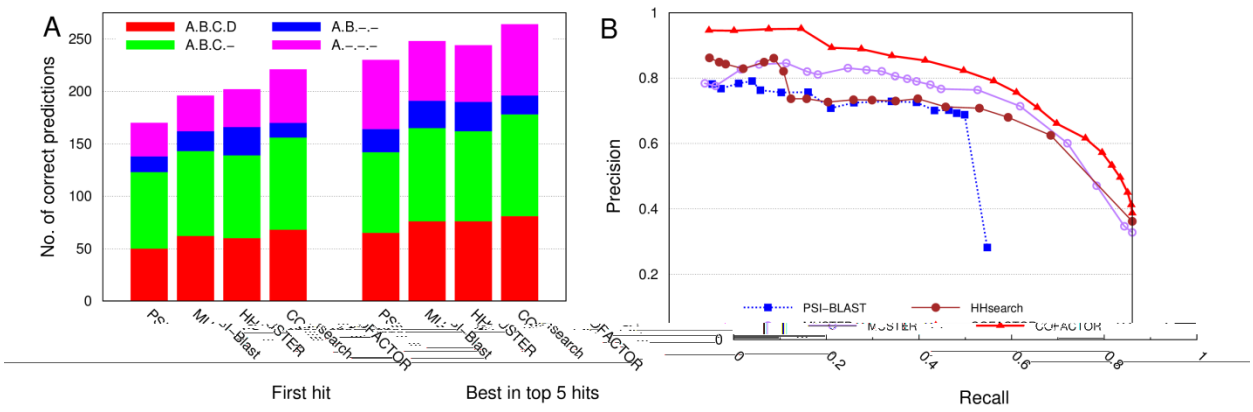
Supplimentary Data

1. Benchmark data set used for evaluating EC number and GO term predictions

EC number and GO term predictions by COFACTOR are evaluated on 450 non-homologous proteins collected from the PDB library with diverse functions, while ensuring that the pair-wise sequence identity is below 30% and there is no self BLAST (1) hit within the dataset. In this set, 318 proteins are enzymes, with EC numbers covering all the 6 enzyme classes. The GO term predictions are evaluated on 337 proteins annotated with at least one GO term. These include 205 enzymes and 132 non-enzymatic proteins. Of these 337 proteins, 308 were annotated with at least one molecular function term; 295 were annotated to be involved in a biological process and cellular location was annotated for 213 proteins in the PDB GOA annotation (2). The Gene Ontology predictions are evaluated on each of the three subsets individually, and also as a combined set.

2. Enzyme Commission number prediction results



Supplimentary Figure S1 Performance comparisons for EC number prediction. (A) Histogram analysis of functional inferences drawn for 318 benchmarking enzymatic proteins at different level of Enzyme Commission number. (B) Precision-Recall analysis for predicting first three digits of EC number.

3. Gene Ontology Term prediction results

Functional similarity between query and template protein for Gene Ontology predictions is evaluated by measuring semantic similarity (SS) between GO terms and functional similarity between gene products (31). Semantic similarity measure relatedness between the GO terms based on the DAG structure of Gene Ontology and information content of the term. In this analysis, we used SS is evaluated based on Relevance similarity Sim_{Rel} (32). Given this way of measuring SS between two GO terms, we evaluate functional similarity (*Fsim*) of predicted GO terms $GO_1 = \{go_{11}, go_{12}, go_{13}, .go_{1m}\}$ with the annotated GO terms of query protein $GO_2 = \{go_{21}, go_{22}, go_{23}, .go_{2n}\}$ using the best match average score strategy (33), which is defined as :

$$FSim(\mathrm{GO}_1, \mathrm{GO}_2) = \frac{\sum_{1 \le i \le m} SS_{\max}(go_{1i}, GO_2) + \sum_{1 \le j \le n} SS_{\max}(go_{2j}, GO_1)}{m+n}, \qquad \text{(S1)}$$

where $SS_{max}(go, GO)$ represents the maximum SS between *go* and any of the terms in the set *GO*. Both *Fsim* and SS range between 0 and 1.

Supplementary Table S1. Coverage of Gene Ontology prediction for the 337 benchmarking proteins, using the top and best in top5 template proteins, identified by homology based functional annotation approaches in the same template library. Coverage of overall GO terms and for the three ontologies (molecular function, biological process and cellular component) is analyzed as the number of query proteins for which template proteins have *Fsim* (Eq. S1) > 0.5.

	Overall		Molecular function		Biological Process		Cellular component	
Method	Top	Best	Top	Best	Top	Best	Top	Best
PSI-BLAST	128	166	154	179	104	130	53	75
MUSTER	150	205	184	235	128	160	71	114
HHsearch	139	189	173	223	116	146	63	109
COFACTOR	165	216	194	237	136	171	67	137

Supplementary References

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.

2. Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C. and Apweiler, R. (2009) The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res*, **37**, D396-403.

31. Pesquita, C., Faria, D., Falcao, A.O., Lord, P. and Couto, F.M. (2009) Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, **5**, e1000443.

32. Schlicker, A., Domingues, F.S., Rahnenfuhrer, J. and Lengauer, T. (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, **7**, 302.

33. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S. and Chen, C.F. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274-1281.