Protein function prediction using homologs

Joseph Bonello

July 16, 2017

Plan

- Introduction
- Set Based Methods
- dcGO-inspired method for CATH
- Next Steps

Motivation

- Newer sequencing machines become more efficient, available and affordable
  - Led to a rapid increase in sequencing and structural genomics initiatives
  - This has led to the number of sequences for analysis have thus been increasing rapidly
  - But also provide us with the opportunity to understand better the complex sequence, structure and function relationships in proteins
- However, an analysis of UniProtKB/Swiss-Prot sequence database (June 2015) shows that less than 15% of human proteins have detailed functional characterisation and only 4% have known structures

Project Summary

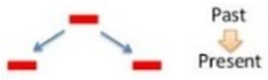
- CATH uses a method called FunFamer that predicts functions for uncharacterised proteins
- Aim to increase its power by combining it with another approach
- Generate scores for presence of GO Terms in CATH Superfamilies and CATH Functional Families
- Four methods are being tested
    - Scores using Jaccard index
    - Scores using Sørensen Index
    - Scores using the Overlap Index
    - Scores using the dcGO inspired method

Background I

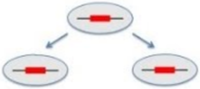
- A protein domain is a distinct functional and/or structural unit of a protein.
- Many proteins consist of several structural domains, and one domain may appear in a variety of different proteins.
- We think that there are less than 10,000 domains covering most proteins in nature.
  - However, there are millions of domain combinations.
  - It is not feasible to characterise all of these combinations.
  - We want to develop a domain grammar of function.

Background II

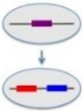
- A Homolog is a gene related to a second gene by descent from a common ancestral DNA sequence as a result of a separation event



  - **Orthologs**: descend from a common ancestor by genome division (speciation)

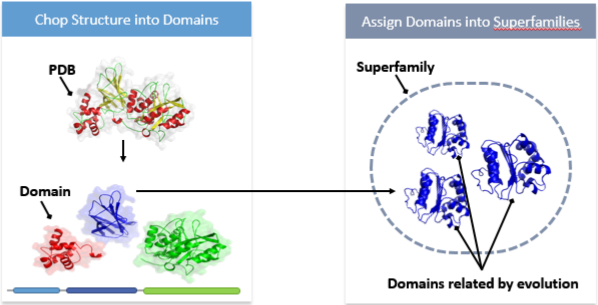


  - **Paralogs**: descent from a common ancestor by duplication within a genome (genetic duplication)



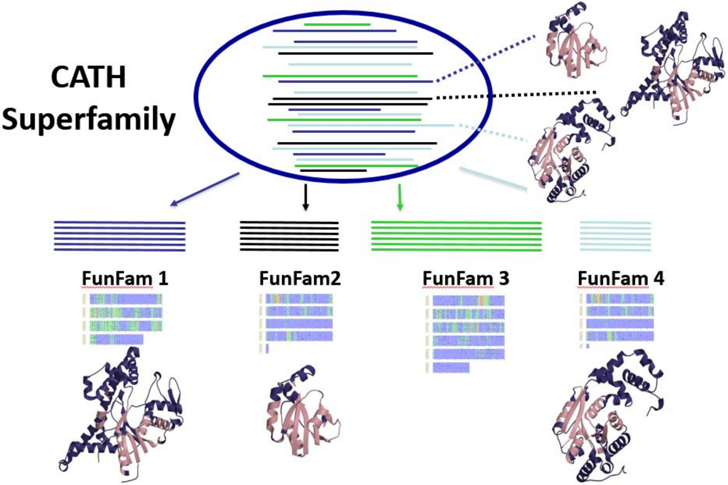
CATH I

- CATH – A Protein Structure Classification Database maintained by the Orengo Group at UCL
- CATH algorithm chops the domains of a protein and assigns them to a superfamily
- Superfamily domains are related by evolution
- This gives us a neat, but solid, grouping of domains


Chop Structure into Domains
PDB
Domain
Assign Domains into Superfamilies
Superfamily
Domains related by evolution

CATH II

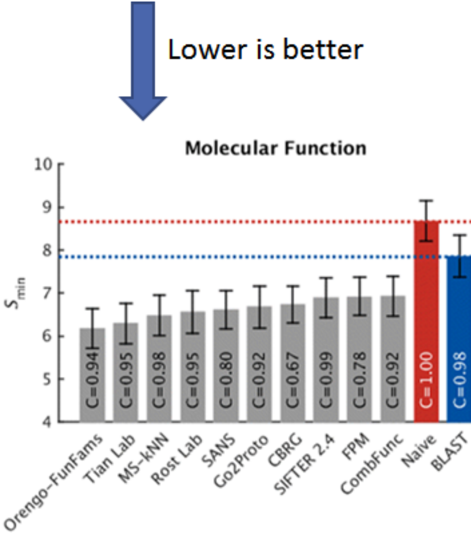
Superfamilies are sub-classified into Functional Families (Funfams)



2638 superfamilies -> ~100,000 functional families (FunFams)

FunFamer

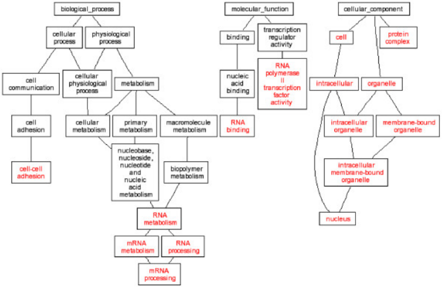
- CATH uses a method called FunFamer to predict functions for uncharacterised proteins
- FunFamer is based on Residue Specificity Information It obtained very good results in the CAFA 2 challenge
- We want to increase the power of FunFamer by combining it with another approach



Lower is better

Gene Ontology I

- The Gene Ontology (GO) attempts to formalise a **standardised "controlled vocabulary of terms"** to describe various aspects of biological systems
- Organised as a **Directed Acyclic Graph (DAG)** whose structure expresses the specialisation of child terms



biological_process

cellular process

physiological process

cell communication

cellular physiological process

metabolism

cell adhesion

cellular metabolism

primary metabolism

macromolecule metabolism

cell-cell adhesion

nucleobase, nucleoside, nucleotide and nucleic acid metabolism

biopolymer metabolism

RNA metabolism

mRNA metabolism

RNA processing

mRNA processing

molecular_function

binding

transcription regulator activity

nucleic acid binding

RNA polymerase II transcription factor activity

RNA binding

cellular_component

cell

protein complex

intracellular

organelle

intracellular organelle

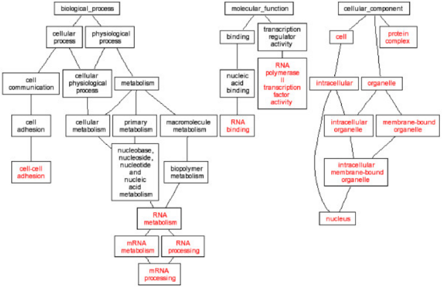
membrane-bound organelle

intracellular membrane-bound organelle

nucleus

Gene Ontology II

- Organised into three sub-ontologies, namely **Molecular Function**, **Cellular Component** and **Biological Process**
- The topmost terms of the DAG are the most generic, becoming increasingly specific as we move down the DAG



Evidence Codes I

Evidence Code	Description
Experimental Evidence Codes	
EXP	Inferred from Experiment
IDA	Inferred from Direct Assay
IPI	Inferred from Physical Interaction
IMP	Inferred from Mutant Phenotype
IGI	Inferred from Genetic Interaction
IEP	Inferred from Expression Pattern

Evidence Codes II

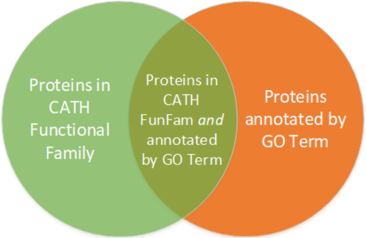
Evidence Code	**Description**
Computational Analysis Evidence Codes	
ISS	Inferred from Sequence or structural Similarity
ISO	Inferred from Sequence Orthology
ISA	Inferred from Sequence Alignment
ISM	Inferred from Sequence Model
IGC	Inferred from Genomic Context
IBA	Inferred from Biological aspect of Ancestor
IBD	Inferred from Biological aspect of Descendant
IKR	Inferred from Key Residues
IRD	Inferred from Rapid Divergence
RCA	Inferred from Reviewed Computational Analysis

Evidence Codes III

Evidence Code	Description
Author Statement Evidence Codes	
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
Curatorial Statement Evidence Codes	
IC	Inferred by Curator
ND	No biological Data available
Automatically Assigned Evidence Codes	
IEA	Inferred from Electronic Annotation

Notation

- We want a method that generates scores based on the **association** of domains in a family with functional annotations in the GO database
- The first set contains proteins in specific Superfamily/FunFam
- The second set represents proteins annotated by a particular GO term
- The *intersection* represents all proteins that are within a Superfamily/FunFam and are annotated by a specific GO term
- The *union* represents all proteins in a Superfamily/FunFam and which are annotated by a particular GO term


Proteins in CATH Functional Family
Proteins in CATH FunFam *and* annotated by GO Term
Proteins annotated by GO Term

The Jaccard Index

- This index provides a comparative score of two sets
- Takes into consideration the members that exist in either set ($A \cup B$), and
- The members that exist in both sets ($A \cap B$)
- Index scores:
  - 0 indicates highly dissimilar sets, while
  - 1 indicates highly similar sets

Jaccard Index Equation

$$J_{AB} = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

The Sørensen-Dice Index

- This index also provides a comparative score of two sets
- It is similar to the Jaccard Index, and provides similar scores
- It is less susceptible to outliers but maintains sensitivity in heterogeneous data sets
- Index scores:
  - 0 indicates highly dissimilar sets, while
  - 1 indicates highly similar sets

Sørensen-Dice Index Equation

$$SD_{AB} = \frac{2|A \cdot B|}{|A|^2 + |B|^2} \tag{2}$$

The Overlap Index

- This index measures the overlap between two sets
- It considers the size of the intersection with respect to the smaller of the two sets being considered
- Scores sets based on how much they overlap, i.e. based on the size of their common members
- Like the Jaccard and Sørensen Indexes, this index scores:
    - 0 indicating highly dissimilar sets, while
    - 1 indicating highly similar sets

Overlap Index Equation

$$O_{AB} = \frac{|A \cap B|}{min(|A|, |B|)} \tag{3}$$

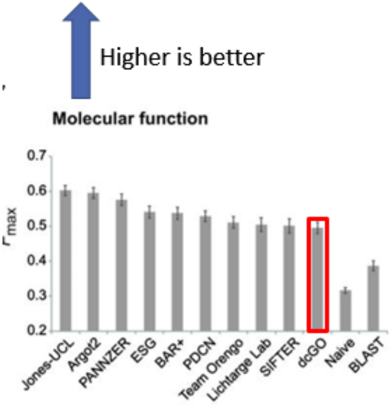
dcGO Background I

As part of our aims, we want to develop a meta-predictor that combines FunFamer with the dcGO (Fang and Gough, 2013) algorithm for CATH

- dcGO (Fang and Gough, 2013) is a statistical approach for scoring the correspondence betwen domains and GO terms
- It takes the following into consideration:
    1. **structural domains** correspond to a **functional unit** of a protein, and therefore GO terms are more likely to correspond to a domain than a whole protein;
    2. if **a domain has more proteins** annotated with a particular GO-term than one would expect **by chance**, then it is possible to **infer functional GO associations**

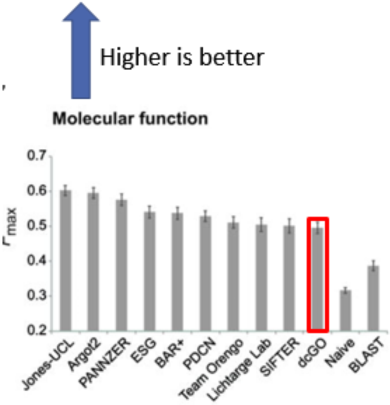
dcGO Background II

- dcGo placed in the top ten methods in CAFA 1, but not as good as FunFamer
- It is the only other method that is domain-based (scalable methods)



dcGO Background III

- dcGO is based on SCOP and SUPERFAMILY databases
- Uses statistical approach to score similarity between domains and GO Terms
- The idea is to use homology to inherit from a characterised protein to an uncharacterised protein



dcGO for CATH I

- dcGO4CATH is an adaptation of this algorithm. It uses two databases to build its scores:
  1. GO annotation for proteins from **UniProtKB-GOA** (experimental or manual evidence codes only are considered from this database to reduce false-positives);
  2. The **CATH database** which contains proteins that are classified as belonging to particular Superfamilies and Functional Families (FunFams)

dcGO for CATH II

- The dcGO4CATH algorithm makes use of a pre-computed matrix, called the **correspondence matrix**
- The matrix stores the observed number of proteins for each GO/(Superfamily/Funfam) domain combination
- The next step in the process is to calculate the **overall and relative p-values** using the hypergeometric distribution
  - It is used to determine the probability of a particular GO term annotates a Superfamily/FunFam and
  - that this probability is not due to chance
- The p-values are then **adjusted using a Benjamini-Hochberg False Discovery Rate (FDR)** correction as a means of discovering the rate of Type I errors (false positives)

Homologs, Orthologs and Paralogs I

A reminder

- A **Homolog** is a gene related to a second gene by descent from a common ancestral DNA sequence either through a separation event due to speciation (ortholog) or separated by genetic duplication (paralog);
- **Orthologs** are genes in different species that have evolved from a common ancestral gene through speciation and which retain the same function;
- **Paralogs** are genes related by duplication within a genome. Paralogs evolve new function even if they are related to the original gene

Homologs, Orthologs and Paralogs II

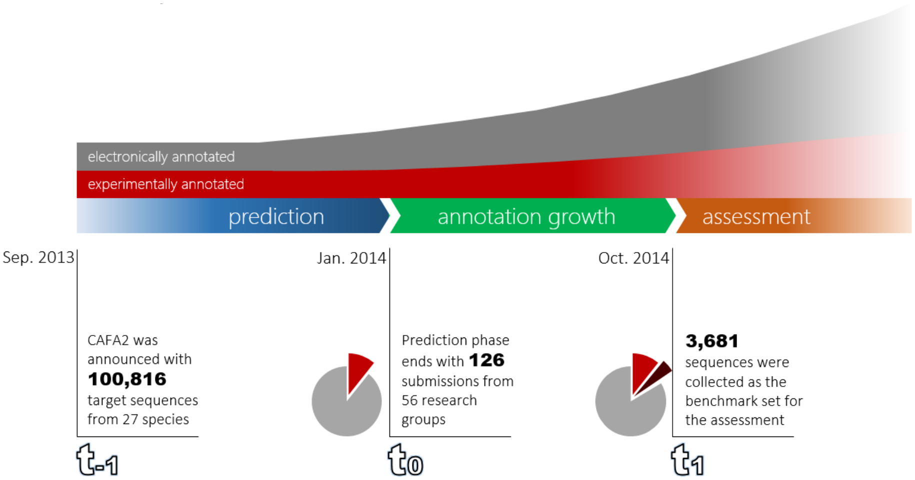
- dcGO is based on coarse clusters of homologs from SCOP
- This makes it weak against paralogs which have modified functions
- Using CATH FunFams we expect to be able to counter this as the subclassification of CATH FunFams is **more refined**, hence we expect it to lead to improved scoring

Critical Assessment of Function Annotation (CAFA) I

- A **community challenge** whose goal is to **improve the understanding** of the current state of computational protein function prediction (Friedberg and Radivojac, 2017)
- CAFA is a timed challenge:
  - During the first stage, a large collection of un-annotated proteins are publicly released, during which time predictors apply their methods
  - During the second stage, biocurators add their results to specialised databases (Swiss-Prot or UniProt-GOA)
  - During the third stage, organisers collect annotations produced in stage two, and analyse the results of the predictions against these results

Critical Assessment of Function Annotation (CAFA) II

- Benchmarks used for evaluation use the CAFA Challenge datasets
- It is a challenge to assess the computational methods used to predict



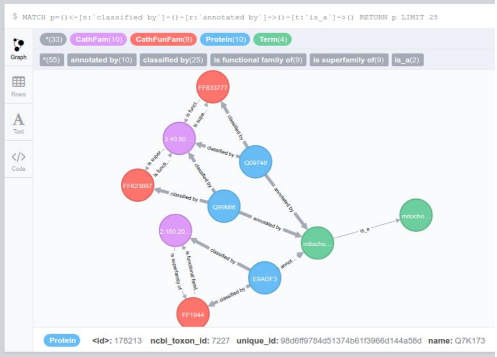
CAFA2 was announced with **100,816** target sequences from 27 species

Prediction phase ends with **126** submissions from 56 research groups

3,681 sequences were collected as the benchmark set for the assessment

Current Setup I

1. Implemented a Graph database (using Neo4J) that contains:
    1. Gene Ontology terms linked
    2. Superfamilies and Functional Families linked
    3. “Relatedness” scores between functional families
    4. Proteins annotated by terms in GO
    5. Proteins classified by Superfamily/Functional Family



```
$ MATCH p=()<-[s:`classified by`]-()-[r:`annotated by`]->()-[t:`is_a`]->() RETURN p LIMIT 25
```

Protein <id>: 178213 ncbi_toxon_id: 7227 unique_id: 98d6ff9784d51374b61f3966d144a58d name: Q7K173

Current Setup II

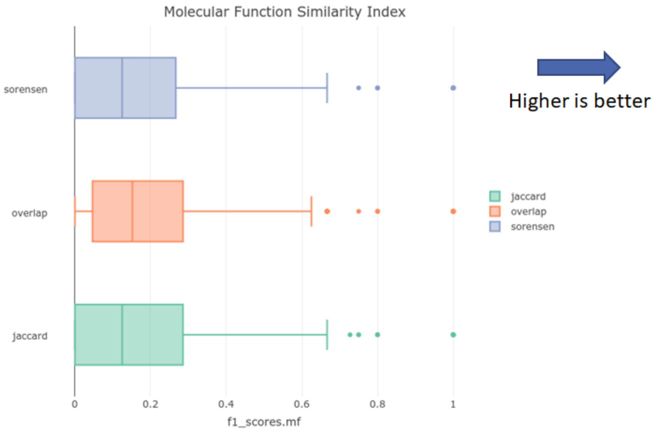
1. Implemented Jaccard Index
2. Implemented Sørensen-Dice Index
3. Implemented Overlap Index
4. Implemented dcGO for CATH (currently generating scores using this method of Superfamilies)
    - Reimplementing the solution to use ALBERT at UoM

Preliminary Results I

1. Benchmarking. The methods were benchmarked using data from CAFA 1 and compared to the results obtained in that challenge.
2. This provides a metric of the strength of the methods with respect to their capability to predict protein function based on the data provided by CATH.

Preliminary Results II

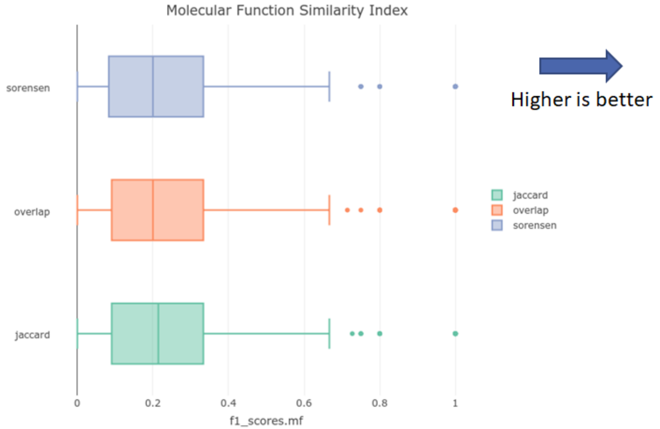
Analysis of the results shows that the predicted terms were too specific, hence low accuracy



Higher is better

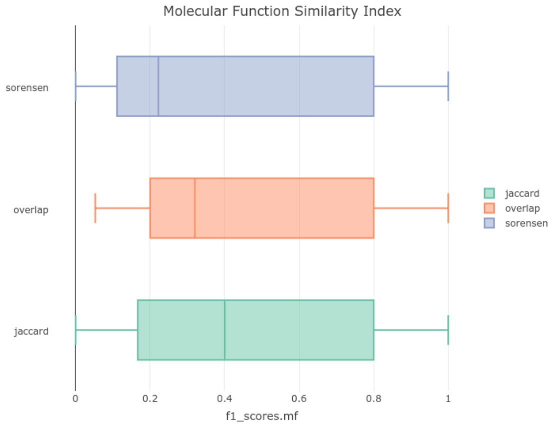
Preliminary Results III

First optimisation performed was to return more generic GO Terms



Preliminary Results IV

Second optimisation performed was to update the scores using experimental annotations only



Future work

1. Optimisation of scoring methods
    - Compare parental GO terms for both predicted and Gold Standard terms
2. Broaden the data used in generating the scores to include data from Gene3D
3. Determine whether the methods can complement methods based on inheritance by homology or whether there is an overlap with these existing methods
4. Investigate a meta-predictor which combines dcGO-inspired methods and FunFamer

References I

Das, S., N. L. Dawson, and C. A. Orengo (2015). "Diversity in protein domain superfamilies". In: *Current Opinion in Genetics and Development* 35, pp. 40–49. ISSN: 18790380.

Das, S. et al. (2015a). "CATH FunFHMMer web server: protein functional annotations using functional family assignments.". In: *Nucleic acids research* 43.W1, W148–53. ISSN: 1362-4962.

Das, S. et al. (2015b). "Functional classification of CATH superfamilies: A domain-based approach for protein function annotation". In: *Bioinformatics* 31.21, pp. 3460–3467.

Fang, H. and J. Gough (2013). "A domain-centric solution to functional genomics via dcGO Predictor". In: *BMC Bioinformatics* 14 Suppl 3.Suppl 3.

References II

Friedberg, I. and P. Radivojac (2017). “Community-Wide Evaluation of Computational Function Prediction”. In: *The Gene Ontology Handbook*. Ed. by Christophe Dessimoz and Nives {\v{S}}kunca. Springer New York, pp. 133–146.

Radivojac, P. et al. (2013). “A large-scale evaluation of computational protein function prediction”. In: *Nature Methods* 10.3, pp. 221–227. ISSN: 1548-7091.

Sillitoe, I. et al. (2015). “CATH: comprehensive structural and functional annotations for genome sequences”. In: *Nucleic Acids Research* 43.D1, p. D376.

Thank you!