Supporting Information

Llewellyn and Eisenberg 10.1073/pnas.0809583105

Materials and Methods

Describing Protein Function. We describe protein function with a collection of probability distributions over Gene Ontology (GO) annotations. Here, the set of possible terms is a new subset of the biological process GO annotations (BP) that form leaves of the GO directed acyclic graph (DAG). To distinguish incompleteness in the ontology from uncertainty in annotation assignment, we create a conditional unknown for each BP, that is, a new term that represents a more specific but undescribed child of the parent BP term. As we allow each BP to have exactly one conditional unknown, each represents the set of all currently undescribed terms. For BP that are currently leaves in the GO hierarchy, the addition of the conditional unknown is only a formality that allows us to dispense with distinguishing between conditional unknown leaves and existing leaves. For general terms (i.e., those that are not leaves), this extension provides the opportunity to represent the uncertainty about function that may have been expressed in annotating a protein with a general term, separately from the possibility that the appropriate more specific GO term did not exist, either due to choices made in structuring the GO DAG (a situation encountered in the mapping between GO and another ontology, such as Interpro), or because the function has yet to be described by biology. Here we make no attempt to determine on an individual protein basis an estimate of the conditional probability, but make a single estimate for each BP by checking for redundancy in the mapping between GO and Interpro, finding the fraction of times both a parent and child Interpro accession mapped (24) to the same GO term. We used a prior count of one so that all annotations had some conditional unknown probability.

With the addition of the conditional unknown leaf, any GO term can be viewed as a distribution over its leaves. We represent protein function as a collection of such distributions over BP

$$\Omega = \{f{:}f \subset C\}$$

where C is the subset of conditional unknown BP leaves (known from here on as simply leaves, unless otherwise specified) of primary annotations among a reference set of proteins, here all annotated proteins found in any Fungi. We define a primary annotation, b, as the most specific annotation previously assigned to a protein along a single path in the GO DAG from among all annotations assigned to the protein in that path with identical GO evidence codes. A protein is often assigned several primary annotations: these could be identical BP with different evidence codes, or different BP with the same evidence codes if they lie on separate DAG paths. By designating primary annotations we remove redundant information and identify C. We assume that each primary annotation represents independent information about a protein: in well-curated organisms such as *S. cerevisiae* this is less troublesome than in organisms that are annotated largely by automated means; these may require additional filtering to select C.

Each primary annotation of a protein generates a probability distribution over its leaves f according to the frequency of the leaves among C and a simple separate procedure to make a first approximation of the probability for conditional unknowns of primary annotations.

$$F = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}$$

where F is a probability vector such that

$$\textstyle\sum_{j \varepsilon c} f_j = 1.$$

The formation of F assumes that each primary annotation b can be represented by a single best leaf. A protein is described by a collection of N such distributions, $\{F_1, F_2, \cdots, F_N\}$, each weighted by the probability that the primary annotation is correct, given the evidence code I: $Pr(b|I_b)$, so that the probability that f is a correct best leaf of b is, assuming independence,

$$\Pr(F_b = f \cap b \ is \ correct) = \Pr(F_b = f)\Pr(b|I_b).$$

The probability that f is a correct annotation for the protein among the N primary annotations is then

$$\begin{aligned} \Pr(f)' &= (F_1 \\ &= f \cap b_1 \ is \ correct) \\ &\cup F_2 = f \cap b_2 \ is \ correct) \\ &\cup \ldots (F_N = f \cap b_N \ is \ correct) \end{aligned}$$

which we calculate by finding the complement of the probability that all f are false

$$\Pr(f)' = 1 - \Pi_{q=0}^{N}[1 - \Pr(F_q = f)\Pr(b_q|I_{b_q})].$$

We also create the error probability vector E for the protein with a uniform distribution over all f. E contributes only to the extent that f is incorrect on all F:

$$\Pr(f) = \Pr(f)' + (1 - \Pr(f)')E_f.$$

The Likelihood. Our likelihoods are estimates of the probabilities of observing a linked protein with its BP and their evidence codes linked with a given score (here zorch) to the target protein, given the hypothesis that a particular leaf BP is the best description of the biological process of the target. The likelihoods are based on empirical counts of linked proteins with given BP in *S. cerevisiae*. To describe the general relationship between the linkage score (e.g., zorch) and function, we first counted each linked pair of annotated proteins in *S. cerevisiae* at their nearest discretized GO distance and linkage score. For each pair of GO annotations, we corrected for the graph structure of GO with the probability of choosing the BP f from among all BP at the linkage score l and GO distance d from H_i

$$\Pr(f,l|d,H_i) \cong \Pr(l|d)\Pr(f|d,H_i)$$

where H_i is the hypothesis, another leaf of BP that best describes the target.

We then estimated the more specific relationship between each individual BP pair and the linkage score, $\Pr(f,l|H_i)$, by counting observations of the pairwise BP of linked proteins, initializing these sparse counts with a prior count guided by the general probability relationship between GO distance and the linkage score.

We followed our principle of measuring the functional similarity of proteins only at their closest BP in estimating the likelihood. Unlike a typical likelihood, we construct ours to deal with the high dimensionality (the many BP of the linked protein) and uncertainty by allowing each hypothesis to focus on the data to its best advantage. We do so by finding the probability of observing the BP f_k of the linked protein that maximizes the normalized likelihood z for the hypothesis H_i,

$$z = \Pr(f,l|H)/\Pr(f,l).$$

The normalization is necessary for fair competition among hypotheses: the fact that a linked protein has a rarely observed BP, for example, should not affect the choice of the best BP for the hypothesis. If the linked protein were to have only one BP rather than a distribution F, this likelihood would lead to a standard Bayesian update with typical normalization by the probability of the data. The practical effect of our likelihood is to translate uncertainty in the data differentially to the hypotheses, rather than to all hypotheses equally.

If the linked protein had a single distribution represented by F_1, we could estimate the likelihood L_i for H_i with the expected value

$$L_i \cong \sum_{j=0}^{n} z_j \Pr(F_1 = f_j)\Pr(b_1|I_{b_1}).$$

To accommodate the many distributions generated by all primary annotations and the error, $\{F_1,F_2,\ \ ,F_NE\}$, we cannot rely on the mutual-exclusivity of a single distribution to find the best leaf f for H_i. Instead we find the expected maximum normalized likelihood over all distributions by first ordering f by z so that $f_g < f_j$ if $z_g > z_j$. For f_k to be the best leaf BP of the linked protein to support the hypothesis H_i, it must be the first correct f along the order of z across all distributions.

We wish to find the probability p_k that f_k is correct from any distribution, and that no other f_j is correct, where $j < k$ ordered by z,

$$p_k = \Pr(f_k \text{ is correct} \cap \text{no other } f_j \text{ is correct where } j < k).$$

Considering many distributions adds combinatorial complexity, especially when dealing with ties. We avoid the computational burden by finding p_k through the change in the cumulative distribution of the probability that f_j in at least one distribution in the set D $\{F_1,F_2,\ \ ,F_NE\}$, is correct where $j < k$ ordered by z.

$$\mathrm{cdf}_k = 1 - \prod_{j=0}^{k}(1 - \Pr(f_j))$$

and

$$p_k = \mathrm{cdf}_k - \mathrm{cdf}_{k-1}$$

where cdf_{-1}=0.

We then estimate the likelihood using the expected maximum normalized likelihood for the hypothesis H_i:

$$L_i \cong \sum_{k=0}^{n} p_k z_k.$$

The construction of the likelihood required binning zorch values and GO distances; we binned manually to balance having adequate counts in each while distinguishing protein pairs with high zorch values and strong functional similarity from the bulk of the data.

Voting. We made two optimized alterations to a majority voting procedure to compare it more favorably to GFL. Each linked protein was given only one vote, so that if it has more than one BP these cast fractional votes. We also limited the linked proteins to those within 75% of the probability of individual success of the best, as voting was less successful when many linked proteins were used.

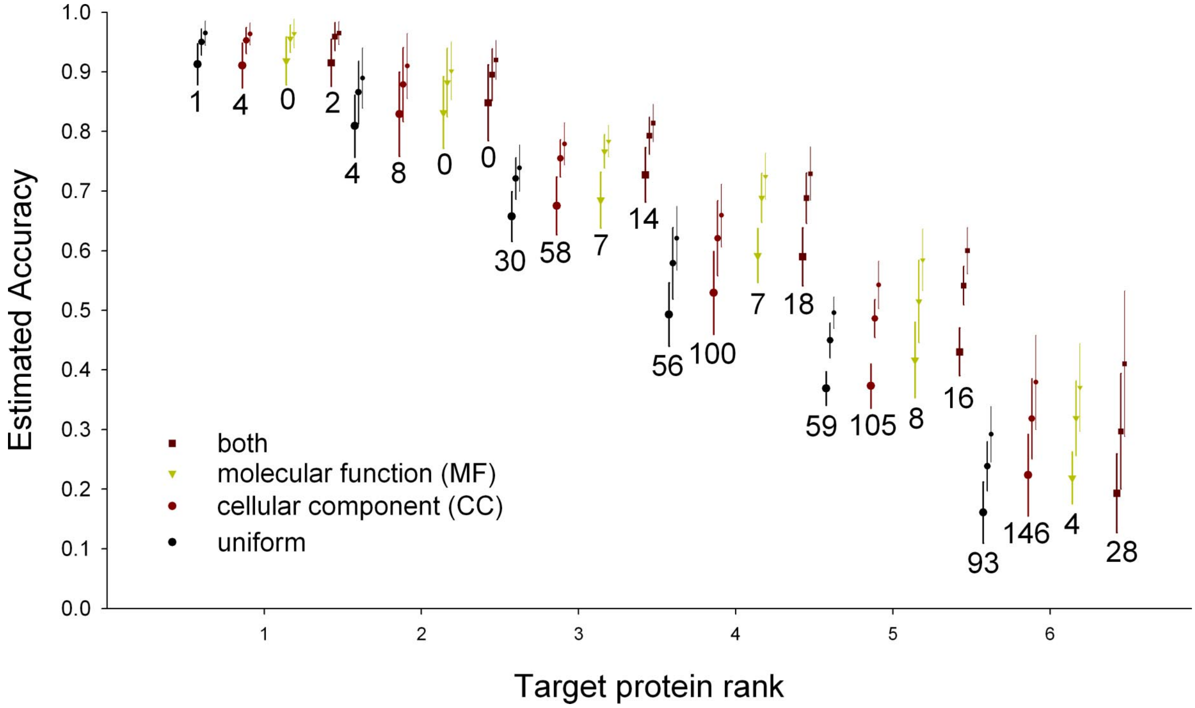

Fig. S1. Predicting unannotated yeast proteins. To estimate the accuracy of zorch-defined links integrated by Generalized Functional Linkages (GFL) for those yeast proteins without biological process annotations (BP), we grouped these by their expected accuracies inferred from the ten-fold cross validation of annotated proteins. The counts of unannotated proteins in each group are shown below the error bars that give one standard deviation from the mean. If more than one choice among the 206 classifiers was chosen from the posterior, the expected accuracy that at least one was correct increased, especially for lower ranked targets: one, two, and three choices, rising, are shown for each prior and ranked group. Many yeast proteins of unknown biological process have cellular component annotations (CC); fewer have molecular function (MF). Those with neither annotation used a uniform prior.

Other Supporting Information Files

Table S1 (PDF)
Table S2 (PDF)
Table S3 (XLS)