Bioinformatics
doi.10.1093/bioinformatics/xxxxxx

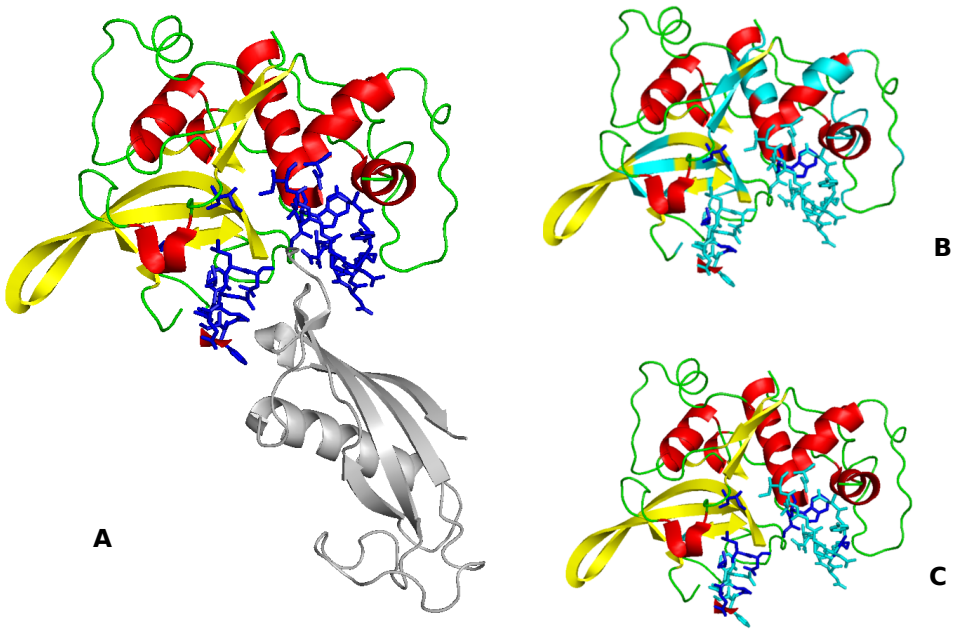
OXFORD

Supplementary Material

SeRenDIP: SEquential REmasteriNg to DerIve Profiles for fast and accurate predictions of PPI interface positions

Qingzhen Hou, Paul De Geest, Christian Griffioen, Sanne Abeln, Jaap Heringa, K. Anton Feenstra *

* To whom correspondence should be addressed: k.a.feenstra@vu.nl.



SI Figure 1. Interface prediction example for PDB 1YVB-A. (A) Cartoon representation of the heterodimeric protein PDB 1YVB (chain A:I). Falcipain is shown in colors and cystatin in white. The interface is indicated in blue sticks. (B) Chain A with all interface sites predicted by our server in cyan. (C) Chain A with the correctly predicted interface sites in cyan. Coverage is 74% and precision 32%. Overall prediction for this target yielded an AUC-ROC of 0.788 and an F1 score of 0.314.

In this supplementary material, we will sumarize the procedure of training new predictors and provide the details of speeding up our previous approach (Hou *et al.*, 2017). We also compare the runtime and accuracy between our webserver and the 'Old' approach.

1 New Predictors

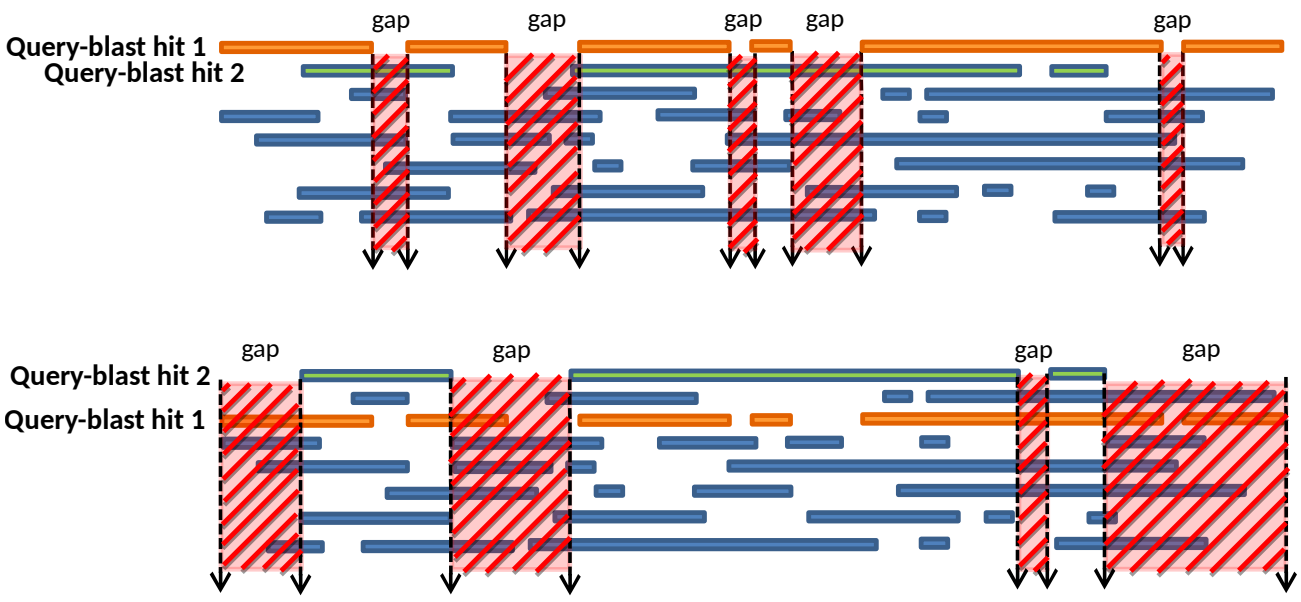
In this research, we use the same sets of proteins as our previous paper (Hou *et al.*, 2017). For clarity, we here briefly summarize the training and testing protocols used to obtain our Random Forest classifiers. The homomeric hm_479 dataset (Hou *et al.*, 2015) was split into 60% training, 20% validation and 20% testing. Datasets Dset_119 and Dset_48 from Murakami and Mizuguchi (2010) were used for heteromeric training and testing. Based on the different training datasets, we train three predictors: RF-hetero (Dset_119), RF-homo (hm_479 training) and RF-combined (hm_479 training and Dset_119).

2 Overall Performance Increase: Speed

The main bottleneck in runtime for the Hou et al. (2017) implementation, is NetSurfP (Petersen *et al.*, 2009) generating profiles (PSSM) for the prediction of solvent accessibility and secondary structure. For the input query sequence, PSI-Blast (Altschul *et al.*, 1997, 2005; Schäffer *et al.*, 2001) is run to obtain homologs from the NR70 database, using a maximum of 3 iterations, a reporting E-value threshold of 10^{-5} and a maximum of 500 hits. All other parameters are used at default value. From the resulting PSI-blast profiles, NetSurfP predicts relative and absolute solvent accessibility (RSA and ASA, resp.) and secondary structure (PA, PB and PC for α-helix, β-sheet and coil propensity, resp.). Additionally, for all the hits, the full length sequences are retrieved from the database and aligned using the Muscle multiple sequence alignment (MSA) tool (Edgar, 2004). From this alignment, sequence entropy values are calculated for each position of the input query sequence.

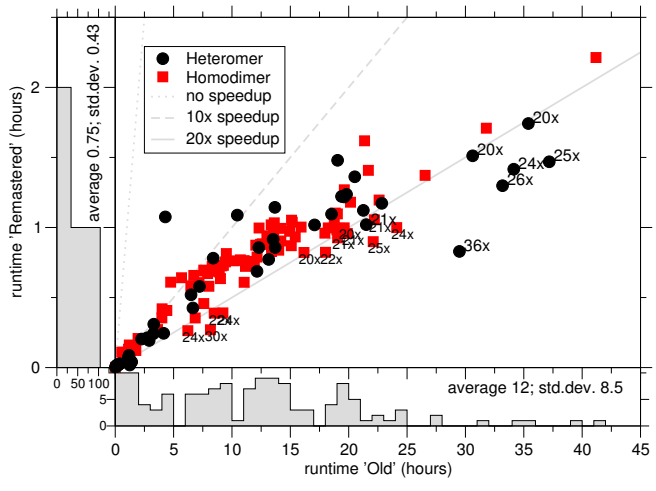
In the previous approach Hou *et al.* (2017), for each of the PSI-Blast hits, NetSurfP would be run, each time creating a PSI-Blast profile for the current (hit) sequence. In addition to the feature values for the input query sequence, features were also generated for a selection of homologs of the query (from PSI-Blast). This means that for each of the homologs (typically a few hundred), another run of PSI-Blast is required. Such runtimes makes it prohibitively expensive for use as a webserver, both because of the investement in CPU time, and because such a delay will be too long for many types of usage.

Here, we explore the insight that (close) homologs of homologs of the query, are mostly also (close) homologs to the query itself. Hence, the useful signal for the predictor may already be present in the homologs of the query. The 'Remastered' approach, takes the MSA of the query sequence and its hits, created by Muscle. Each of the sequences is in turn taken as a 'master' sequence, and by removing all positions from the alignment where there are gaps in the master sequence, a new master-slave alignment is created. This process is known as 'remastering' of a multiple-sequence alignment; This is illustrated in SI figure 2. Depending on the gap distribution in the master sequence, a different master-slave alignment is created for each of the sequences in the alignment. For each of the master-slave alignments a profile (PSSM) is created, which differ only by their selection of columns to retain. The PSSMs are generated by the 'alignment restart' module of PSI-Blast, which reads an MSA and writes a PSSM, taking the first sequence in the alignment as the master sequence. Then, each of the PSSMs are given to NetSurfP for prediction of solvent accessibility and secondary structure. SI figure 3 shows a comparison of runtimes for the 'Remastered' profile generation dropping from one to two days, down to usually less than one hour. The average gain



SI Figure 2. The concept of 'Remastered' profile generation. Choosing one sequence in an MSA as 'master', subsequently all columns are removed from the alignment where there are gaps in the master sequence, indicated by the red hashed 'gap' areas. From the remaining columns, the profile is generated. The top panel shows the master-slave generation when choosing Query blast-hit 1 as master sequence. The bottom panel shows the master-slave for blast-hit 2. These sequences are clearly different in their gap distribution in the alignment, and hence lead to different profiles being created; in practice differences may be more subtle.

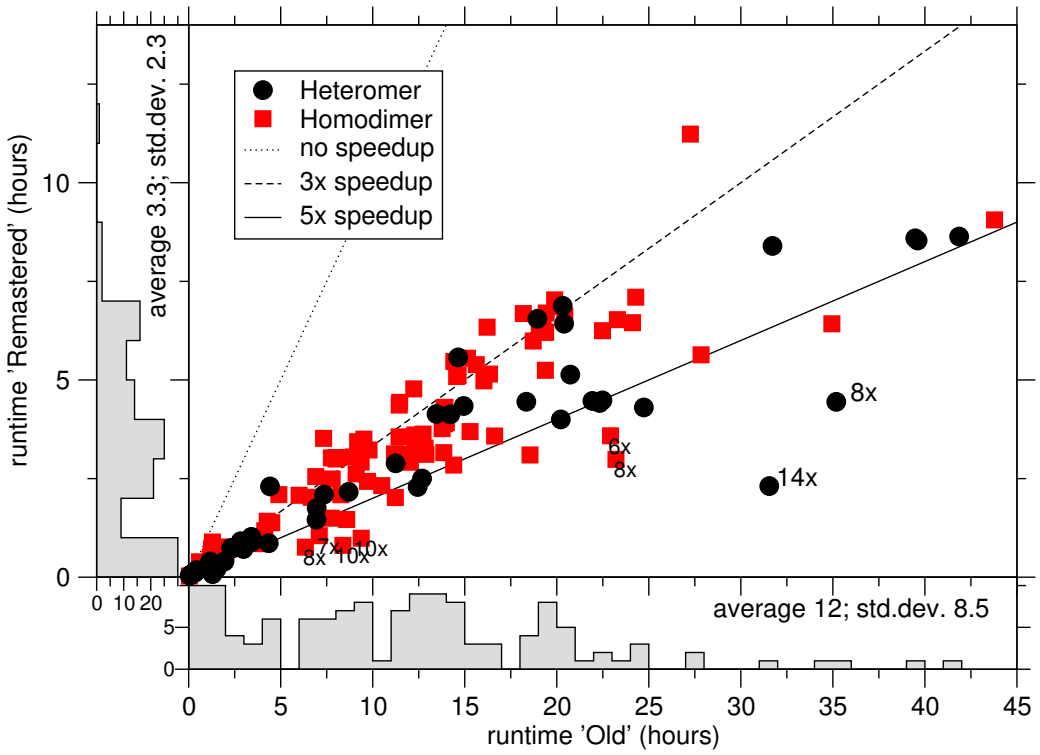
was 15-fold (± 7). Overall, to predict the interface of each protein in our training dataset, average runtimes were reduced about four-fold (± 2), as shown in SI figure 3.



SI Figure 3. Comparison of profile generation runtimes between the 'Old' and 'Remastered' approach on a single CPU. The main panel shows the correlation between runtimes of the profile generation in both approaches. Results for the heteromeric test-set are shown in black circles, for the homomeric test-set in red squares. Dotted, dashed and drawn lines show no, 10-fold and 20-fold speedup, respectively. Grey bar charts on both axes show the respective distribution of runtimes. Average speedup is 15-fold.

3 Overall Performance Increase: Accuracy

We then compared the performance measures with those obtained for the 'Old' models of Hou *et al.* (2017). This is summarized in SI Table1. The 'Remastered' RF models perform better or equal in almost all cases on almost all measures. ROC and the Precison/Recall plots are SI figure 4 and the overlap of predicted and true sites between methods is in SI figure 5.



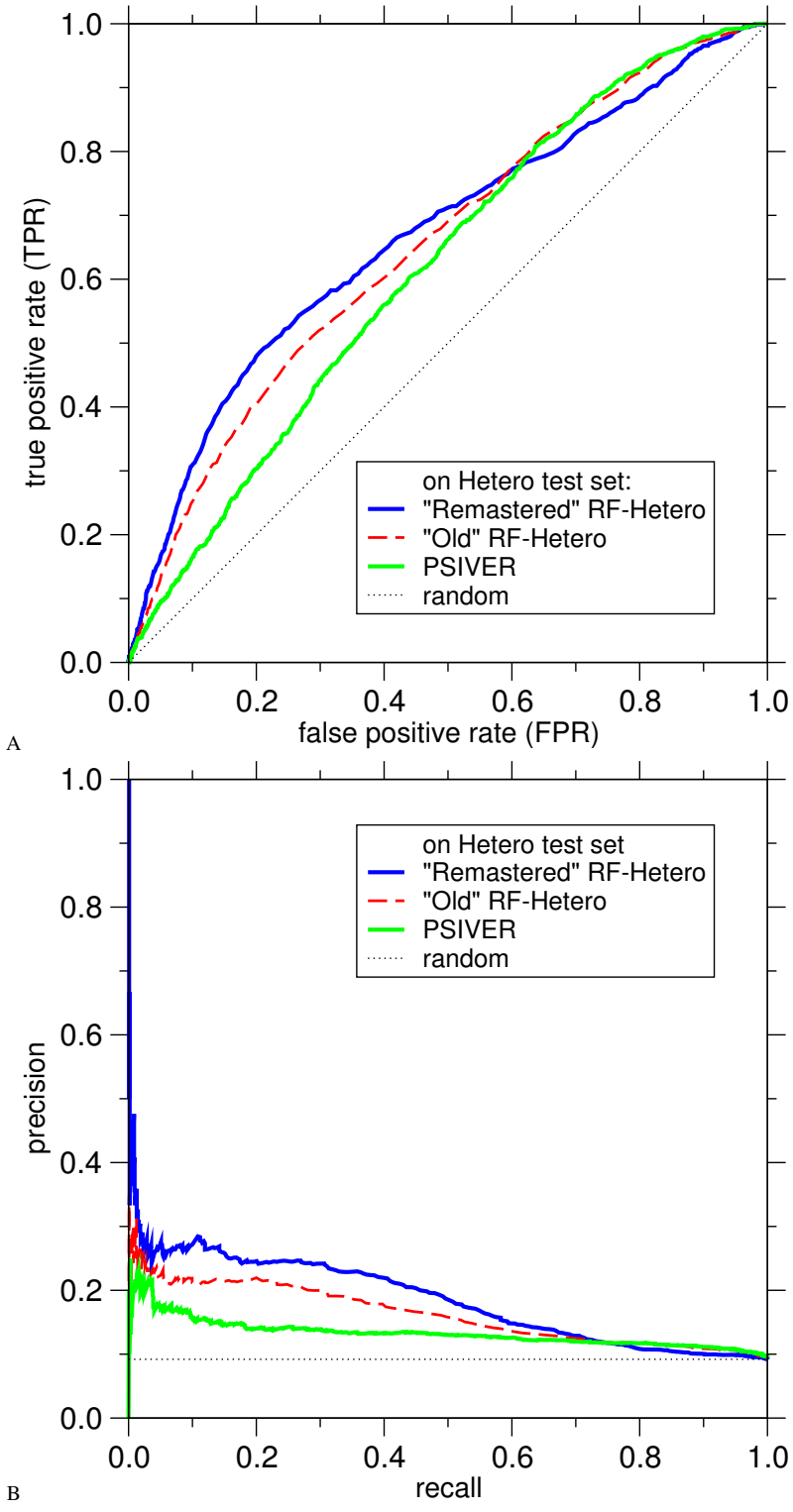
SI Figure 4. Comparison of runtimes between the 'Old' and 'Remastered' approach. The main panel shows the correlation between overall runtimes in both approaches. Results for the heteromeric test-set are shown in black circles, for the homomeric set in red squares. Dotted, dashed and drawn lines show no, 3-fold and 5-fold speedup, respectively. Grey bar charts on both axes show the respective distribution of runtimes.

4 The Webserver

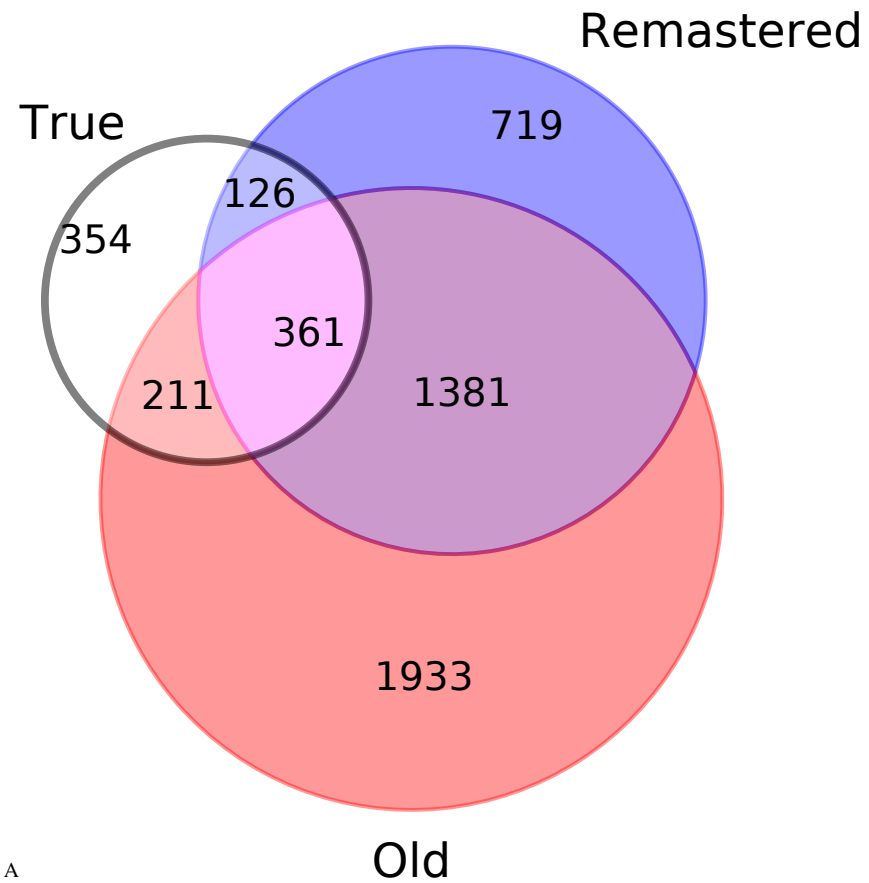
Our 'Remastered' profile approach for feature generation, which avoids the recursive calling of PSI-Blast on all homologs of the input query sequence, yielded a speed-up of more than ten-fold for the profile generation, and over four-fold overall, bringing it down to a practical level of a few hours. Surprisingly, the RF models trained on the features generated with the new approach, achieved a small performance improvement, compared to the previous approach. Together, this now allows the method to be practically usable, and we make it available both as source code and webserver for easy access.

Training set	Test set	Features	MCC	F1	Precision	Recall	Specificity	AUC ROC
Combined	Heteromeric	'Old'	0.122	0.226	0.146	0.500	**0.695**	0.636
		'Remastered'	**0.135**	0.228	0.142	**0.589**	0.639	**0.655**
Heteromeric		'Old'	0.131	0.230	0.146	**0.547**	0.667	0.652
		'Remastered'	**0.193**	**0.278**	**0.196**	0.480	**0.800**	**0.668**
Homomeric		'Old'	0.103	0.213	**0.140**	0.446	**0.716**	0.619
		'Remastered'	**0.110**	0.213	0.132	**0.554**	0.630	**0.625**
Combined	Homomeric	'Old'	0.277	0.462	0.383	0.581	**0.734**	0.724
		'Remastered'	**0.289**	**0.474**	0.379	**0.633**	0.703	**0.732**
Heteromeric		'Old'	0.064	0.297	0.263	**0.343**	0.727	0.552
		'Remastered'	**0.085**	0.299	**0.284**	0.316	**0.772**	**0.568**
Homomeric		'Old'	0.265	0.454	0.373	0.581	**0.722**	0.720
		'Remastered'	**0.286**	**0.472**	0.376	**0.632**	0.700	**0.731**

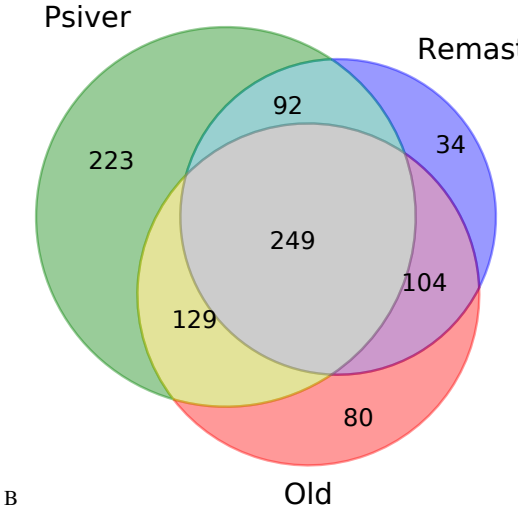
Table 1. **Benchmark results for the 'Old' and 'Remastered' RF models.** RF models were trained on the homomeric, heteromeric, and both (combined) training sets, and were tested on the homomeric and heteromeric test sets. For both Training and Testing, the 'Old' (Hou et al., 2017) or 'Remastered' feature sets were used, as indicated. For each train/test set combination, the highest score for each benchmark criterium is shown in bold; differences of less than 0.005 are not highlighted.



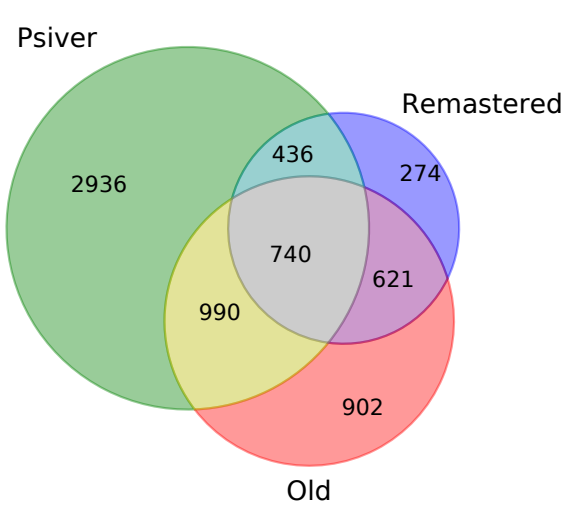
SI Figure 5. Benchmark of the RF interface prediction models. Performance of the 'Remastered' and 'Old' RF-Hetero models compared with that of PSIVER, on the heteromeric test-set using (A) ROC and (B) a P/R plots. The 'Remastered' Hetero model is shown in blue, the 'Old' model in dashed red and PSIVER in green.



A



B



C

SI Figure 6. Overlap of predicted and true sites. (A) Overlap of predicted sites of the 'Remastered' and 'Old' RF-Hetero models and true interface sites according to PISA, on the heteromeric test-set using a Venn diagram. The 'Remastered' Hetero model is shown in blue, the 'Old' model in red and true sites in white. (B) For the correctly predicted sites (true positives), the overlap between 'Remastered' and 'Old' RF-Hetero models and Psiver (shown in green). (C) Like B, for the incorrectly predicted sites (false positives).

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**(17), 3389–402.

Altschul, S. F., Wootton, J. C., Gertz, E. M., Agarwala, R., Morgulis, A., Schaffer, A. A., and Yu, Y.-K. (2005). Protein database searches using compositionally adjusted substitution matrices. *FEBS Journal*, **272**(20), 5101–5109.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**(1).

Hou, Q., Dutilh, B., Huynen, M., Heringa, J., and Feenstra, K. (2015). Sequence specificity between interacting and non-interacting homologs identifies interface residues - a homodimer and monomer use case. *BMC Bioinformatics*, **16**(1).

Hou, Q., De Geest, P., Vranken, W., Heringa, J., and Feenstra, K. (2017). Seeing the trees through the forest: Sequence-based homo- and heteromeric protein-protein interaction sites prediction using random forest. *Bioinformatics*, **33**(10).

Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M., and Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC structural biology*, **9**(1), 1.

Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., and Altschul, S. F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic acids research*, **29**(14), 2994–3005.