INVESTIGATING SEMANTIC SIMILARITY MEASURES ACROSS THE GENE ONTOLOGY: THE RELATIONSHIP BETWEEN SEQUENCE AND ANNOTATION

P. W. Lord, et al (2003)
Presented by Yuji Mo

Outline

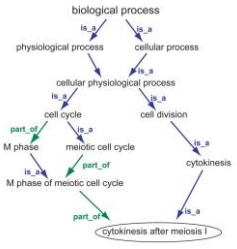
- Introduction
- Related work
- Method and Results
- Application
  - Error checking
  - Search Tool
- Summary
- Future Work

Introduction

- Bioinformatics Data
  - Gene product sequence
    - Enzymes and other proteins encoded in DNA
    - Not human readable
  - Annotation
    - Labels in scientific language
    - Not easy to interpret computationally

Introduction

- Gene Ontology (GO)
  - Three independent ontology
    - Molecular function
    - Biological process
    - Cellular component
  - Shared vocabulary
    - 33059 defined so far
  - Descriptive logic (DL)
    - Transitive relation like is_a, part_of



Introduction

- Number of relations in GO

	Is_a relation	Part_of relation
Biological Process	6207	35
Molecular Function	5697	989
Cellular Component	543	619

Introduction

- Different ratio of *Part_of* and *is_a* relation
- *Part_of* and *is_a* are often exclusive
- They are treat equally in this paper

Related Work

- Various Measure for quantifying the semantic similarity
  - Text based
    - Similar description has similar semantic
  - Path distance between term
    - Assume all semantic link equal weight
  - Information content
    - Used in this paper

Introduction

- SWISS-PROT KB
  - Collaboratively annotated by GO
  - Evidence code
    - Traceable Author Statement (TAS)
    - Inferred from Sequence Similarity (ISS)
    - Not Recorded (NR)

Method

- Semantic similarity between terms
  - Probability of minimum subsume.

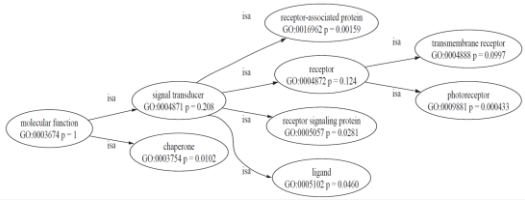
$$p_{ms}(c1, c2) = \min_{c \in S(c1,c2)} \{p(c)\}.$$

$$\mathrm{sim}(c1, c2) = -\ln p_{ms}(c1, c2).$$

  - S(c1,c2) is the set of common parent
  - p(c) the probability that c or its children occur

Method

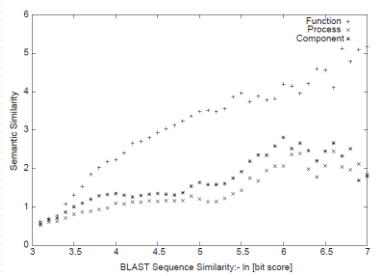
- Example



Method

- Semantic similarity between proteins
  - Average similarity between all pair of terms
- Sequence similarity between proteins
  - Validation Metric
  - Bit score from BLAST results
    - Independent of database size
    - Logarithmic normalized measure

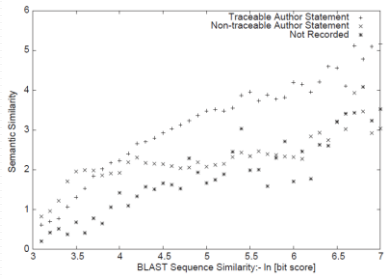
Different aspect of GO



Different aspect of GO

- As sequence similarity increasing, semantic similarity increases in all aspect
  - Similar sequences are likely to be annotated identically
- Molecular Function has higher semantic similarity
  - Homologue proteins must has high Sequence Similarity

Different Evidence Code



Different Evidence Code

- Records with all evidence code are all appeared to be correlated
- TAS shows much greater correlation
- Not all records are equally reliable

Results

Compare sequence and semantic similarity	
Aspect	**Covariance**
Function	0.58
Process	0.28
Component	0.38

Evidence of molecular function	
Evidence code	**Covariance**
TAS	0.58
NAS	0.26
NR	0.49

Application

- Error checking
  - Find protein pairs not obeying the trend above
    - low sequence but high semantic similarity
    - low semantic but high sequence similarity
- Search Tool
  - Find similar protein to a query protein

Error checking

- Proteins with low sequence similarity but high semantic similarity
  - Category specific
    - Groups where different class of protein involved in a same process, like “polymorphic group”
  - Mis-annotation
    - Terms with similar spell
    - Used an obsolete term
    - Errors in the GO structure

Error checking

- Proteins pair with low semantic similarity but high sequence similarity
  - Lack of biological knowledge
  - Lack of a more specific term
  - Mis-annotation

Search Tool

- Similar search tool like BLAST
  - Give a ranked list of semantic related proteins
  - Useful to retrieve related protein by interest aspect
  - Offer an alternative view of the protein at other dimension

Summary

- Strong correlation between semantic and sequence similarity
- Correlation are not distributed in the same way
  - In different aspect or evidence code
- Semantic similarity has useful application

Future work

- Examine effect of different semantic links
- Investigate aspect other than molecular function
- Explore other validation metrics
  - For example, microarray results

Thanks You

Questions and Comments ?