Supplementary Material S7

The negative data set

Our gold standard consists of curated interaction pairs found in the DIP database; it does not contain negative interaction pairs. The negative interaction pairs have been randomly generated at a proportion of five times the number of positive interaction pairs. Of these negative interaction pairs, only 3.18% (10,848/341,357) were mapped to the pDBs and, therefore, had metrics extracted from the Blast+ alignments. The remaining 96.82% (330,509/341,357) did not map to the pDBs (Table 1). Because these random pairs have negative interactions, it is natural to expect that they are not going to be mapped to the pDBs. This shows that the random selection of negative interaction pairs is efficient for use in studies of interaction network prediction through ortholog interaction mapping. Moreover, regarding the 3.18% mapped negative interactions, the hypothesis that these are biologically true interaction pairs found in the pDBs that have not yet been included in our gold standard (DIP) can be raised, but further investigation is required to confirm it.

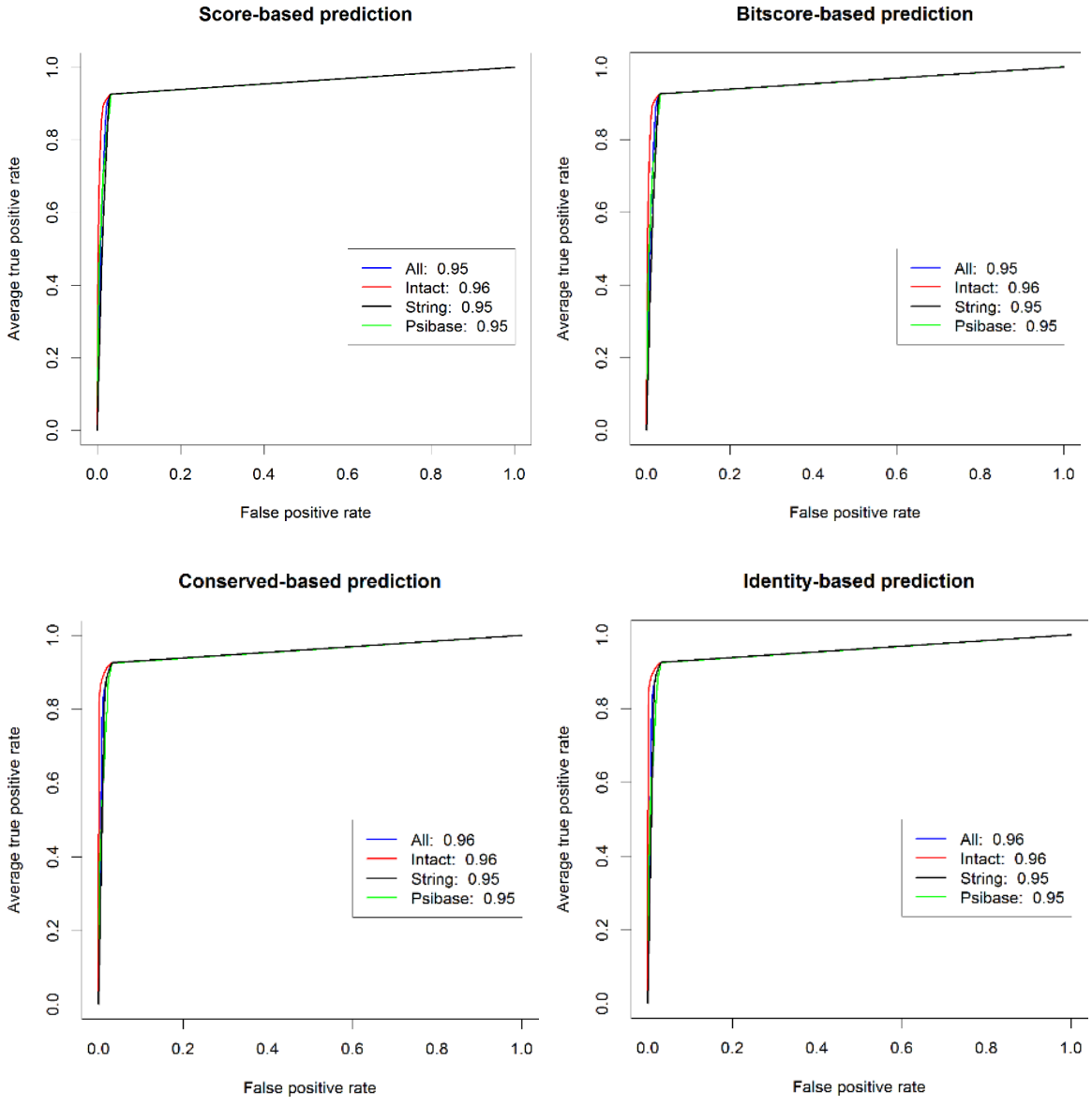
Table 1 - Distribution of DIP interaction pairs in pDBs

DIP	Negative	Positive	Negative(*)	Positive(*)
String	6,584	43,216	200,597	3,457
Intact	4,212	50,116	128,328	4,020
Psibase	52	1,910	1,584	153
Total interaction pairs	10,848	95,242	330,509	7,630

(*) Number of DIP interaction pairs that were not mapped to any pDB

To validate our metrics, we can use the negative data sets of our gold standard (DIP) that were mapped (10848+95242) or not mapped (330509+7630) to any pDB in two different ways: (i) using only the interaction pairs that map ortholog interaction pairs to the pDBs, or (ii) considering the biological point of view that the negative set really should not contain ortholog interaction pairs in the pDB, use this negative set, proportionally distributing the non-mapped interaction pairs among the pDBs. However, for those interaction pairs that were not mapped to the pDBs, it was not possible to extract any values from the Blast+ alignment, and therefore, the null value (0) was attributed to each metric for these pairs.

The previous analyzes (Suplementary Material Figure S3 and Suplementary Material Figure S4) were generated using only the interaction set that was mapped to the pDBs. By also using the non-mapped dataset, we generated ROC curves for each metric/pDB (Figure 1). The bias that these unmapped interaction pairs introduce into the ROC curves is observable, especially the negative interaction set which, by being the majority (330,509) and having null value (0), influences the metrics, causing any metric evaluated to yield good results (Figure 1). Any cut-off point that is above zero (0), for any metric/pDB, would yield good results, showing that to have good, realistic metrics, the unmapped dataset should not be included in the analyses.



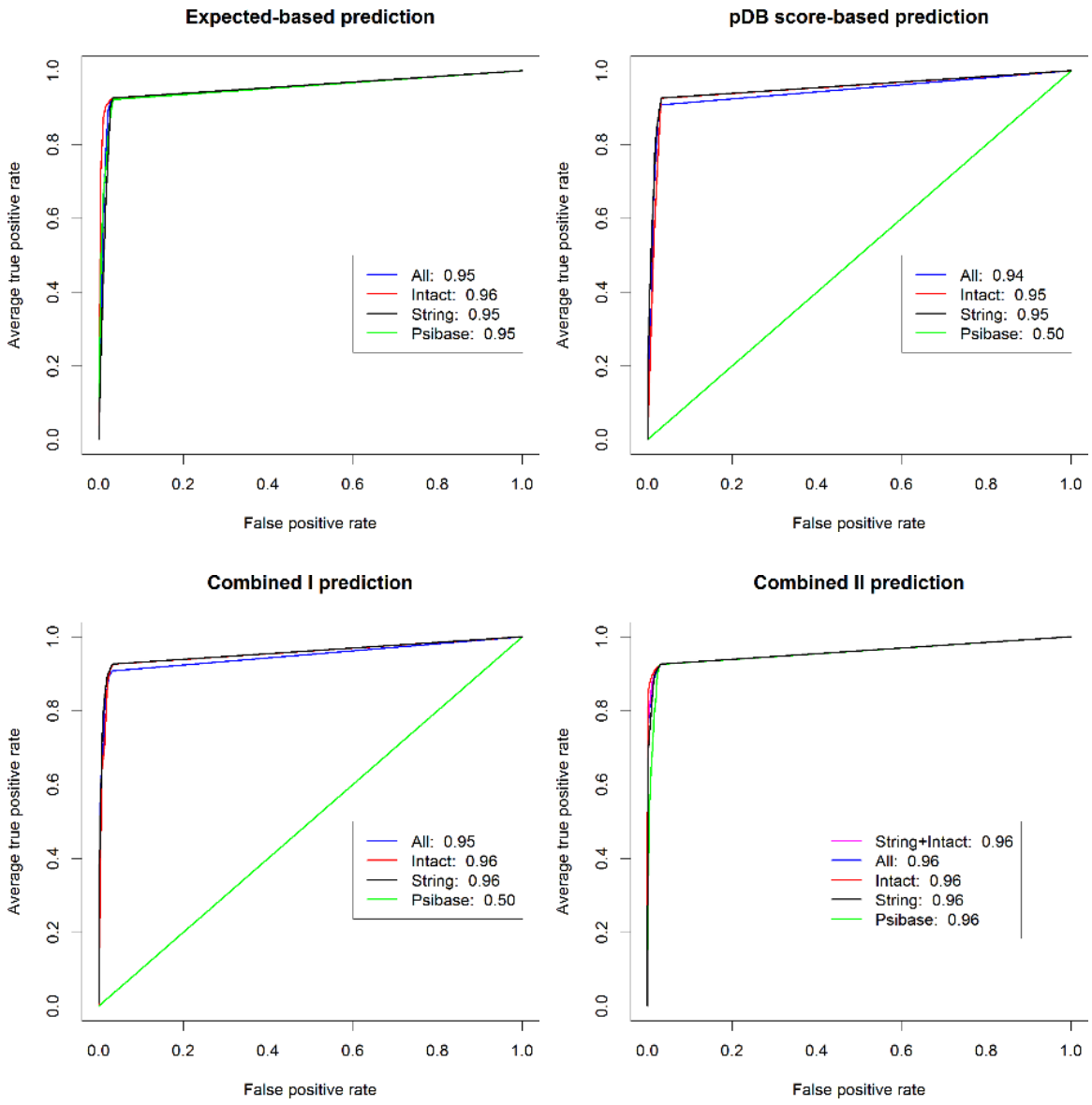


Figure 1 - ROC curves corresponding to the metrics generated with the Blast+ parameter num_alignments set to 20 and minimum interaction pair metric value min(a,b). The interaction pairs that had no values extracted from Blast+ are included in this dataset, for which the value 0 (zero) was attributed to each metric. Conserved uses the metric Conserved/100*Coverage/100. Identity uses the metric Identity/100*Coverage/100. Combined I uses the metric pDB score*qt_pDB. Combined II uses the metric pc_identity/100*pc_coverage/100 for the Intact pDB and the metric pc_identity/100*pc_coverage/100*qt_pDB/2 for the String and Psibase pDBs.