CAFA3 Submission Rules and Format

(26-SEP-2016)

1. The evaluation will be performed for the target sequences that accumulate experimental annotations between the submission deadline (January 22, 2017) for predictions and the time of evaluation (July 2017) (and other times in later re-assessments). The initial evaluation will take place in July 2017, however, CAFA 3 assessors will carry out evaluations following this date, as more experimentally annotated results accumulate. We expect the final evaluation to done on or about October 2017. The evaluation will be carried out for all targets as well as in several separate categories. Some of the categories are expected to be identical or similar to those in the CAFA 2010-2011 experiment [1]. However, we will also conduct new evaluations for the proteins that contained incomplete experimental annotation prior to the submission deadline. Specifically, targets that were not experimentally annotated before the submission deadline in any of the major databases and across all ontologies (considered in CAFA) will constitute the first evaluation group. This part is identical to the evaluation in CAFA 2010-2011 [1]. The second group of targets will consist of proteins evaluated in one ontology (e.g. MFO) that before the submission deadline contained experimental annotation in another ontology (e.g. BPO and/or CCO). Finally, the third group of targets will consist of sequences that were experimentally annotated in one ontology (and potentially others) before the submission deadline and have accumulated additional experimental terms in the same ontology. Important Note: To facilitate participation of teams with different levels of access to computational resources each team will be evaluated in two modes: (1) on all targets that accumulated experimental annotation, and (2) only on the subset of targets for which the group submitted predictions, provided the group made automated annotation of at least 5000 proteins. In the first mode of evaluation, the targets on which the group does not submit predictions will be treated as incorrect predictions (appropriately for each evaluation metric). In the second mode, those targets will be ignored.

2. One team may test up to 3 different prediction models (named MODEL 1, MODEL 2, and MODEL 3) during submission. MODEL 1 will be officially evaluated by the organizers, but other models will also be considered. A group should use its best algorithm as MODEL 1.

3. A team may choose to predict using any of the following ontologies: MFO, BPO, CCO and (for human) HPO. The evaluation will be performed separately for each ontology. A team may choose to predict function using one or more of the above ontologies, and does not have to predict using all of them.

```
AUTHOR TEAM_NAME
MODEL  1
KEYWORDS literature, ortholog.
ACCURACY  1  PR=0.75; RC=0.31
ACCURACY  2  PR=0.55; RC=0.66
ACCURACY  3  PR=0.92; RC=0.51
T00001    GO:000173   1.00
T00203    GO:000123   0.01
T05203    GO:000123   0.91
.
.
END
```

Figure 1. File format for GO predictions.

```
AUTHOR TEAM_NAME
MODEL  1
KEYWORDS clinical data, synteny.
ACCURACY   1  PR=0.76; RC=0.32
ACCURACY   2  PR=0.92; RC=0.51
T00001     HP:0010268  0.71
T00001     HP:0010669  0.73
T00002     HP:0012050  0.90
T00003     HP:0012050  0.90
.
.
END
```

Figure 2. File format for HPO predictions.

4. The prediction output file format is shown in Figure 1. The file a group submits should be in text format (*.txt) or compressed (*.zip) text file. Predictions can be uploaded any number of times. The ones with the most recent time stamp in the system at the submission deadline will be used for evaluation.

- The AUTHOR line lists the team name that the team leader used during registration.
- The MODEL line contains numbers 1, 2, or 3 and corresponds to the prediction model used as described in bullet (2)
- The KEYWORDS line contains a list of keywords that describe the methodology used. Keywords line uses a comma-separated list, ending with a full stop, of one or more of the following pre-specified keywords: sequence alignment, sequence-profile alignment, profile-profile alignment, phylogeny, sequence properties, physicochemical properties, predicted properties, protein interactions, gene expression, mass spectrometry, genetic interactions, protein structure, literature, genomic context, synteny, structure alignment, comparative model, predicted protein structure, de novo prediction, machine learning, genome environment, operon, ortholog, paralog, homolog, hidden Markov model, clinical data, genetic data, natural language processing, other functional information.
- The ACCURACY lines are optional. If present, they must contain the group's estimate of the accuracy of their method for each of the three modes of evaluation (see (1) above). Each line contains estimated precision (PR) and recall (RC) for the Fmax point exactly as evaluated in the

CAFA 2010-2011 manuscript [1]. The number indicates one of the three evaluation modes as described under point (1) above. Both numbers must be in the interval [0.00, 1.00]. Two significant figures are required (e.g. 0.70 is valid but 0.7 or .70 are not). The ACCURACY lines may be different in each submitted file (all targets are broken up based on species). If so, a weighted average will be used to estimate the final accuracy of the model. Weights will be determined by the number of proteins from each target file that accumulate experimental terms between the submission deadline and the time of evaluation.

- 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 target files (in the FASTA header for each sequence). The Gene Ontology ID must correspond to valid terms in GO from June 1, 2016. MFO, BPO, and CCO are to be combined in the prediction files, but they will be evaluated independently by the assessment team. The score must be in the interval (0.00, 1.00] and contain two significant figures. A score of 0.00 is not allowed; that is, the team should simply not list such pairs. In case the predictions are not propagated to the root of ontology, the assessors will recursively propagate them 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 MFO, BPO, and CCO combined (same for HPO submissions). The assessors will provide software so the groups will be able to check the format of their prediction files. Please submit only files that are verified for correctness. The assessors will not analyze submissions that are in incorrect format. If your method does not output a score associated with predicted terms, but rather just a set of terms, the team can set scores for all such predictions to the same value (e.g. 1.00). Such methods will be characterized by a single precision/recall point, instead of a precision/recall curve.
- The Human Phenotype Ontology annotations for human targets shall be submitted separately using the HPO annotation. HPO build #1701 shall be used July 4, 2016. Figure 2 shows the format for HPO submissions. Only two accuracy estimations are necessary: one for the proteins/genes that have not been associated with any HPO terms before the submission deadline; and the other for the remaining targets.
- The prediction file must end with the keyword END in a line of its own.
- Allowed delimiters are tab and whitespace only.
- Prediction file name format:

Format for GO predictions

```
Use team ID, model number, and taxon IDs as follows:
teamID_modelNo_taxonID.{txt/zip}

Example for team Doe group, model 1, human sequences GO prediction
file name format: (doegroup_1_9606.txt)

Example for team Doe group, model 3, mouse sequences GO prediction
file name format: (doegroup_3_10090.txt.zip)
```

Format for HPO predictions

```
Use team ID, model number, and acronym HPO as follows:
teamID_modelNo_hpo.{txt/zip}

Example for team Doe group, model 2, HPO prediction:
(doegroup_2_hpo.txt)
```

5. Prediction and evaluation types (new in CAFA3)

5.1 **Protein-centric predictions**

A protein-centric prediction addresses the following question: “given a protein, what are all the ontology terms associated with it” (see Radivojac et al. *Nat. Methods*, 2013 and Jiang et al. *Genome Biol.* 2016). This is the main mode of evaluation in CAFA.

5.2 **General term-centric predictions**

A term-centric prediction means: “given a specific ontological term, which genes in an organism fit that term?” All protein centric predictions will also be evaluated in term-centric mode. However, predictors can submit files with term-centric predictions and ask not to be evaluated in the protein-centric mode. **The filenames of predictions designated for term-centric only evaluations *must* be prefixed with “TC_”, or they will be evaluated both term-centric and protein-centric.** For example, if a team has an algorithm that predicts beta-amylases, they can submit a file that *only* predicts what genes in a given organism are beta amylases. For more about term-centric evaluations, see the CAFA1 and CAFA2 papers, as well as http://biofunctionprediction.org/cafa-targets/Introduction_to_protein_prediction.pdf

5.3 **Specific term-centric predictions**

In addition, we will be experimentally screening three organisms for specific functions. The organisms are *Drosophila melanogaster, Candida albicans*, and *Pseudomonas aeruginosa*. For each of these genomes, we will be collecting protein-centric predictions (as for the rest of the CAFA files), but we will also be checking for specific term-centric predictions. The screens are

being run by Deborah Hogan's and Gio Bosco's labs in the Geisel School of Medicine, Dartmouth.

The predictions for the following specific terms will evaluated in the following organisms:

Pseudomonas aureginosa UCBPP-PA14
Biofilm formation GO:0042710
Motility: GO:0001539
Phenazine biosynthetic process: GO:0002047

Candida albicans SC 5314 Assembly 22:
Biofilm formation GO:0042710
Hyphal growth: GO:0030448
Cell growth mode switching, budding to filamentous: GO:0036187

Drosophila melanogaster:
Long term memory: GO:0007616

5.4 Moonlighting proteins

"A moonlighting protein is a single protein that has multiple functions that are not due to gene fusions, multiple RNA splice variants or multiple proteolytic fragments." (Taken from: http://www.moonlightingproteins.org/ where you can read more about moonlighting proteins.) We provide a set of 40 moonlighting proteins, for which one or more functions are yet undocumented, yet have been experimentally determined. The moonlighting proteins are provided courtesy of Constance Jeffery, University of Illinois, Chicago.

5.5 Binding site predictions

In CAFA3 we are introducing predictions of macromolecular binding sites in a protein (DNA, RNA) and metal binding sites. The predictions can be made for any CAFA target, but the evaluation will be performed only on proteins that gained experimental validation of binding sites after the CAFA3 submission deadline in January 2017.

5.5.1 Prediction format for binding site predictions

The binding site predictions should be provided for each protein in the following format:

```
AUTHOR TEAM_NAME
MODEL 1
KEYWORDS phylogeny, literature
>T12345
TYPE
s1, s2,..., sn
TYPE
s1, s2,..., sn
.
.
END
```

Where the line prefixed by ">" is the CAFA ID of the protein. TYPE is one of: DNA, RNA, or METAL (at this time, we are grouping all metals into a single field). The line below the type line is *a single physical line* comprising n scores separated by commas, n being the length of that protein. Each number s_i is a confidence value [0,1] and has exactly two significant digits (e.g. 0.40, not 0.4).

Example:

```
AUTHOR ARISTOTLE_TEAM
MODEL 1
KEYWORDS phylogeny, literature
>T123567
RNA
0.00, 0.00, 0.11, 0.50, 0.80, ..., 0.00
DNA
0.00, 0.01, 0.52, 0.86, 0.12, ...,0.54
>T456789
DNA
0.00, 0.01, 0.51, 0.81, ...,0.50
RNA
0.41, 0.41, 0.42, 0.24, ...,0.34
METAL
0.00, 0.00, 0.92, 0.01, ..., 0.03
.
.
END
```

The header (AUTHOR, MODE, KEYWORDS, ACCURACY records) and footer (END) of the file are the same as for the other CAFA files, see Fig. 1.

6. For the policies on data sharing, anonymity, de-identification, or withdrawal from the entire experiment (after the submission deadline) see the Data Sharing, Anonymity, and Withdrawal policy page.

7. A group can appeal their prediction evaluation to the CAFA assessment team, and to the CAFA organizers. Appeals will be discussed and ruled upon by the assessment team and the organizers. All such rulings are final.