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: Fal and Eisenberg (2005): Set of protein sequences from the FSSP database
  - GoChica: Martin et al (2004): Sequence data from seven complete genomes

Protein Test Set
- NEED: Select or one more "gold standard" test sets ξ 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 non-redundant test set covering GO function space accepted by the community to support comparative evaluation across systems
- POSOC: 4530 Swiss-Prot protein sequences with both known PDB structures and known GO annotations

Annotation Mappings
- ISSUE: Which annotation mappings to use?
- ISSUE: Common standards to mean's of comparing various studies
- ISSUE: Filtering on annotation evidence codes (e.g., IC = inferred by curator vs. IEA = inferred from electron 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 (Fal and Eisenberg 2005)
- POSOC: GO UniProt annotation set for Swiss-Prot protein sequences, used for both neighbor mappings to GO annotations

POSet Ontology Categorizer (POSOC)
- Joslyn et al (2003): Given the Gene Ontology (GO) and mappings to GO nodes:
  "Splitter" 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
  - POSETC categorizes the structure of the GO, percolating hits upwards, and calculating scores for GO nodes.
  - Scores to rank-order nodes with respect to gene functions, 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, d, e
    - Which nodes to attend to?
    - (C), (H), (A-H), (H)
    - Depending on balance of specificity and coverage

POSET Ontology Laboratory Environment (POSOLE)
- General environment for ontology experimentation
  - Graph representation of an ontology as a partially ordered set (poset)
  - Protein labelations as (g, w) within average rank
  - Algorithms for node categorization utilizing the structure of the ontology
  - First deployment: Ontology categorization for automated protein function annotation
  - Second deployment: Ontology node labeling
  - Protein label 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

References
- Pali and G Eisenberg: David, (2000) "Discrete Pairwise Prediction of Protein function from Protein Structure", Structure 8, 13, pp. 121-130
- KM Verspoor, JD Cohn, CA Joslyn, SM Mniszewski, A Reichsteiner, LM Rocha, and T Sinex: (2005) "Protein Annotation as Term Categorization in the Gene Ontology Using Word Proximity Networks", BMC Bioinformatics 6(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 a 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{\text{Correct \ predictions}}{\text{Predictions}}
- Recal = \frac{\text{Correct \ predictions}}{\text{Annotations \ in \ GO}}
- But how do you calculate f(x) = g(x) in the GO?
- When does a GO node p in G(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 Krititchenko et al. 2005:
- Precision vs Rank (Cellular Component)

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

Contact
Cliff Joslyn
M5 8265 LANL
Los Alamos, NM 87545
joslyn@lanl.gov
-503-667-5096
http://www.c3.lanl.gov/posoc