Manual annotation of a subset of the CAFA3 target set.

What is the CAFA Challenge?
The Critical Assessment of protein Function Annotation algorithms (CAFA) challenge is a large-scale assessment whose purpose is to evaluate new computational methods that are capable of predicting Human Phenotype Ontology and Gene Ontology (GO) terms for proteins, based on their sequence or structure.

What are Intrinsically Disordered and Moonlighting Proteins?
For the latest CAFA challenge (CAFA3), the Gene Ontology Annotation (GOA) team has contributed and curated two target sets: the first consists of intrinsically disordered proteins, and the second consists of moonlighting proteins.

Intrinsically disordered proteins (IDP) are proteins that lack a fixed or ordered 3D structure. They range from fully unstructured to partially structured and include random coils, molten globules, and flexible linkers. These are identified using structural or biophysical techniques. A good example is p53, which is partially structured.

Moonlighting proteins are proteins that can perform more than one function that are not due to gene fusions, RNA splice variants or proteolytic degradation. For example, ERK2, whose primary function is being a MAP kinase, also has a secondary function as a transcriptional repressor.

How were targets sets created and curated?
To create our target sets, we used data from the DisProt database for intrinsically disordered proteins (http://www.disprot.org/) and the MoonProt database of moonlighting proteins (http://www.moonlightingproteins.org/) to generate a potential list of proteins for CAFA. For each candidate protein we found appropriate literature to provide experimental evidence of whether it was intrinsically disordered or had a secondary function. The decision trees below were used while reading the literature to determine whether a protein fits the criteria of being an IDP or moonlighting protein as well as whether a paper was suitable for GO curation.

How will these annotations be used?
After evaluation, the IDP dataset comprised 472 proteins and the moonlighting dataset comprised 156 proteins. In total, 766 papers were curated, which resulted in the creation of 6981 new GO annotations to be used as a benchmark for evaluating the predictions submitted by CAFA3 contestants. The submitted annotations will be scored against the manually curated annotations to see how well contestants’ prediction algorithms performed.

Where can I access the dataset?
The annotations for the target sets are publicly accessible through the QuickGO website (http://www.ebi.ac.uk/QuickGO/).

To access the dataset in QuickGO, Select ‘View GO Annotations’ from the main page, and select CAFA (Critical Assessment of Protein Function Annotation) in the ‘Assigned by’ column.

This work is sponsored by a joint grant from the BBSRC and the NSF.