COFACTOR: improved protein function prediction by combining structure, sequence, and protein-protein interaction information

Chengxin Zhang, Peter L. Freddolino and Yang Zhang

Supplementary Materials

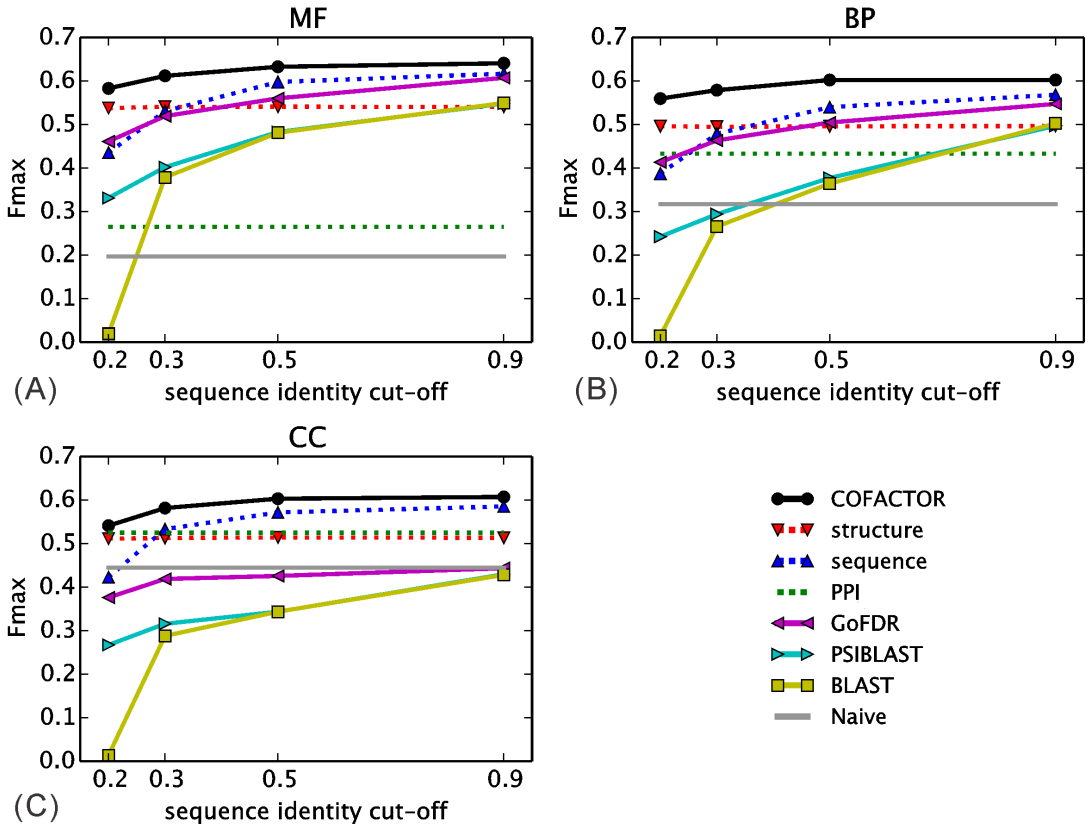

Figure S1. Accuracy of GO annotations by COFACTOR and the control methods at different sequence identity cut-offs on a test set of 1,224 non-redundant proteins. (A) molecular function (MF); (B) biological process (BP); (C) cellular component (CC). Accuracy is evaluated by maximum F-measure (Fmax). No sequence identity cut-off is imposed on Naïve, as it is not relevant. 'structure' represents the structure-based pipeline used in the former version of COFACTOR server (Roy et al, Nucleic Acid Res, 40: W471-7, 2012) but based on the new template database; 'sequence' and 'PPI' represent the newly developed sequence- and PPI-based pipelines; 'COFACTOR' represents the consensus prediction combining the 'structure', 'sequence', and 'PPI' pipelines used in the current server. Only GO terms annotated by UniProt-GOA with experimental evidence codes (EXP, IDA, IMP, IGI, IEP, TAS, or IC) are taken as 'gold standards'. All parent GO terms of annotated GO terms are also considered annotations of each target. For the predicted GO terms, all their parent terms are recursively propagated toward the root such that each parent term receives the highest confidence score among its children terms. The root term of the three GO aspects and the extremely common 'protein binding' term are excluded. The corresponding precision and recall values are listed in Table S1.

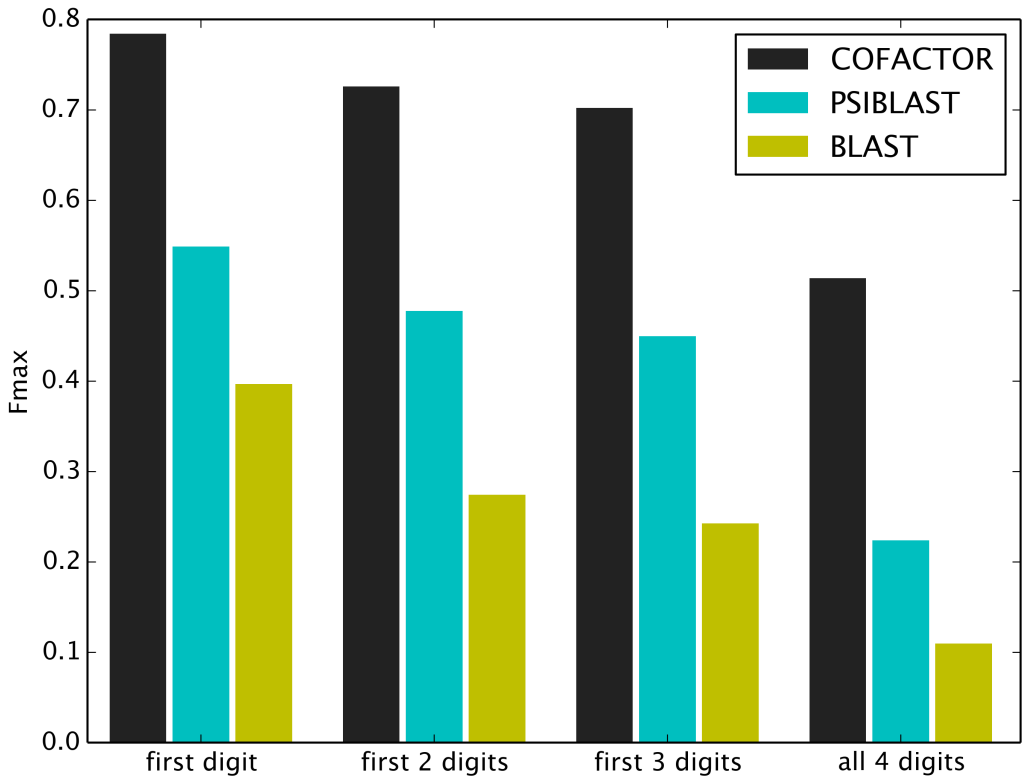

Figure S2. Accuracy of EC number prediction by COFACTOR and the control methods at 30% sequence identity cut-off. Accuracy is evaluated by maximum F-measure (Fmax). The BLAST and PSIBLAST baseline methods are implemented as in Figure S1, but use the same EC library as COFACTOR.

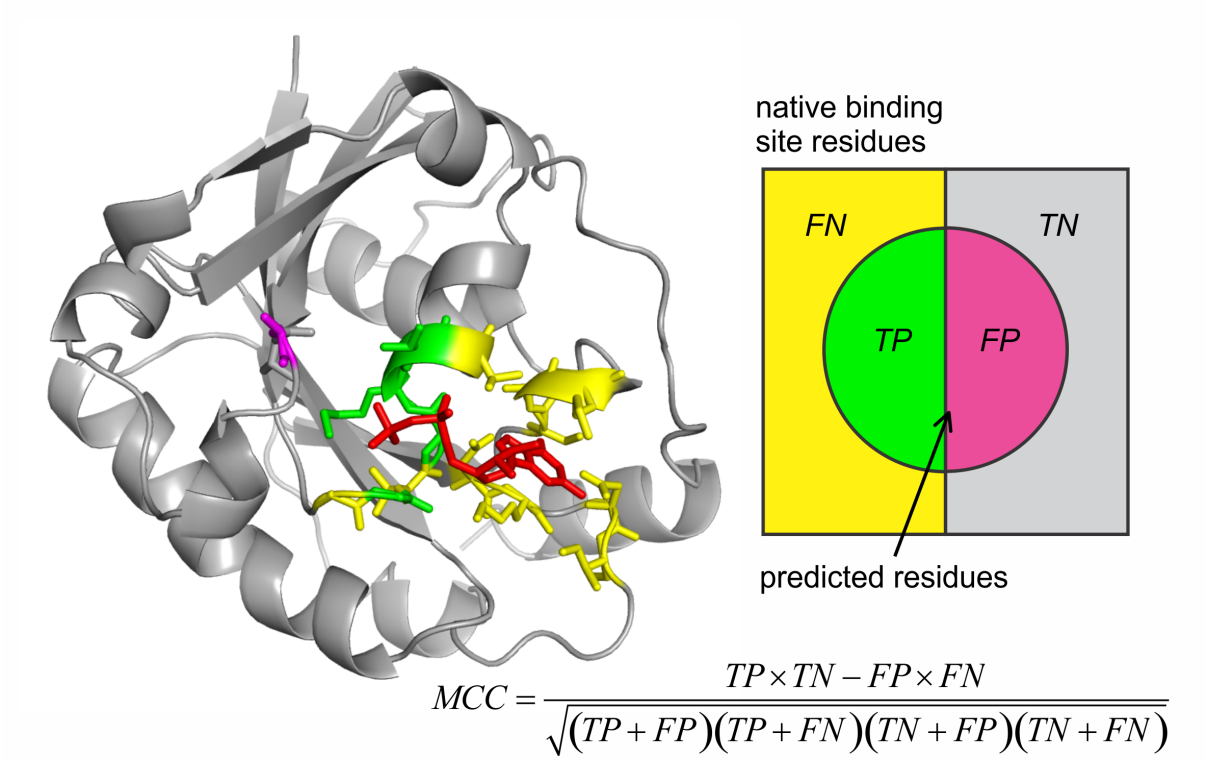

Figure S3. An illustrative example of ligand binding site prediction on chain C of the GDPRan-NTF2 complex (PDB ID: 1a2k). Red: native ligand GDP. Green: residues correctly predicted (TP). Magenta: residues incorrectly predicted (FP). Yellow: native binding site residues that are not predicted (FN). The prediction is evaluated by Mathews Correlation Coefficient (MCC). The definition of binding site is the same as that used in the community-wide CASP experiment (Schmidt et al, Proteins, 79: 126-136, 2011), i.e., a residue is defined as a binding site if it has at least one atom whose distance from the closest ligand atom is within 0.5 Å plus the sum of the van der Waals radii of the two atoms.

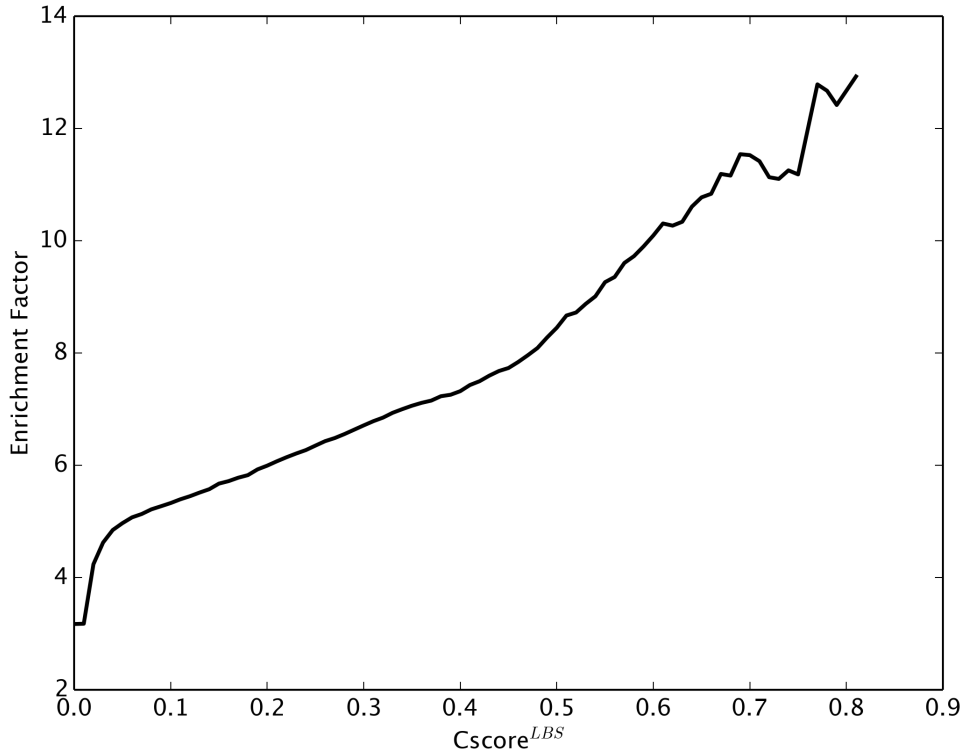

Figure S4. Enrichment factor of ligand binding site prediction by COFACTOR at different confidence score cut-offs over simply picking pockets on the protein structure using the Fpocket program.

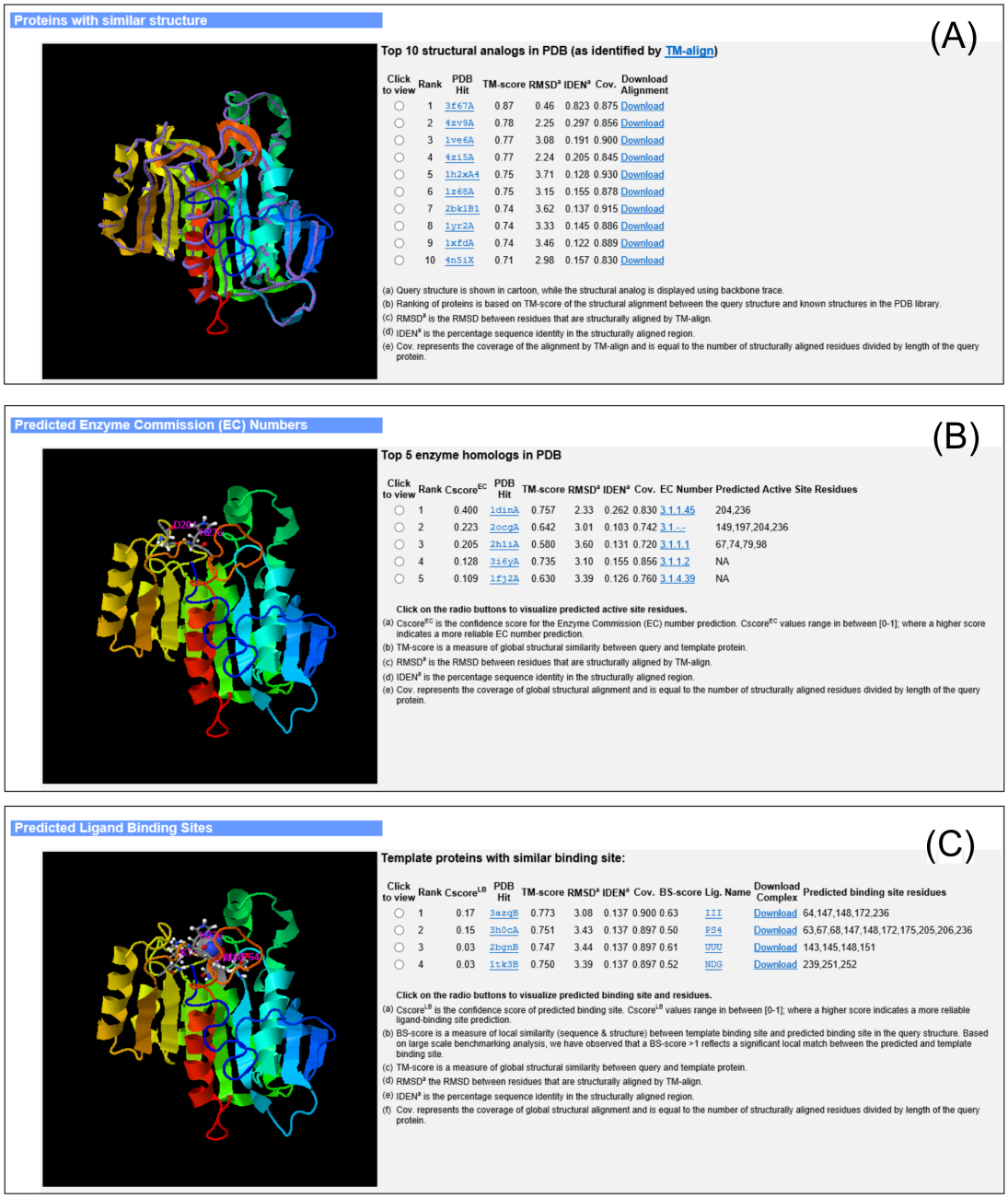

Figure S5. Illustrative examples of COFACTOR webserver output for (A) structures that are structurally closest to the query structure (B) predicted Enzyme Commission numbers, and (C) predicted ligand-binding sites. The images are same as those in Figures 3A, C, D of the main text, but with a higher resolution.

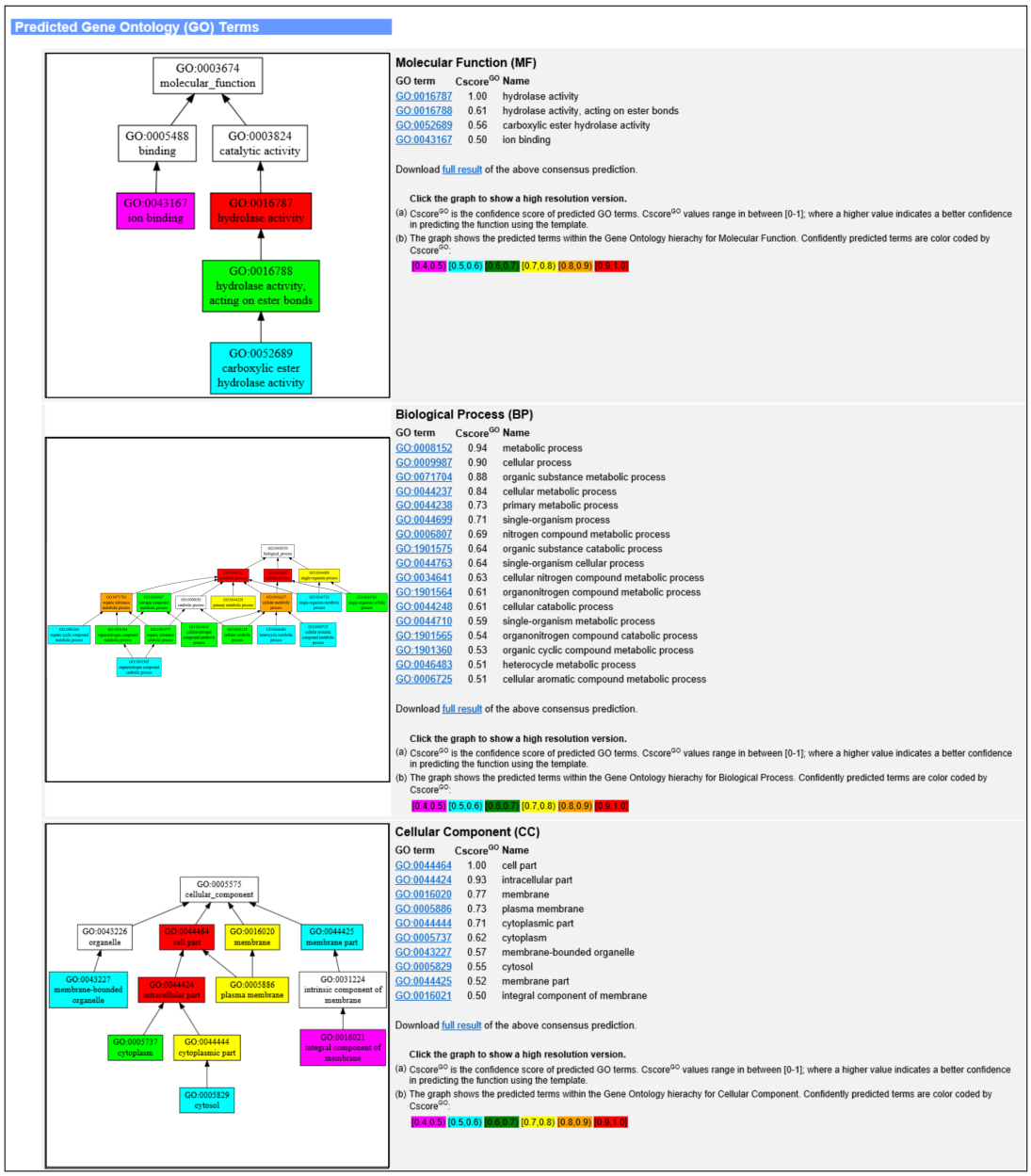

Figure S6. Illustrative examples of COFACTOR webserver output for Gene Ontology (GO) term prediction. The three panels correspond to the three aspects of GO terms: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). The images are same as those in Figure 3B of the main text, but with a higher resolution.

Table S1. Precision, Recall, and F-measure for GO annotations by COFACTOR and the control methods at different sequence identity cut-offs on a test set of 1,224 non-redundant proteins. Precision, Recall and F-measure (Fmax) are reported for Cscore cut-offs at which the maximum F-measure is obtained. 'structure', 'sequence' and 'PPI' refers to the individual structure, sequence and PPI based pipelines in COFACTOR, with 'structure' corresponding to the original COFACTOR server (Roy et al, Nucleic Acid Res, 40: W471-7, 2012). 'COFACTOR' refers to consensus prediction that combines the three pipelines, as used in the new version of the COFACTOR server. 'Naïve' refers to the baseline method in which a GO term is scored with the relative frequency of this term in Swiss-Prot over all annotated proteins.

Sequence identity cutoff	Methods	MF			BP			CC		
		Precision	Recall	Fmax	Precision	Recall	Fmax	Precision	Recall	Fmax
20%	COFACTOR	0.563	0.602	0.582	0.530	0.592	0.559	0.553	0.529	0.541
	structure	0.461	0.645	0.538	0.504	0.488	0.496	0.517	0.507	0.512
	sequence	0.497	0.387	0.435	0.457	0.336	0.387	0.546	0.345	0.423
	PPI	0.185	0.350	0.243	0.420	0.444	0.432	0.438	0.658	0.526
	GoFDR	0.510	0.420	0.461	0.454	0.380	0.414	0.423	0.339	0.376
	PSIBLAST	0.278	0.410	0.332	0.206	0.294	0.242	0.217	0.348	0.267
	BLAST	0.645	0.010	0.019	0.511	0.008	0.015	0.425	0.007	0.014
30%	COFACTOR	0.582	0.643	0.611	0.586	0.572	0.579	0.655	0.524	0.582
	structure	0.461	0.654	0.541	0.501	0.489	0.495	0.516	0.510	0.513
	sequence	0.603	0.474	0.531	0.515	0.447	0.479	0.659	0.448	0.533
	PPI	0.192	0.339	0.245	0.428	0.436	0.432	0.438	0.659	0.526
	GoFDR	0.525	0.514	0.519	0.521	0.417	0.464	0.411	0.427	0.419
	PSIBLAST	0.382	0.4243	0.402	0.276	0.315	0.294	0.317	0.315	0.316
	BLAST	0.342	0.424	0.379	0.229	0.316	0.266	0.238	0.365	0.288
50%	COFACTOR	0.619	0.647	0.633	0.616	0.589	0.602	0.675	0.545	0.603
	structure	0.461	0.654	0.541	0.501	0.491	0.496	0.517	0.512	0.514
	sequence	0.656	0.549	0.598	0.582	0.504	0.540	0.664	0.502	0.572
	PPI	0.192	0.340	0.245	0.422	0.443	0.432	0.444	0.648	0.527
	GoFDR	0.624	0.508	0.560	0.549	0.467	0.505	0.406	0.448	0.426
	PSIBLAST	0.497	0.470	0.483	0.376	0.379	0.377	0.321	0.370	0.344
	BLAST	0.467	0.496	0.481	0.347	0.384	0.365	0.316	0.377	0.344
90%	COFACTOR	0.644	0.636	0.640	0.616	0.588	0.602	0.632	0.584	0.607
	structure	0.461	0.655	0.541	0.502	0.492	0.497	0.516	0.511	0.513
	sequence	0.716	0.542	0.617	0.614	0.529	0.568	0.647	0.535	0.587
	PPI	0.192	0.340	0.245	0.422	0.443	0.432	0.444	0.649	0.527
	GoFDR	0.688	0.544	0.607	0.573	0.524	0.548	0.460	0.428	0.444
	PSIBLAST	0.594	0.509	0.548	0.625	0.413	0.497	0.652	0.321	0.430
	BLAST	0.638	0.483	0.549	0.628	0.419	0.497	0.677	0.313	0.430
na	Naïve	0.210	0.186	0.197	0.305	0.329	0.317	0.479	0.416	0.445