	319	320	327	328	329	330	331
dL score	2.92	2.27	1.45	3.49	2.22	1.65	1.90	2.20	2.47	1.97	-0.94	-0.61	0.04	0.11	1.04
dL Z-score	1.79	1.33	0.74	2.19	1.29	0.89	1.06	1.28	1.47	1.11	-0.95	-0.72	-0.26	-0.21	0.45
tL score	-2.98	-2.52	-1.69	-3.64	-2.05	-1.69	-1.54	-1.68	-1.39	-1.86	0.14	1.18	-0.06	-0.98	-1.19
tL Z-score	-1.88	-1.46	-0.71	-2.47	-1.04	-0.71	-0.58	-0.70	-0.44	-0.87	0.94	1.88	0.76	-0.07	-0.26

Fig. 5 Example of BindML+ prediction

predicted by BindML+, prediction of permanent or transient interactions, is shown in the two more rows (tL score and tL *Z*-scores) in the right-hand side table. As described in Subheading 2.2, predicted interface resisdues are classified into either permanent or transent types by their tL *Z*-scores. Residues with tL *Z*-scores greater than or equal to zero correpond to permanent binding site predictions, while tL *Z*-scores below zero represent transient binding site predictions. In the table, predicted protein binding site residues, which have negative dL *Z*-score values, are shown in gray columns. When the binding site predictions are confident (dL *Z*-score < -2.0), letters in the colums are colored in the same way as the BindML output page. Classification of permanent and transient interactions for residues at predicted binding sites (i.e., those highlighted in gray columns) is colored with red or blue, for permanent or transient interactions, respectively.

The top of the BindML+ result page shows the overall prediction of interaction types, either permanent or transient with a confidence score. Note that this overall classification of interaction is computed using information of predicted interaction types of individual residues thus, it is possible that individual residues have different predicted types than the overall interaction type. The overall classification has a score that ranges from 0.0 to 1.0 with 1.0 being the highest confidence. This score is based on the output of the logistic regression used in the interaction type classification.

In a BindML+ page, the structural view on the left of the page visualizes predicted interaction types of individual residues in colors, permanent (red) to transient (blue), according to the tL *Z*-score in the table. The source of the visualized structure in the PDB format, which can be downloaded, has the predicted binding interface scores (dL *Z*-scores) mapped to the B-Factors and the interaction type score, tL *Z*-scores, mapped to the occupancy field.

The protein used as an example in Fig. 5 is staphostatin-staphopain complex (PDB ID: 1pxv). This protein has a permanent interaction. BindML+ correctly predicted its interaction type as permanent with a score of 0.274. The structure panel in Fig. 4 shows binding interface residues of staphostatin (chain on the left) to its inhibitor, staphopain (smaller gray structure on the right side of the complex), are almost all predicted to have permanent interaction (red), while the opposite side of the residues is predicted to have transient interaction properties (blue). Ala283 is emphasized in volume representation. This is a successful example of prediction with the area under the curve (AUC) value of 0.84 for binding residue prediction with 63 predicted binding interface residues out of 73 predicted to have permanent interaction.

4 Conclusion

BindML and BindML+ provide prediction of residues at protein binding interface for a query protein structure entirely from evolutionary information embedded in the MSA of the protein. The algorithms are based on a novel idea of constructing a phylogenetic tree of mini-MSA of local surface regions of the query protein. The performance of these two methods were rigorously benchmarked and compared favorably with related existing methods [10, 11]. The methods can be applied for experimentally solved high-resolution structures, computationally modeled structures, and artificially designed proteins. Also, the methods will be useful in designing protein-protein interactions at desired sites in the query protein and controlling strength of interactions.

Acknowledgments

The authors thank Lyman Monroe for proofreading the manuscript. This work has been supported by grants from the National Institutes of Health (R01GM075004 and R01GM097528), National Science Foundation (IIS1319551, DBI1262189, IOS1127027), and National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-220-C00004).

References

1. Rajagopala SV, Sikorski P, Kumar A, Mosca R, Vlasblom J, Arnold R, Franca-Koh J, Pakala SB, Phanse S, Ceol A, Hauser R, Siszler G, Wuchty S, Emili A, Babu M, Aloy P, Pieper R, Uetz P (2014) The binary protein-protein interaction landscape of Escherichia coli. Nat Biotechnol 32(3):285–290. doi:10.1038/nbt.2831
2. Hauser R, Ceol A, Rajagopala SV, Mosca R, Siszler G, Wermke N, Sikorski P, Schwarz F, Schick M, Wuchty S, Aloy P, Uetz P (2014) A second-generation protein-protein interaction network of Helicobacter pylori. Mol Cell Proteomics 13(5):1318–1329. doi:10.1074/mcp.O113.033571
3. Mosca R, Ceol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. Nat Methods 10(1):47–53. doi:10.1038/nmeth.2289
4. Venkatraman V, Yang YD, Sael L, Kihara D (2009) Protein-protein docking using region-based 3D Zernike descriptors. BMC Bioinformatics 10:407. doi:10.1186/1471-2105-10-407
5. La D, Kihara D (2008) Predicting binding interfaces of protein-protein interactions. In: Li XL, Ng SK (eds) Biological data mining in protein interaction netrworks. IGI-Global, Hershey, PA, pp 64–79
6. Zhou HX, Qin S (2007) Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics 23(17):2203–2209. doi:10.1093/bioinformatics/btm323
7. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A (2005) Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 61(Suppl 7):27–45
8. Li B, Kihara D (2012) Protein docking prediction using predicted protein-protein interface. BMC Bioinformatics 13:7. doi:10.1186/1471-2105-13-7
9. Fleishman SJ, Whitehead TA, Strauch EM, Corn JE, Qin S, Zhou HX, Mitchell JC,

Demerdash ON, Takeda-Shitaka M, Terashi G, Moal IH, Li X, Bates PA, Zacharias M, Park H, Ko JS, Lee H, Seok C, Bourquard T, Bernauer J, Poupon A, Aze J, Soner S, Ovali SK, Ozbek P, Tal NB, Haliloglu T, Hwang H, Vreven T, Pierce BG, Weng Z, Perez-Cano L, Pons C, Fernandez-Recio J, Jiang F, Yang F, Gong X, Cao L, Xu X, Liu B, Wang P, Li C, Wang C, Robert CH, Guharoy M, Liu S, Huang Y, Li L, Guo D, Chen Y, Xiao Y, London N, Itzhaki Z, Schueler-Furman O, Inbar Y, Potapov V, Cohen M, Schreiber G, Tsuchiya Y, Kanamori E, Standley DM, Nakamura H, Kinoshita K, Driggers CM, Hall RG, Morgan JL, Hsu VL, Zhan J, Yang Y, Zhou Y, Kastritis PL, Bonvin AM, Zhang W, Camacho CJ, Kilambi KP, Sircar A, Gray JJ, Ohue M, Uchikoga N, Matsuzaki Y, Ishida T, Akiyama Y, Khashan R, Bush S, Fouches D, Tropsha A, Esquivel-Rodriguez J, Kihara D, Stranges PB, Jacak R, Kuhlman B, Huang SY, Zou X, Wodak SJ, Janin J, Baker D (2011) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol 414 (2):289–302. doi:10.1016/j.jmb.2011.09.031

10. La D, Kihara D (2012) A novel method for protein-protein interaction site prediction using phylogenetic substitution models. Proteins 80(1):126–141. doi:10.1002/prot.23169
11. La D, Kong M, Hoffman W, Choi YI, Kihara D (2013) Predicting permanent and transient protein-protein interfaces. Proteins 81 (5):805–818. doi:10.1002/prot.24235
12. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R (2005) Optimal docking area: a new method for predicting protein-protein interaction sites. Proteins 58(1):134–143. doi:10.1002/prot.20285
13. Xu D, Tsai CJ, Nussinov R (1997) Hydrogen bonds and salt bridges across protein-protein interfaces. Protein Eng 10(9):999–1012
14. Tjong H, Qin S, Zhou HX (2007) PI2PE: protein interface/interior prediction engine. Nucleic Acids Res 35(Web Server issue): W357–W362. doi:10.1093/nar/gkm231
15. Jones S, Thornton JM (1997) Analysis of protein-protein interaction sites using surface patches. J Mol Biol 272(1):121–132. doi:10.1006/jmbi.1997.1234
16. Jones S, Thornton JM (1997) Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 272(1):133–143. doi:10.1006/jmbi.1997.1233
17. Morcos F, Hwa T, Onuchic JN, Weigt M (2014) Direct coupling analysis for protein contact prediction. Methods Mol Biol 1137:55–70. doi:10.1007/978-1-4939-0366-5_5
18. Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28(2):184–190. doi:10.1093/bioinformatics/btr638
19. Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18(4):309–317
20. Kuipers RK, Joosten HJ, Verwiel E, Paans S, Akerboom J, van der Oost J, Leferink NG, van Berkel WJ, Vriend G, Schaap PJ (2009) Correlated mutation analyses on super-family alignments reveal functionally important residues. Proteins 76(3):608–616. doi:10.1002/prot.22374
21. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52 (5):696–704
22. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42(Database issue):D222–D230. doi:10.1093/nar/gkt1223
23. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39(Web Server issue):W29–W37. doi:10.1093/nar/gkr367
24. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32 (5):1792–1797