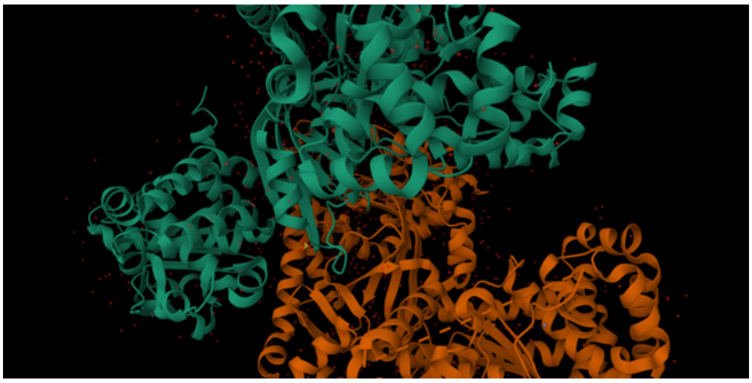
CAFA 6 Protein Function Prediction

Predict the biological function of a protein



Overview

Proteins are large molecules that are responsible for many activities in our cells, tissues, organs, and bodies and they also play a central role in the structure and function of cells. Proteins are composed of 20 types of smaller molecules known as amino acids, which are ordered in a long chain known as the protein amino acid sequence. Each protein has its own sequence that determines its structure and its function. You will build a model that predicts what a protein does based on its amino acid sequence. These predictions will help researchers understand how proteins function, and could lead to the development of new medical treatments and therapies.

Start
2 months ago

Close
2 months to go
Merger & Entry

Context

Proteins are responsible for many activities in our tissues, organs, and bodies and they also play a central role in the structure and function of cells. Proteins are large molecules composed of 20 types of building-blocks known as amino acids. The human body makes tens of thousands of different proteins, and each protein is composed of dozens or hundreds of amino acids that are linked sequentially. This amino-acid sequence determines the 3D structure and conformational dynamics of the protein, and that, in turn, determines its biological function. Due to ongoing genome sequencing projects, we are inundated with large amounts of genomic sequence data from thousands of species, which informs us of the amino-acid sequence data of proteins for which these genes code. The accurate assignment of biological function to the protein is key to understanding life at the molecular level. However, assigning function to any specific protein can be made difficult due to the multiple functions many proteins have, along with their ability to interact with multiple partners. More knowledge of the functions assigned to proteins—potentially aided by data science—could lead to curing diseases and improving human and animal health and wellness in areas as varied as medicine and agriculture.

Research groups have developed many ways to determine the function of proteins, including numerous methods based on comparing unsolved sequences with databases of proteins whose functions are known. Other efforts aim to mine the scientific literature associated with some of these proteins, while even more methods combine sophisticated machine-learning algorithms with an understanding of biological processes to decipher what these proteins do. However, there are still many challenges in this field, which are driven by ambiguity, complexity, and data integration.

Important Note

This is a prospective (i.e., future) data competition. Many proteins in the Test data do not currently have any assigned functions. Proteins having one or more of their functions published by researchers during the curation phase of the competition will comprise the future test set. Final leaderboard scores will be calculated after the curation phase of the competition.

Background

The organizers provide a set of protein sequences on which the participants are asked to predict Gene Ontology (GO) terms in each of the three subontologies: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). This set of sequences is referred to as the test superset.

The proteins from the test superset that (1) originally had no experimentally assigned functions in a particular subontology and accumulate experimental annotations, or (2) originally had experimentally assigned functions in all three subontologies and accumulate experimental annotations in any subontology between the submission deadline and the time of evaluation in that subontology, are referred as the test set for that subontology. There will be three different test sets, one for each subontology, and the participants will be scored on each. The final performance accuracy will be computed by combining the three scores, as described below under Evaluation Metrics.

The organizers also provide the training set containing protein sequences that have at least one experimentally determined GO term in at least one subontology, together with those experimental annotations. These proteins may also appear in the test superset.

Evaluation Metrics

Submissions will be evaluated on proteins that have accumulated experimentally-validated functional annotations in any subontology between the submission deadline and the time of evaluation. For example, a protein that had no experimental terms in, say, the Molecular Function (MF) subontology of

GO and has accumulated experimental annotations in MF after the submission deadline will be included in the test set for evaluating the MF term predictions. In addition, a protein that already had experimental terms in all three subontologies before the submission deadline and has accumulated experimental annotations in MF after the submission deadline will also be included in the test set for evaluating the MF term predictions. The same holds for the Biological Process (BP) or Cellular Component (CC) subontologies of GO. The proteins that qualify will create three different test sets, one for each subontology of GO. The same protein can appear in more than one test set if it accumulates experimentally-validated annotations in more than a single subontology.

The maximum F1-measure based on the weighted precision and recall will be calculated on each of the three test sets, and the final performance measure will be an arithmetic mean of the three maximum F-measures (for MF, BP, and CC). The formulas for computing weighted F1-measures are provided in the supplement (page 31) of the following paper: *Jiang Y, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. (2016) 17(1): 184*, in the full evaluation mode. The weights (i.e., information content ic(f), where f is a term in any subontology) for each term f of each subontology are provided by the challenge organizers. Note that we equivalently refer to those weights as ia(f), called information accretion for the functional term f. The rationale for using weighted precision and recall is that GO is hierarchical and thus, the terms on top of the hierarchy are implied by their descendants. The weight for a term is determined by the logarithm of the frequency of occurrence of that term in a large pool of proteins. The root terms appear in every protein's annotation and thus, their weights are 0. Terms deep in the ontology tend to appear less frequently, be harder to predict, and thus their weights are larger (Clark & Radivojac, 2013). However, this does not always hold true, as highlighted in the following discussion.

Using the terminology from Jiang et al. (2016), the evaluation will be carried out for no-knowledge and limited-knowledge protein targets combined, in the full evaluation mode, using maximum F-measures of information-accretion weighted precision and recall, one for each subontology. Note that in this competition, we also include the evaluation of proteins that already had experimental terms in all three subontologies, and have accumulated more experimental terms after the submission deadline, this is known as partial-knowledge protein targets. The three maximum F-measures of the three subontologies (Molecular Function, Biological Process, and Cellular Component) will be combined as an arithmetic mean for each subtype of knowledge gain. Finally, the three F-measures from the three subtypes no-knowledge, limited-knowledge, and partial-knowledge will be combined again as an arithmetic mean to compute the final performance. The evaluation code is available on this GitHub repository.

Leaderboard

The participants are cautioned that the leaderboard was designed to display method performance on a relatively small selection of proteins from the test superset (see Data), provided to us by the UniProtKB team, but not available in UniProtKB or other public databases. These proteins will not be included in the test set for the subontologies used for the leaderboard evaluation. The final test set will consist of proteins that will have accumulated functional terms after the submission deadline, and therefore, some distribution shift between the sample of proteins used for the leaderboard and the final evaluation sample is to be expected. Overall, the participants are encouraged to maximize the generalization performance and use the leaderboard only as a rough indicator of their model's performance.

Submission File

The list of predictions contains a list of pairs between protein targets and GO terms, followed by the probabilistic estimate of the relationship (one association per line). The target name must correspond to the target ID listed in the test set (in the FASTA header for each sequence). The GO ID must correspond to valid terms in GO's version listed in the Data section—invalid terms are automatically excluded from evaluation. Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) subontologies of GO are to be combined in the prediction files, but they will be evaluated independently and combined at the end as described above. The score must be in the interval (0, 1.000] and contain up to 3 (three) significant figures. A score of 0 is not allowed; that is, the team should simply not list such pairs. In case the predictions in the submitted files are not propagated to the root of ontology, the predictions will be recursively propagated by assigning each parent term a score that is the maximum score among its children's scores. Finally, to limit prediction file sizes, one target cannot be associated with more than 1500 terms for MF, BP, and CC subontologies combined.

For any protein ID in the test superset, you must list a set of GO terms and assign your estimated probability. If a protein ID is not listed in your submitted file, the organizers will assume that all predictions are 0. The file should not contain a header; columns must be tab-separated. An example submission file may look as follows:

```
P9WHI7   GO:0009274   0.931
P9WHI7   GO:0071944   0.540
P9WHI7   GO:0005575   0.324
P04637   GO:1990837   0.23
P04637   GO:0031625   0.989
P04637   GO:0043565   0.64
P04637   GO:0001091   0.49
etc.
```

The participants can manually investigate the UniProtKB entries for P9WHI7 and P04637 to familiarize themselves with biological databases.

Optional Free Text Prediction

Optionally, predictors may also include text in English that describes the function of any of the proteins in the test superset. The free text prediction task is optional. It will not be evaluated during the time of competition, it will not be included in the leaderboard calculation, and it will not be considered for winning prizes. Text predictions will be evaluated at a later time than GO term predictions, once a sufficient number of human-written textual paragraphs accumulate in UniProt (e.g., 9-12 months after the submission deadline). The assessment will be used to inform future directions of protein function prediction.

Each protein target is allowed up to five lines of free text that will combine to make the text paragraph. Each line of text may only contain ASCII printable characters (ASCII codes: 33-126), with the space (ASCII code: 32) character used as a (word) delimiter. ASCII printable characters include letters, digits, punctuation marks and symbols. The text paragraph cannot have any tabs in it. The list of text predictions should be in the following format: target name, followed by the word "Text" in the second field, followed by a probabilistic estimate of the text line, and lastly the text string. The text prediction for each protein is limited to 3,000 characters over all lines used for that protein, including spaces (longer submissions will be truncated to the first 3,000 characters). The breakdown of the entire textual description into up to five lines is to allow for differential confidence levels for different textual assertions, with up to five different confidence levels per protein. All limitations are imposed to control the overall size of the submission files and allow for an efficient accuracy assessment by the organizers.

Only one file can be submitted for both GO term and text prediction tasks; that is, GO term predictions should be combined with text predictions in a single submission file. If participants opt in for text prediction, an example submission file will look as follows:

```
P9WHI7   GO:0009274   0.931
P9WHI7   GO:0071944   0.540
P9WHI7   GO:0005575   0.324
P9WHI7   Text    0.123  P9WHI7 is involved in homologous recombinational repair, a high-fidelity pathway for fixing double-strand breaks. This pro
P04637   GO:1990837    0.23
P04637   GO:0031625    0.989
```

```
P04637   GO:0043565   0.64
P04637   GO:0001091    0.49
P04637   Text        0.234 Multifunctional transcription factor that induces cell cycle arrest, DNA repair or apoptosis upon binding to its target
P04637   Text        0.570 Interaction with BANP was reported to enhance phosphorylation on Ser-15 upon ultraviolet irradiation
P04637   Text        0.570 Regulates the circadian clock by repressing CLOCK-BMAL1-mediated transcriptional activation of PER22
```

Teams that choose to participate in the text prediction, but not in GO term prediction, can simply include only those lines that contain the word "Text" in the second field. They will not be scored in the GO term prediction.

Evaluation of Textual Predictions

The evaluation of textual predictions will contain two phases, which may depend on the number of participating teams and proteins that accumulate new textual descriptions between the submission deadline and the time of evaluation. In phase 1, large language models will be used to evaluate the accuracy of text paragraphs against human-written paragraphs; for example, in UniProt. The best teams will be identified using conventional metrics for text summarization. In phase 2, we anticipate that the human evaluators will compare paragraphs from the best teams identified in phase 1 against human descriptions to obtain the final rankings. Some of the lower-scoring teams in phase 1 may be randomly included in the human evaluation for calibration.

It is important to mention two different scenarios in which textual predictions will be evaluated. In the first scenario, a large body of literature may already exist about the function of a given protein in the public domain, but it may not have yet been summarized in UniProt at the time of the submission deadline. In this case, the evaluation is effectively testing for the quality of text summarization as it is a classical problem in the natural language processing community. When possible, the predictors should also include the traceable evidence for particular statements; e.g., using PubMed IDs of the corresponding publications as in the example submission above. In the second scenario, there may not be any literature in the public domain about the function of a particular protein. In those cases, the predictors must predict function based on sequence and any other available data (e.g., expression), from which text needs to be generated. The second scenario is different and potentially more difficult than the first scenario. It will be separated from text summarization to the extent possible during evaluation (e.g., participants can combine the two scenarios, which will need sophistication during assessment). Participants should note that protein function prediction is carried out in an open world; that is, certain predictions (of GO terms or sentences) may be correct, but the experimental data may not support them at the time of assessment.