Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation

---

by P. W. Lord, R. D. Stevens, A. Brass and C. A. Goble
Bioinformatics 19(10) 1275–1283
http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/10/1275

presented by Christopher Maier for
INLS 279: Bioinformatics Research Review
2006-02-01

Overall Concept

- Use the addition of ontological annotations to create a new search layer on top of biological databases: semantic querying, to find entries that “mean” the same thing

What is an Ontology?

“A Conceptualization of a Specification”

- Originally a tool from philosophy to convey the existence and relationships of all that exists
- Now used as a formal method to define important concepts and relationships in a particular domain
- More powerful than controlled vocabularies due to added logical infrastructure; more powerful than taxonomies due to additional relationships

The Gene Ontology

- Contains three different “sub-ontologies”: molecular function, cellular component, and biological process
- 20,349 total terms as of December 2005
- Annotations in numerous databases
- http://www.geneontology.org, http://www.godatabase.org/

Defining and Validating Semantic Similarity

Approaches to Ontological Similarity

- Path Distance
- Depth
- These approaches don’t seem to perform well in the biological domain

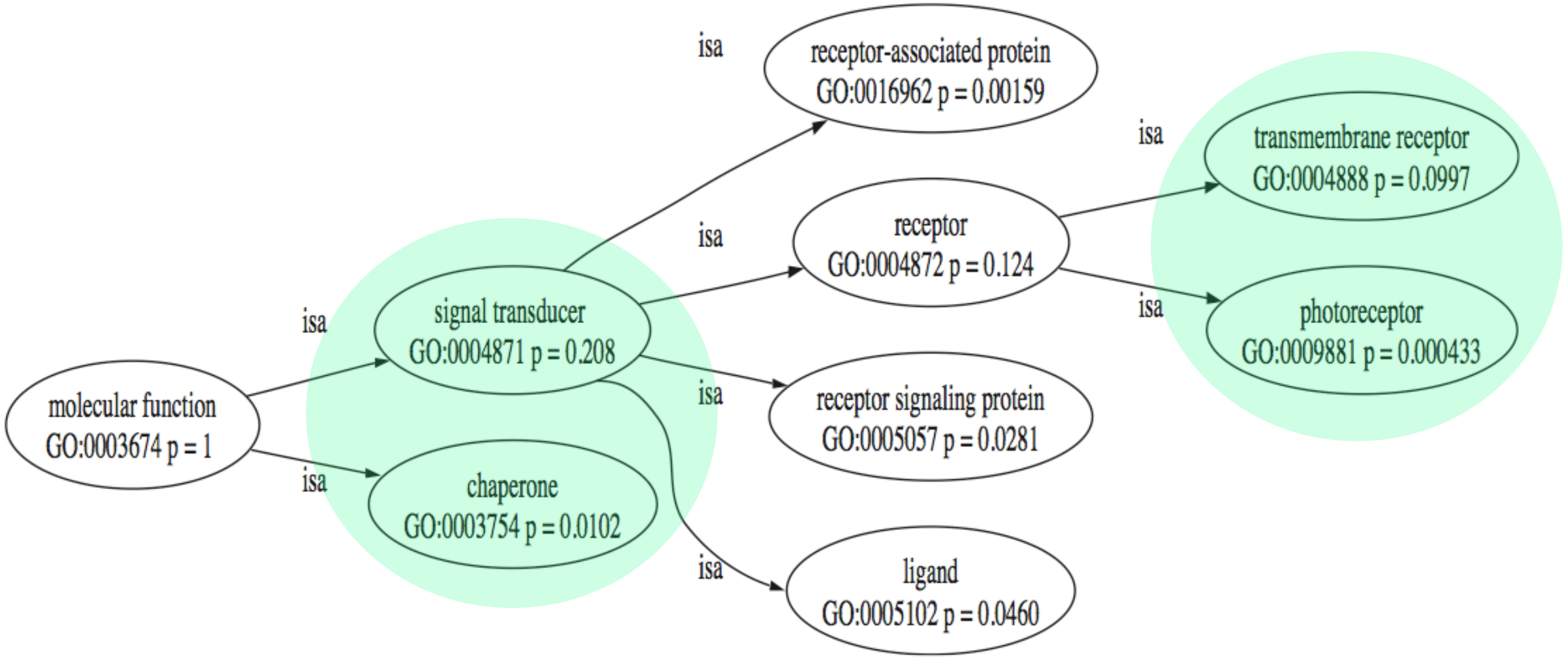

Figure 1 | GO Fragment

Our Definition of Similarity

- Count number of times a term appears (including implicit appearances due to subsumption relationships)
- The less frequent a term, the more informative it is
- Probability of the minimum subsumer for multiple parentage
- Similarity is a negative log function

Validation of Semantic Similarity

- Hard to use traditional validation approaches
- See if sequence similarity tracks with semantic similarity

Why Sequence Similarity?

- Properties of biological macromolecules such as DNA and proteins ultimately derive from their sequence

- Thus, proteins with very similar sequence will generally fold into a very similar 3D shape, allowing them to perform similar functions

- This serves as an empirical measure of similarity, against which our ontological measure can be proven

Adapting to SWISS-PROT

- Orphan Terms
  - “part-of” terms do not participate in “is-a” relationships!
  - Link these back to the ontology root, despite semantic impoverishment
- Link Type Bias
  - Large majority of “molecular function” is “is-a”; over half of “cellular component” is “part-of”
- Multiple Annotations
  - Take average

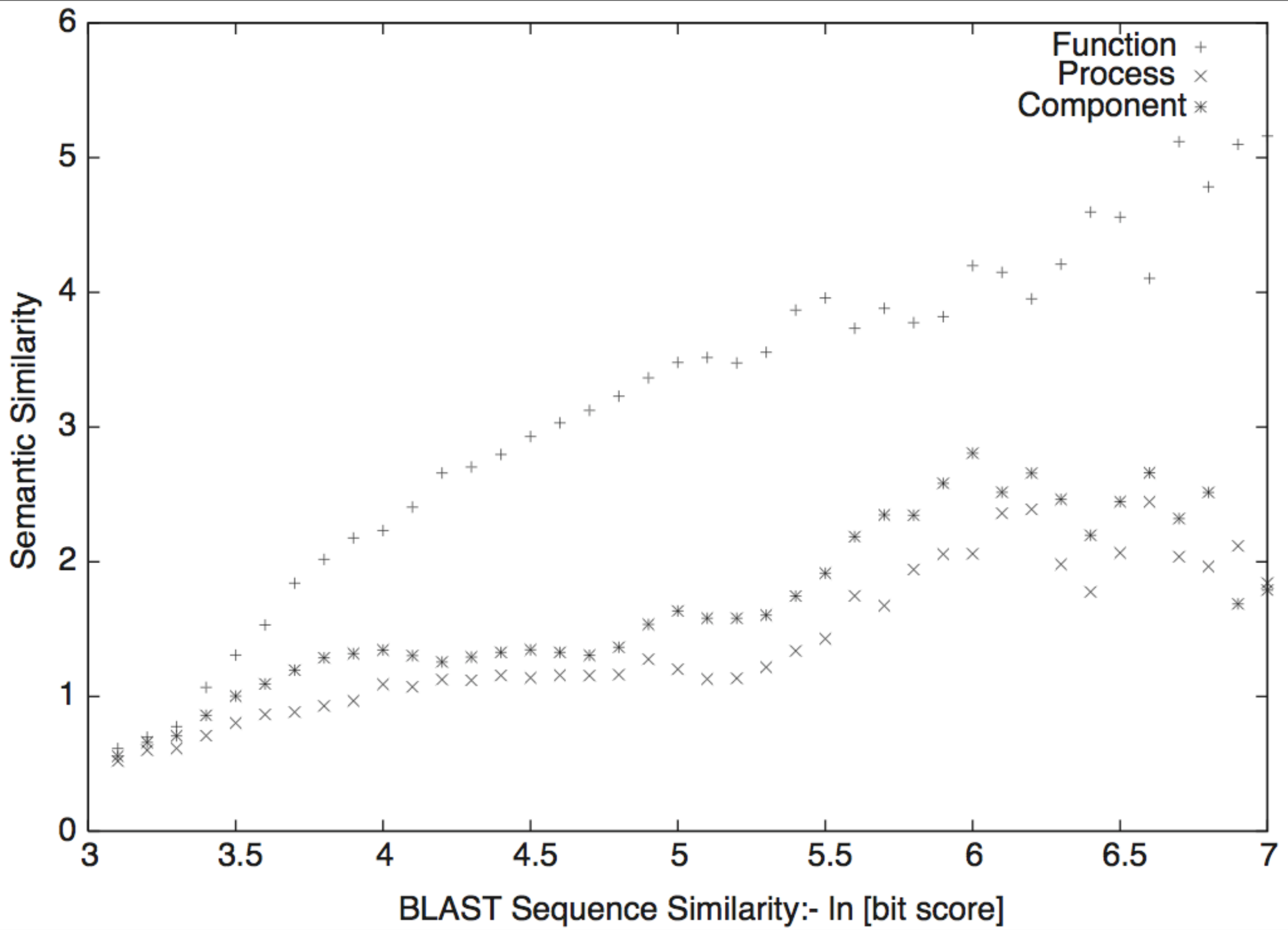

Figure 2

Similarity Correlations in GO

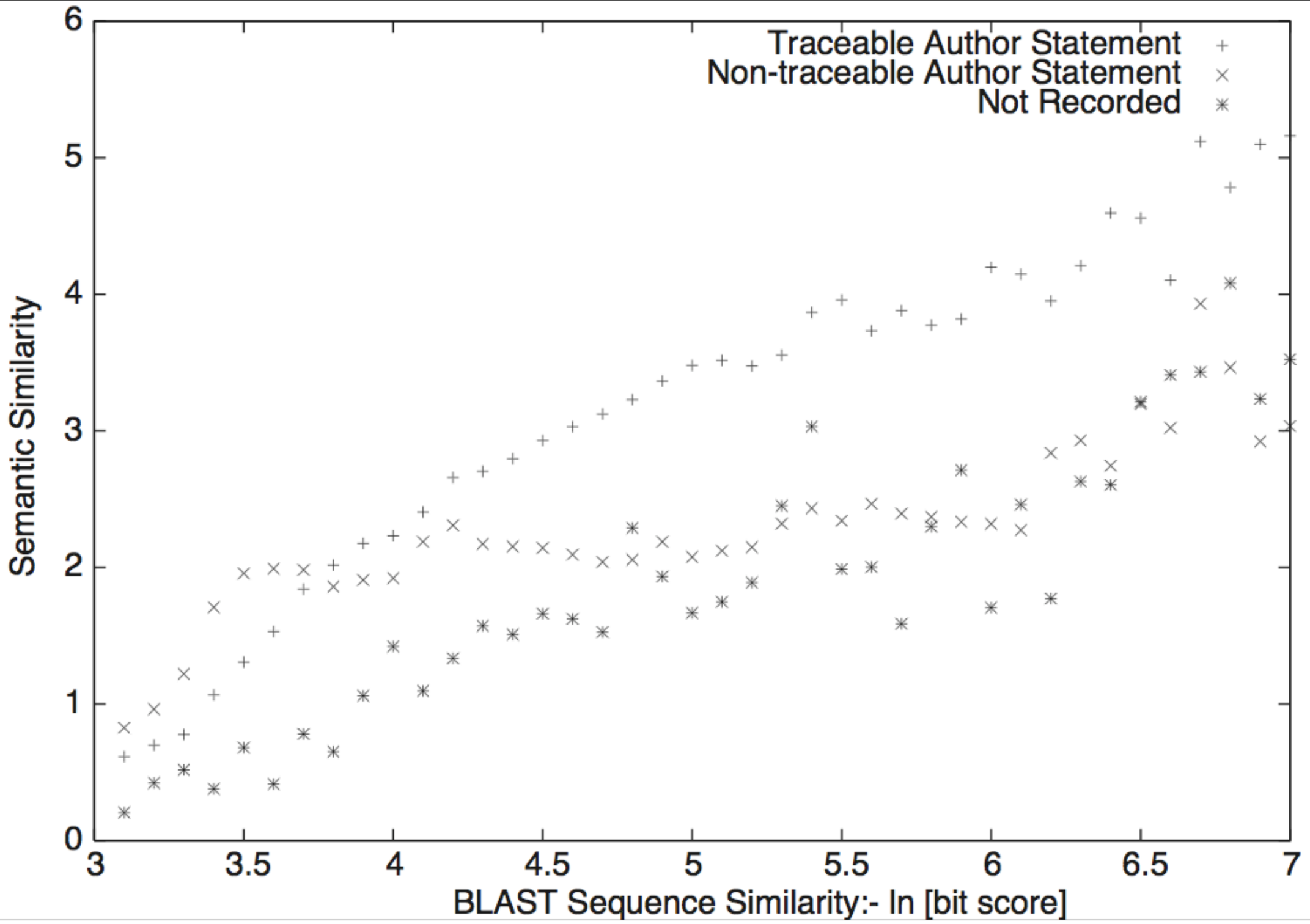

Figure 3

Similarity and Evidence Codes

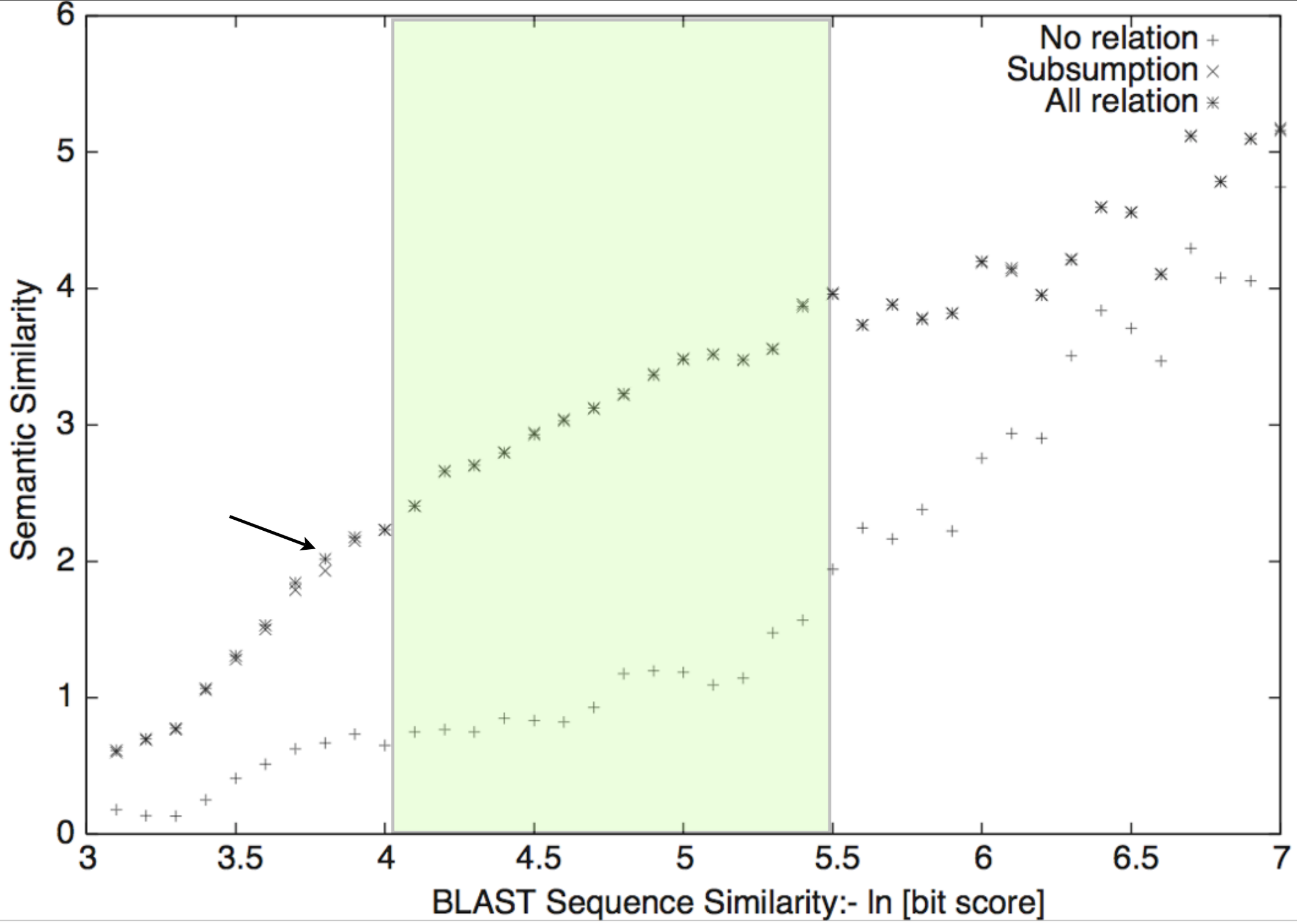

Figure 4

Correlation with links removed

Outliers

- Polymorphic groups: different proteins participate in the same process
- Hyper-variable families
- Mis-annotations
- Under-annotation

Application:
Semantic Search

Search

- Utilize semantic similarity to provide alternative search axes
- Each of the three sub-ontologies of GO retrieves a different kind of “similar” proteins

Swissprot ID	Description	Similarity
(a) Molecular Function		
OPSG_HUMAN	Green-sensitive opsin (Green cone photoreceptor pigment).	8.15
OPN4_HUMAN	Opsin 4 (Melanopsin).	7.23
OPSB_HUMAN	Blue-sensitive opsin (Blue cone photoreceptor pigment).	4.92
5H6_HUMAN	5-hydroxytryptamine 6 receptor (Serotonin receptor)	3.92
A1AA_HUMAN	Alpha-1A adrenergic receptor (Alpha 1A-adrenoceptor)	3.92
A1AB_HUMAN	Alpha-1B adrenergic receptor (Alpha 1B-adrenoceptor).	3.92
(b) Biological Process		
AIPL_HUMAN	Aryl-hydrocarbon interacting protein-like 1.	2.89
CNCG_HUMAN	Retinal cone rhodopsin-sensitive cGMP	2.89
CNRA_HUMAN	Rod cGMP-specific 3′,5′-cyclic phosphodiesterase	2.89
CNRC_HUMAN	Cone cGMP-specific 3′,5′-cyclic phosphodiesterase	2.89
CNRD_HUMAN	Retinal rod rhodopsin-sensitive cGMP	2.89
CRB1_HUMAN	Beta crystallin B1.	2.89
(c) Cellular Component		
1A01_HUMAN	HLA class I histocompatibility antigen	1.86
5H1A_HUMAN	5-hydroxytryptamine 1A receptor (5-HT-1A)	1.86
A1A2_HUMAN	Sodium/potassium-transporting ATPase alpha-2 chain	1.86
A1AA_HUMAN	Alpha- 1A adrenergic receptor	1.86
A33_HUMAN	Cell surface A33 antigen precursor	1.86
ACHA_HUMAN	Acetylcholine receptor protein	1.86

Table 4 | Semantic Search Results

Conclusion

What have we learned?

- Semantic similarity is valid concept
- Ontology structure adds value above controlled vocabulary
- Possible uses: semantic search, error detection

The Future

- As GO grows both in size and in use, the value of semantic searching on GO annotations will increase
- What other similarity functions could be used?
- Are there other measures with which cellular component and biological process similarity are correlated?