MetaGO: Predicting Gene Ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping

Chengxin Zhang, Wei Zheng, Peter L Freddolino, and Yang Zhang

Supplementary Material

Text S1: Control methods

Following the definitions used in the CAFA experiments, we implemented three baseline methods, "Naive", "BLAST", and "PSI-BLAST", to compare with the MetaGO pipelines. In "Naïve", regardless of the query, GO term *q* is predicted with a confidence score that equals the relative frequency of this term in the UniProt database over all annotated proteins. In "BLAST", a query is searched against all UniProt proteins annotated with known GO terms using BLAST with E-value cutoff 0.01, while in "PSI-BLAST", a query is first searched against the UniRef90 database using PSI-BLAST with three iterations and an E-value cutoff 0.01 to generate an initial sequence profile, which is again used to search against the annotated UniProt library in one iteration. For both "BLAST" and "PSI-BLAST", the default confidence score for GO term *q* is the highest local sequence identity to any (PSI-)BLAST hit annotated with *q* at the aligned region.

In addition to the standard baseline methods, we include two separate GO term prediction methods, GoFDR and GOtcha, in our control method set. GoFDR is designed to generate GO predictions from the functionally discriminating residues (FDRs) in multiple sequence alignments (Gong et al, Methods, 93:3, 2016); whereas GOtcha combines and recalibrates function prediction from sequence homologs detected across different species (Martin et al, BMC Bioinformatics, 5, 2004). GoFDR was ranked as the top predictor in the 2nd Critical Assessment of Function Annotation (CAFA2) experiment (Jiang et al, Genome Biology, 17:184, 2016) and is the only CAFA2 algorithm that provides a standalone program. To our knowledge, GoFDR and GOtcha are the only two commonly used and publicly available GO prediction methods that can be downloaded and run locally on our computers. This allows us to compare them with MetaGO on benchmark proteins with databases modified to restrict the sequence identity between query protein and available functional templates.

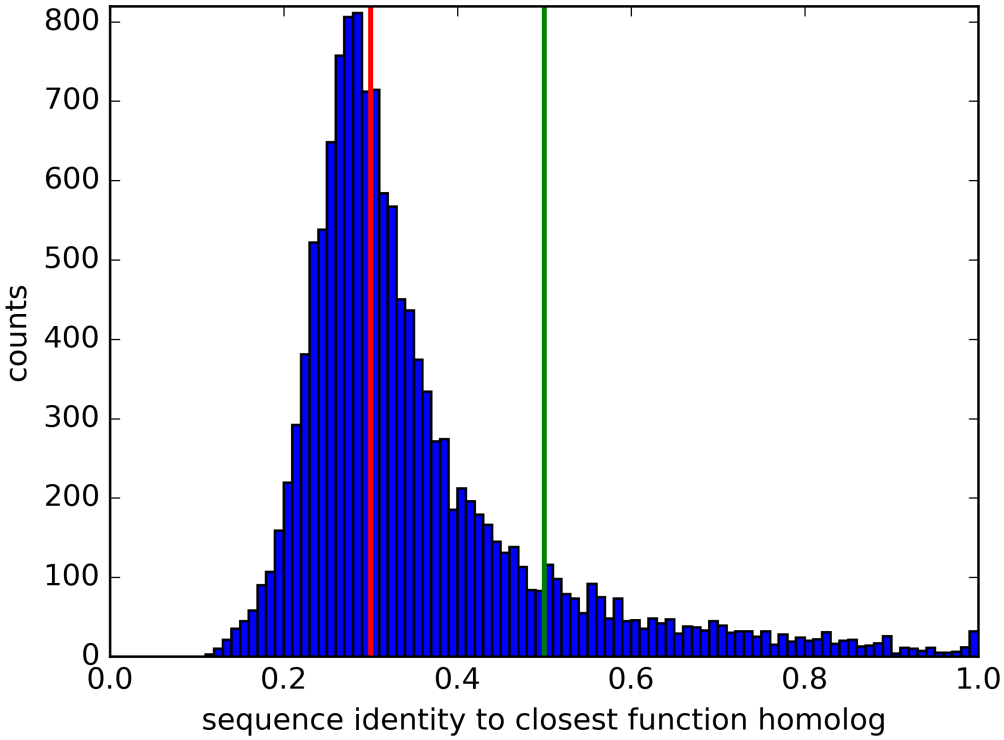

Figure S1. Distribution of sequence identity (number of identical residues divided by query length) between 1,400 randomly sampled UniProt sequences and their closest function homologs with known GO annotations. These UniProt proteins are sampled with two criteria: sequence length is between 30 and 700 amino acids, and pairwise sequence identities between different sampled sequences are < 40%. The red and green vertical lines indicate that 41% and 87% of sampled proteins share less than 30% and 50% sequence identity to their closest function homolog, respectively.

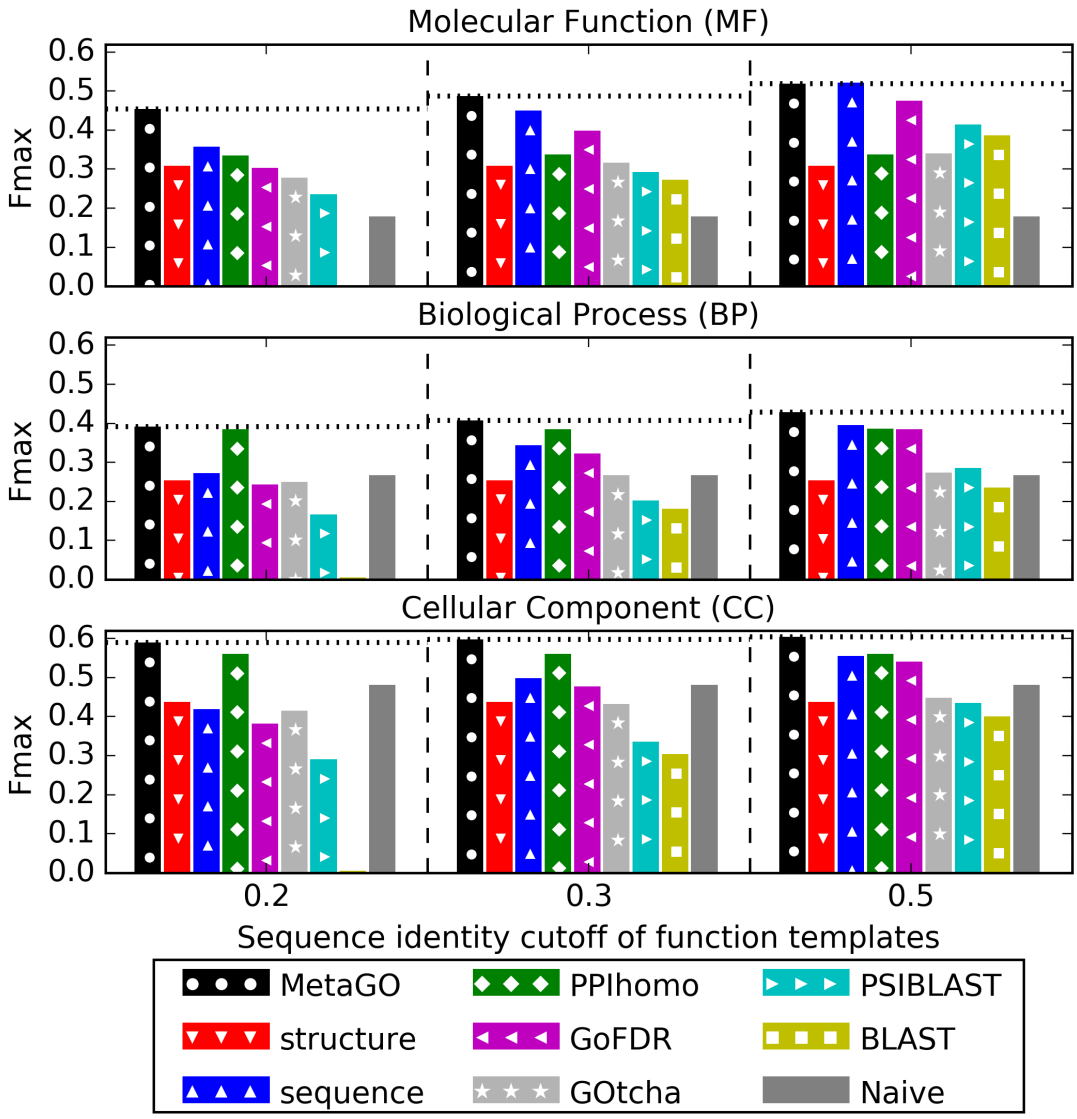

Figure S2. Color version of Figure 2 from the main text, showing Fmax score of the GO predictions by MetaGO, compared to that by the three component pipelines (structure, sequence, and PPI-homolog), and five control methods (GoFDR, GOtcha BLAST, PSI-BLAST, and Naïve) at different sequence identity cut-offs for filtering function templates. The dotted horizontal lines label the performance of MetaGO.

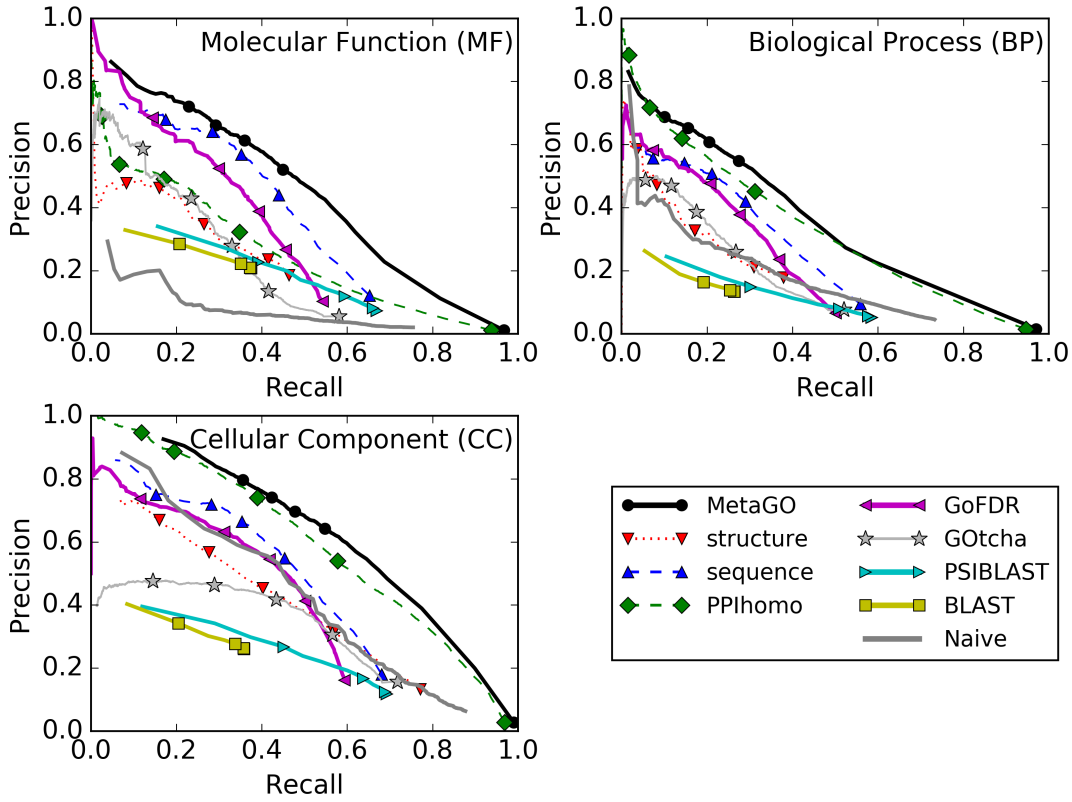

Figure S3. Color version of Figure 3 from the main text, showing precision-recall curves of GO predictions by MetaGO, compared to that by the three component pipelines (structure, PPIhomo, and sequence), and five control methods (GoFDR, GOtcha, BLAST, PSI-BLAST, and Naïve) at a 30% sequence identity cut-off for functional templates.

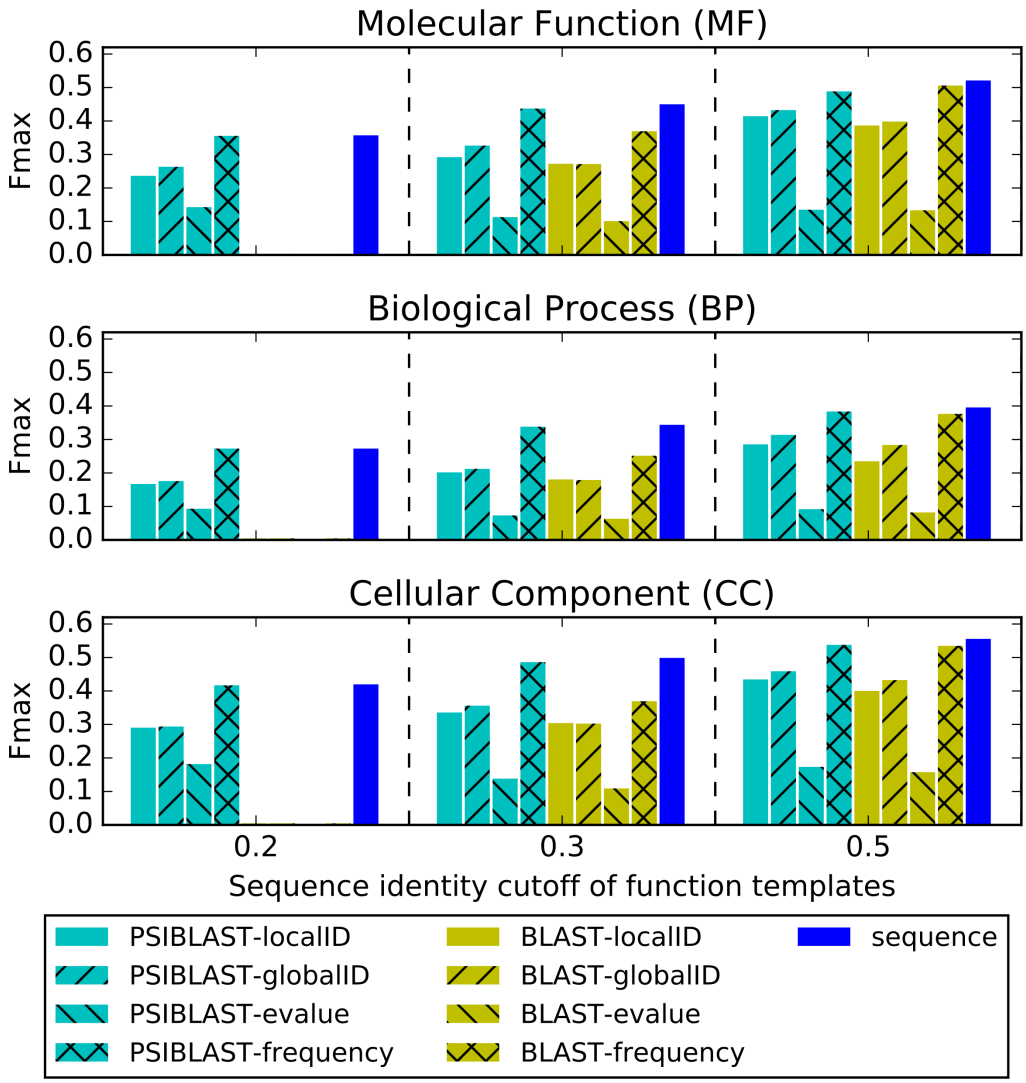

Figure S4. The Fmax score of the GO prediction by PSI-BLAST and BLAST using four different scoring functions (*localID*, *globalID*, *evalue*, and *frequency*) for selecting the functional templates. "*sequence*" indicates the sequence-based pipeline developed in MetaGO, which combines the prediction results from PSI-BLAST and BLAST hits.

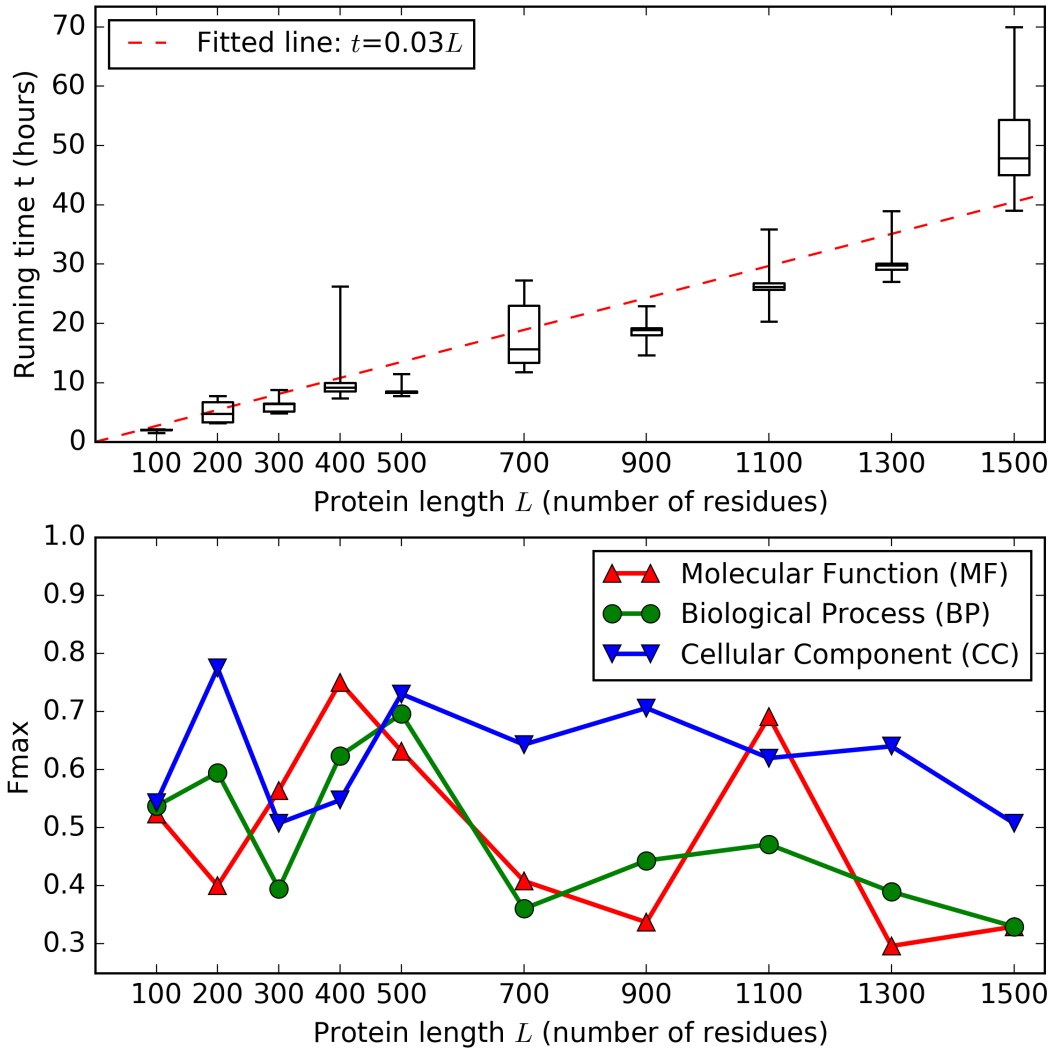

Figure S5. Performance of MetaGO on different size of proteins. From the CAFA3 targets, five proteins are selected for each of the following ten protein lengths: 100, 200, 300, 400, 500, 700, 900, 1100, 1300, and 1500 residues. All the 50 selected proteins have experimental GO annotations, and share less than 30% sequence identity to each other and to the MetaGO training set of 1,224 *E. coli* proteins. The upper panel displays the distribution of MetaGO running time for the five proteins with a given sequence length, where the total running time nearly linearly increases with the protein length (with a Pearson Correlation Coefficient 0.96). It should be noted that within the MetaGO webserver, the three component methods (sequence, structure, PPI-homolog) are run in parallel, and the sequence and PPI-homolog methods are much faster than the structure-based method. Therefore, the total running time of the webserver depends almost exclusively on the speed of structure-based method. The lower panel lists the average Fmax score versus the protein length. While there is strong correlation between speed and protein length, the prediction accuracy (Fmax) of MetaGO for any of the three GO aspects does not have a clear dependency on protein length.

Table S1. Fmax score of the GO prediction on our test set by MetaGO, its three component methods (structure, sequence, and PPI-homolog), and five control programs at different sequence identity cutoffs. As the prediction from "Naïve" is independent of input sequence, only one Fmax value is shown for each GO aspect, which does not correspond to any specific sequence identity cutoffs.

Method	GO Aspect	Sequence identity cutoffs 0.2	0.3	0.5
MetaGO	MF	0.454	0.487	0.518
	BP	0.391	0.408	0.428
	CC	0.589	0.598	0.605
structure	MF	0.308	0.309	0.308
	BP	0.254	0.254	0.253
	CC	0.438	0.438	0.438
sequence	MF	0.357	0.450	0.521
	BP	0.273	0.344	0.396
	CC	0.420	0.499	0.556
PPI-homolog	MF	0.335	0.337	0.338
	BP	0.385	0.386	0.386
	CC	0.561	0.561	0.561
GoFDR	MF	0.303	0.399	0.475
	BP	0.244	0.323	0.385
	CC	0.382	0.478	0.541
GOtcha	MF	0.278	0.316	0.340
	BP	0.250	0.267	0.273
	CC	0.416	0.433	0.449
PSIBLAST	MF	0.236	0.292	0.414
	BP	0.167	0.202	0.285
	CC	0.290	0.336	0.434
BLAST	MF	0.003	0.272	0.386
	BP	0.005	0.181	0.235
	CC	0.005	0.304	0.400
Naïve	MF		0.179	
	BP		0.267	
	CC		0.481	

Table S2. Weight parameters for different pipelines in Eq. (9) decided by logistic regression, based on a set of 1,224 training proteins taken from the *E. coli* genome which are non-homologous to the test proteins of this study. These proteins have experimental GO annotations, and their length ranges from 38 to 968 residues.

Weights	**MF**	**BP**	**CC**
w_{sequence}	4.828	3.455	2.716
w_{PPIhomo}	11.374	8.465	19.453
$w_{\text{structure}}$	4.878	3.921	3.395
w_{Naive}*	-6.305	-0.827	-13.384
w_0	-5.659	-5.317	-5.805

*We note that the weight for "Naive" is negative, and therefore the over-prediction of uninformative GO terms that are close to the root of GO hierarchy is suppressed. It is of interest to note that this piece of data seems to be counter-intuitive in that the "Naïve" component, which is a strong predictor for CC, is given a negative weight by logistic regression in MetaGO. To understand the reason for the weight w'_{Naive} being negative, we can re-write the logistic regression expressed by Eq. (9) into:

$$Cscore^{MetaGO}(q) = 1/\left\{1 + \frac{exp[-x(q)]}{exp[-x'(q)]} \cdot exp(-w_0)\right\} \quad (S1)$$

In this equation, the argument in the denominator, i.e.,

$$x(q) = \sum_{m\epsilon\{structure,sequence,PPIhomo\}} w_m \cdot Cscore^m(q) \quad (S2)$$

is a linear combination of the three component methods for predicting GO term q, while the argument in the nominator,

$$x'(q) = w'_{Naive} \cdot Cscore^{Naive}(q) \quad (S3)$$

is the expected value of Eq. (S2), where $w'_{Naive} = -w_{Naive} > 0$ is the (positive) weight for q's background probability $Cscore^{Naive}(q)$. Thus, the term $\frac{exp[-x(q)]}{exp[-x\prime(q)]}$ quantifies the deviation of combined confidence for the three components from the expected value (background).