AUTOMATING ONTOLOGICAL FUNCTION ANNOTATION: TOWARDS A COMMON METHODOLOGICAL FRAMEWORK

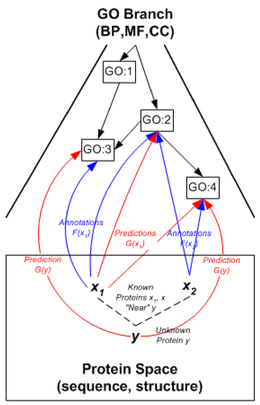
Cliff A Joslyn, Judith D Cohn, Karin M Verspoor, and Susan M Mniszewski
Los Alamos National Laboratory

Motivation

- Annotate protein function as GO node assignment
- Map previously unknown proteins to GO nodes
- Construct mappings from sequence, structure, literature, and/or pathways space to GO function space
- Some existing approaches:
  - **Proknow**: Pal and Eisenberg (2005): Set of protein sequences from the FSSP structure library
  - **GOtcha**: Martin *et al* (2004): Sequence data from seven complete genomes
- Our approach:
  - Determine near BLAST neighbors of unknown proteins
  - Select GO node(s) using the POSOC categorization algorithm
- Questions:
  - How do we know how well we did?
  - How do we measure performance in the context of the particular properties of the Gene Ontology?

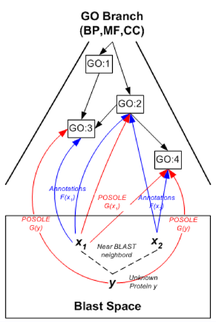
Generic Automated Ontological Protein Function Annotation

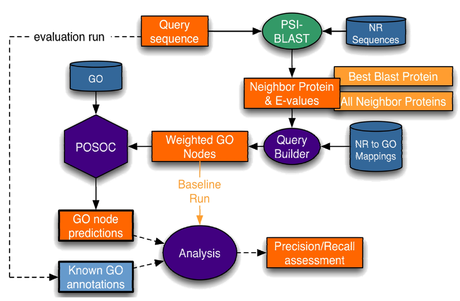
- Known proteins *x*, unknown proteins *y*
- Each protein *x* has known annotations *F(x)*, a set of GO nodes
- Induce new set of GO nodes *G(y)* for unknown protein
- Testing: compare predictions of known proteins G(x) against known annotations F(x)
- ***ISSUES:***
  - How to identify known proteins *x*?
  - How to identify annotation mappings *F(x)* of known proteins?
  - How to compare *F(x)* against *G(x)*: ***generalized precision and recall***
    - When *F* and *G* live in the GO structure?
    - When *G(x)* might return a ranked list?
  - How to account for "near misses" in the GO?
  - How to measure the "spread" and "location" of result sets *F(x),G(x),G(y)* in the GO?



POset Ontology Laboratory Environment (POSOLE)

- General environment for ontology experimentation
  - Graph representation of an ontology as a partially ordered set (poset)
  - Poset statistics analysis (e.g. depth, width, average rank)
  - Algorithms for node categorization utilizing the structure of the ontology
- **First Deployment**: Ontology categorization for automated protein function annotation
  - Function: Gene Ontology node
  - Protein: target sequence or Swiss-Prot identifier
  - Map proteins to sets of potential Gene Ontology nodes
  - Ontology categorization: "clustering" nodes in ontology space to identify the most likely node assignment





POSOC

`http://www.c3.lanl.gov/posoc`

Contact

Cliff Joslyn
MS B265 LANL
Los Alamos, NM 87545
joslyn@lanl.gov
+1-505-667-9096

Protein Test Set

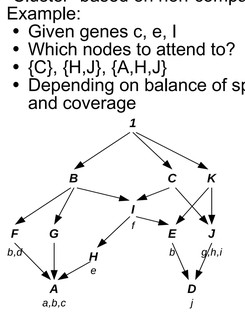
- **NEED**: Select one or more "gold standard" test sets *X* of proteins with trusted annotations in the GO to be used for performance evaluation
- **ISSUE**: Test sets should be non-redundant and should evenly represent the test space
- **GOAL**: A nonredundant test set covering GO function space accepted by the community to support comparative evaluation across systems
- **POSOLE**: 4530 Swiss-Prot protein sequences with both known PDB structures and known GO annotations

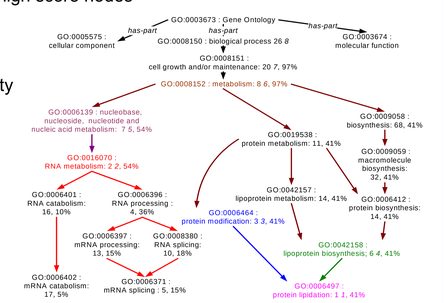
Annotation Mappings

- **ISSUE**: Which annotation mappings to use?
- **ISSUE**: Community standard to provide a means of comparing various studies
- **ISSUE**: Filtering on annotation evidence codes (e.g. IC = inferred by curator vs. IEA = inferred from electronic annotation) may be necessary to support evaluation over only trusted data
- **ISSUE**: Common ranking of the evidence codes can be used to assess annotation quality (Pal and Eisenberg 2005)
- **POSOLE**: GOA UniProt annotation set for SwissProt protein sequences, used for both neighbor mappings to GO annotations

POSet Ontology Categorizer (POSOC)

- Joslyn *et al.* 2004: Given the Gene Ontology (GO) . . . And mappings to GO nodes . . .
- "Splatter" them over the GO . . . Where do they end up?
  - Concentrated? -- Dispersed?
  - Clustered? -- High or low?
  - Overlapping or distinct?
- Pseudo-distances between comparable nodes to measure vertical separation
- POSOC traverses the structure of the GO, percolating hits upwards, and calculating scores for GO nodes.
- Scores to rank-order nodes with respect to gene locations, balancing:
  - **Coverage**: Covering as many genes as possible
  - **Specificity**: But at the "lowest level" possible
- "Cluster" based on non-comparable high score nodes
- Example:
  - Given genes c, e, I
  - Which nodes to attend to?
  - {C}, {H,J}, {A,H,J}
  - Depending on balance of specificity and coverage





References

- CA Joslyn: (2004) "Poset Ontologies and Concept Lattices as Semantic Hierarchies", in *Conceptual Structures at Work, LNAI*, v. 3127, ed. Wolff *et al.*, pp. 287-302, Springer-Verlag, Berlin
- CA Joslyn and WJ Bruno: (2005) "Weighted Pseudo-Distances for Categorization in Semantic Hierarchies", *2005 Int. Conf. on Conceptual Structures*, to appear in *Lecture Notes in AI*
- CA Joslyn, SM Mniszewski, AW Fulmer and GG Heaton: (2004) "The Gene Ontology Categorizer", *Bioinformatics*, v. 20:s1, pp. 169-177
- S Kiritchenko, S Matwin, and AF Famili: (2005) "Functional Annotation of Genes Using Hierarchical Text Categorization", to appear in Proc. BioLINK SIG on Text Data Mining
- D Martin, M Berriman, and G Barton: (2004) "GOtcha: A New Method for Prediction of Protein Function Assessed by the Annotation of Seven Genomes". *BMC Bioinformatics* 5:178
- D Pal and D Eisenberg, David: (2005) ``Inference of Protein Function from Protein Structure", *Structure*, v. 13, pp. 121-130
- KM Verspoor, JD Cohn, SM Mniszewski, and CA Joslyn: (2004) "Nearest Neighbor Categorization for Function Prediction" In CASP 06 abstract book.
- KM Verspoor, JD Cohn, CA Joslyn, SM Mniszewski, A Rechtsteiner, LM Rocha, and T Simas: (2005) :"Protein Annotation as Term Categorization in the Gene Ontology Using Word Proximity Networks", *BMC Bioinformatics* 2005 vol 6(suppl 1)

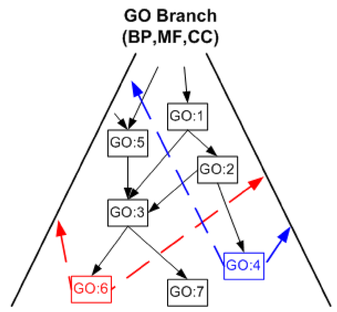
POSOLE Evaluation Runs

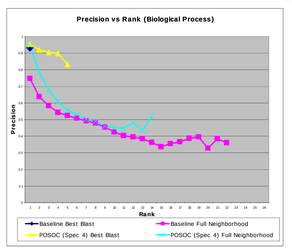
- **Baseline Best BLAST**: GO nodes associated with non-identical protein scoring highest in the PSI-BLAST analysis (all rank 1)
- **Baseline Full Neighborhood**: GO nodes associated with *all* proteins matched in the PSI-BLAST analysis (evalue < 10); ranked by evalue of the corresponding PSI-BLAST match
- **POSOC Best BLAST**: Inputs to POSOC are GO nodes associated with non-identical protein scoring highest in the PSI-BLAST analysis, weighted by evalue of the match. POSOC categorizes and ranks these inputs to produce the predictions.
- **POSOC Full Neighborhood**: Inputs to are the GO nodes associated with **all** proteins matched in the PSI-BLAST analysis, weighted by evalue of the match. POSOC categorizes and ranks these inputs to produce the predictions

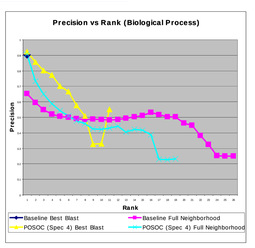
Hierarchical Evaluation Metrics

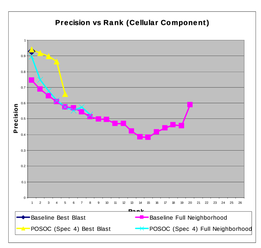
- Compare answers *F(x)* against predictions *G(x)*
- Precision = $\frac{|F(x)\cap G(x)|}{|G(x)|}$, Recall = $\frac{|F(x)\cap G(x)|}{|F(x)|}$
- But how do you calculate $F(x)\cap G(x)$ in the GO?
- When does a GO node *p* in *F(x)* count as a "match" against a *q* in *G(x)*?
- What if *p* matches *q* whenever *p* is an ancestor of *q* in the GO?
- But what about siblings? Don't "near misses" count?
- Adapt approach of Kiritchenko *et al.* 2005:

$$P = \sum_{q\in G(x)} \max_{p\in F(x)} \frac{|\uparrow p \cap \uparrow q|}{|\uparrow q|} \qquad R = \sum_{p\in F(x)} \max_{q\in G(x)} \frac{|\uparrow p \cap \uparrow q|}{|\uparrow p|}$$









Ontology Distance Metrics

- How "far apart" are nodes *p* and *q*?
- "Geneological" approach:
  - Radius 0: **Equals**: Direct match
  - Radius 1: **Nuclear family**: Parents, children, siblings
  - Radius 2: **Extended family**: Grandparents, grandchildren, cousin, aunt/uncle, niece/nephew
- Towards a general formulation of metric-based poset distances and evaluation functions: under development (Joslyn and Bruno 2005)