Semantic Similarity

Measuring Similarity across the Gene Ontology

Phillip Lord, Robert Stevens, Andy Brass, Carole Goble

`p.lord@russet.org.uk`

Department of Computing Science, University of Manchester

What is GO for?

“The original intent of the group was to construct a set of vocabularies comprising terms that we could share with a common understanding of the meaning of any term used, and that could support cross-database queries.”

What do want to ask?

- What proteins are *semantically similar* to a query protein?
- Or what proteins have *semantically similar* annotation?

Judging Semantic Distance

- Direct matches. Two proteins are semantically similar if they are annotated with the same terms.
- But what of “transmembrane receptor”, (GO:0004888), and “photoreceptor”, (GO:0009881)
- Probability of a direct match depends on the size of GO.

Edge Distance

- The further GO terms are away in the Directed Acyclic Graph (DAG), the less related they are.
- “photoreceptor”, (GO:0009881) and “transmembrane receptor”, (GO:0003754) share a common parent.
- “chaperone”, (GO:0003754) and “signal transducer”, (GO:0004871) also share a common parent.

Edge counting with Weighting

- Each edge can have a weight, perhaps based on depth, to scale the distance calculation.
- “high-affinity tryptophan transporter”, (GO:0005300) is 14 terms deep.
- “anticoagulant”, (GO:0008435) is 3 terms deep.
- Hand annotating GO would be a significant task.

Edge counting with Weighting

- Each edge can have a weight, perhaps based on depth, to scale the distance calculation.
- “high-affinity tryptophan transporter”, (GO:0005300) is 14 terms deep.
- “anticoagulant”, (GO:0008435) is 3 terms deep.
- Hand annotating GO would be a significant task.

Even if we knew how to do it

How is GO Used?

- GO has already been used to annotate many databases. Can we use the information in the corpus?
- Can we define similarity extensionally rather than intentionally?

Information Content

The less frequently a term occurs, the more informative it is.

Information Content

The less frequently a term occurs, the more informative it is.

“Alpha Mating Factor”

Rosetta Inpharmatics: Pubs: Signaling and Circuitry of Multiple MAPK Pathways...

Zymo Research’s new products are for E. coli transformation, bubble-free gel casting,

ALPHA-MATING FACTOR H-TRP-HIS-TRP-LEU-GLN-LEU-LYS-PRO-GLY-GLN-PRO-MET-TYR-OH. Yeast P values

The alpha project @ tMSI: Mating response

Information Content

The less frequently a term occurs, the more informative it is.
“Sex Pheromone”

Primal Instinct Pheromones - Pheromone The secret formula to get girls!

PHEROMONE POWER human sex pheromones PHEROMONE POWER The most powerfull love potion! Human Pheromone the proven ingredient

PHEROMONE ATTRACTION building self confidence PHEROMONE ATTRACTION Primal Instinct pheromones - Incredible

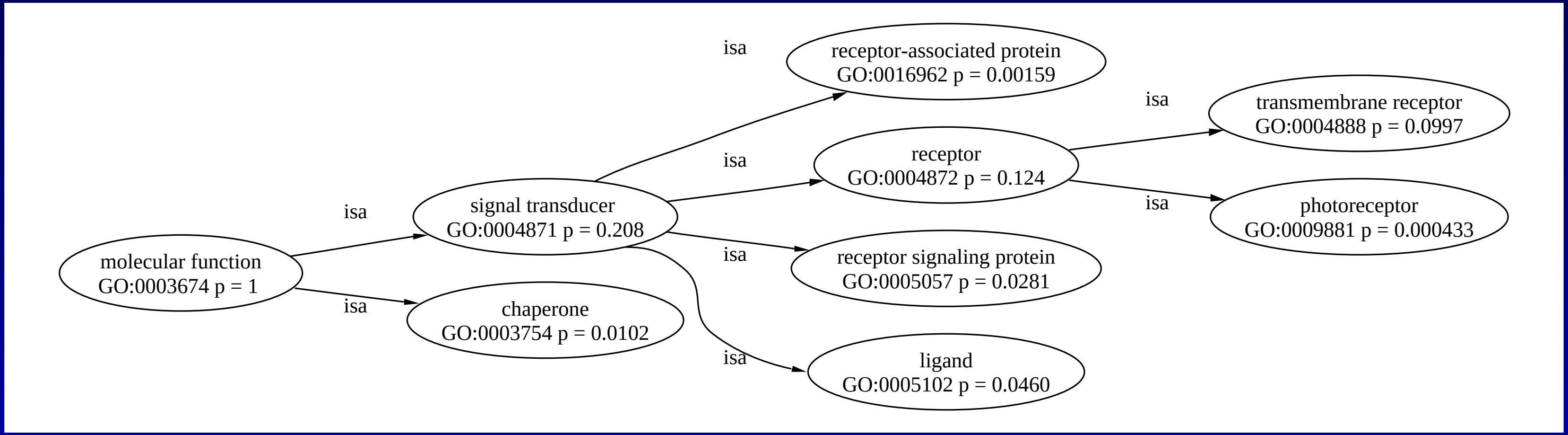
Learn the art of SEDUCTION. All Free Information. sex pheromone – aphrodisiac – pheromone smell !!

Information Content and GO

We define $p(c)$ as the number of times each term, or any of its children occur, divided by the number of times any term occurs.

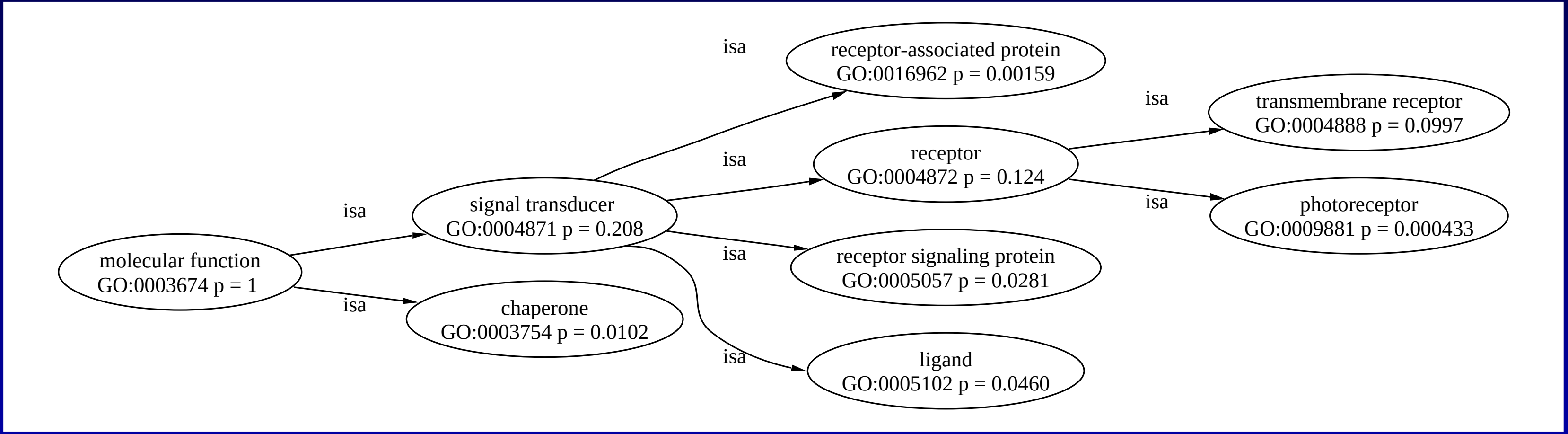
Information Content and GO

We define $p(c)$ as the number of times each term, or any of its children occur, divided by the number of times any term occurs.



Information Content and GO

We define $p(c)$ as the number of times each term, or any of its children occur, divided by the number of times any term occurs.



Because the GO aspects are disconnected sub-graphs, we can calculate this probability for any aspect, or for GO as a whole.

Probabilities to Similarity

We define *probability of the minimum subsumer* p_{ms} as

$$p_{ms}(c1, c2) = \min_{c \in S(c1,c2)} \{p(c)\} \tag{1}$$

where $S(c1, c2)$ is the set of parental concepts shared by the query terms $c1$, $c2$.

Probabilities to Similarity

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

after Resnik, 1995

Probabilities to Similarity

$$\mathrm{sim}(c1,c2) = \frac{2\times[\ln p_{ms}(c1,c2)]}{\ln p(c1)+\ln p(c2)}$$

after Lin, 1998

$$\begin{aligned}\mathrm{dist}(c1,c2) = -2\ln p_{ms}(c1,c2) - \\ (\ln p(c1) + \ln p(c2))\end{aligned}$$

after Jiang and Conrath, 1998

Validation!

- but does it work?

Validation!

- but does it work?
- or rather is it sensible?

Validation!

- but does it work?
- or rather is it sensible?

How can we test this measure?

Validation!

- but does it work?
- or rather is it sensible?

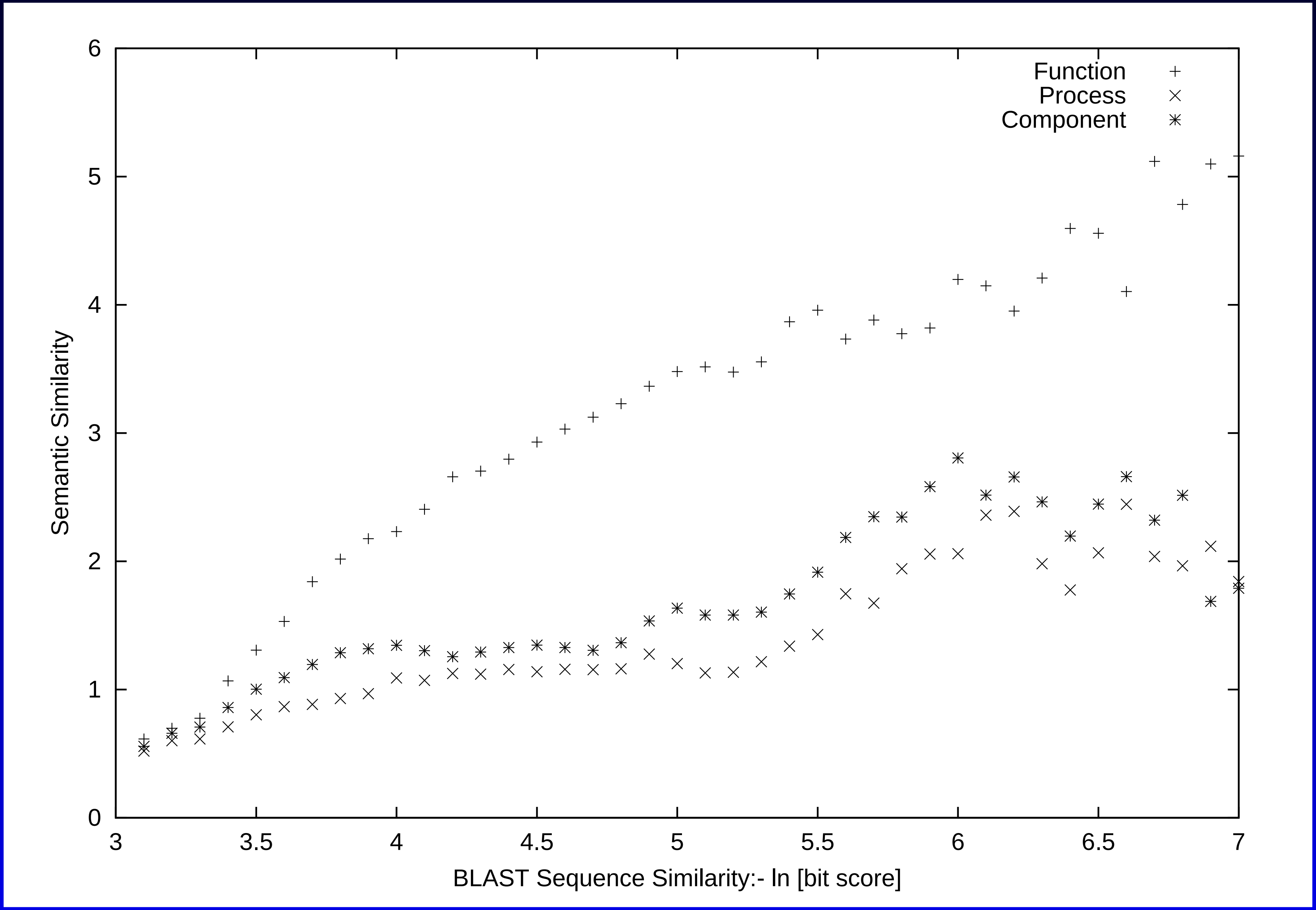
How can we test this measure?

If two sequences are similar, the annotation should also be similar.

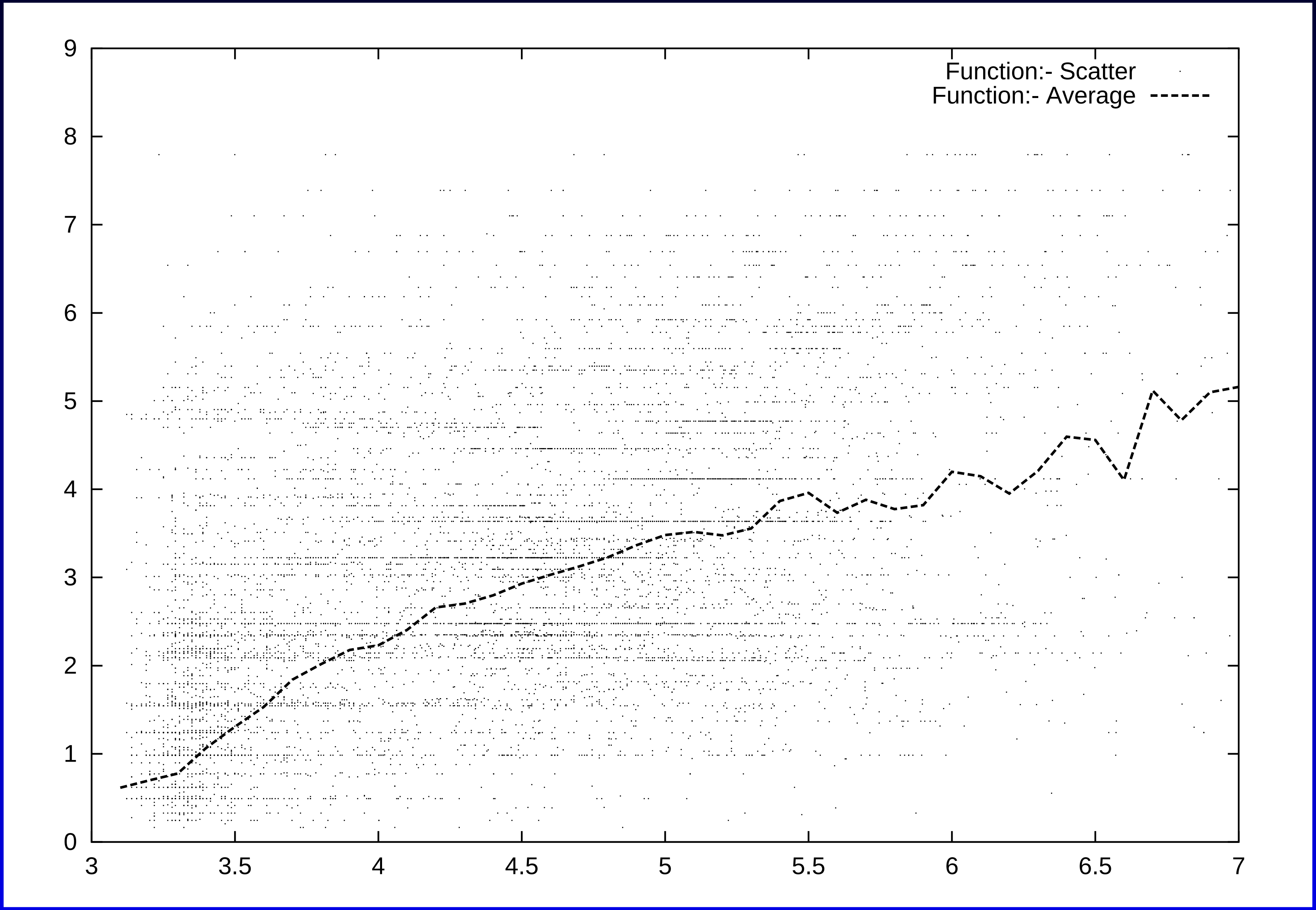
Validation!

- BLAST all SWISS-PROT sequences.
- For each, take all pairs (query and hit).
- Compare semantic similarity, with $\ln[bitscore]$.
- Average semantic similarity for intervals of $ln[bitscore]$

Validation!


Function +
Process ×
Component ✳
Semantic Similarity
0
1
2
3
4
5
6
3
3.5
4
4.5
5
5.5
6
6.5
7
BLAST Sequence Similarity:- ln [bit score]

Scatter


Function:- Scatter
Function:- Average
0
1
2
3
4
5
6
7
8
9
3
3.5
4
4.5
5
5.5
6
6.5
7

Outliers

SPEE_HUMAN (Spermidine synthase (EC 2.5.1.16))

SPSY_HUMAN (Spermine synthase (EC 2.5.1.22))

Both annotated as spermidine synthase.

Searching SWISS-PROT

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

Searching with OPSR_HUMAN

Searching SWISS-PROT

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

Searching with OPSR_HUMAN

Searching SWISS-PROT

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

Searching with OPSR_HUMAN

Conclusions

- Information Content Based measures appear to producing biologically “sensible” results.
- They can be used to check GO annotation.
- They can be used to search GO.

Future Work

- A Web based search tool.
  `http://gosst.cs.man.ac.uk`
- User studies with different measures.
- Differentiating link types.
- Performance optimisation.

Acknowledgements

Robert Stevens, Andy Brass, Carole Goble

David Hoyle, Paul Kirby

Midori Harris, Mike Ashburner, Evelyn Camon

The GO database, and perl API

bioperl

References

[Jiang and Conrath, 1998] Jiang, J. J. and Conrath, D. W. (1998). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, Taiwan. ROCLING X.

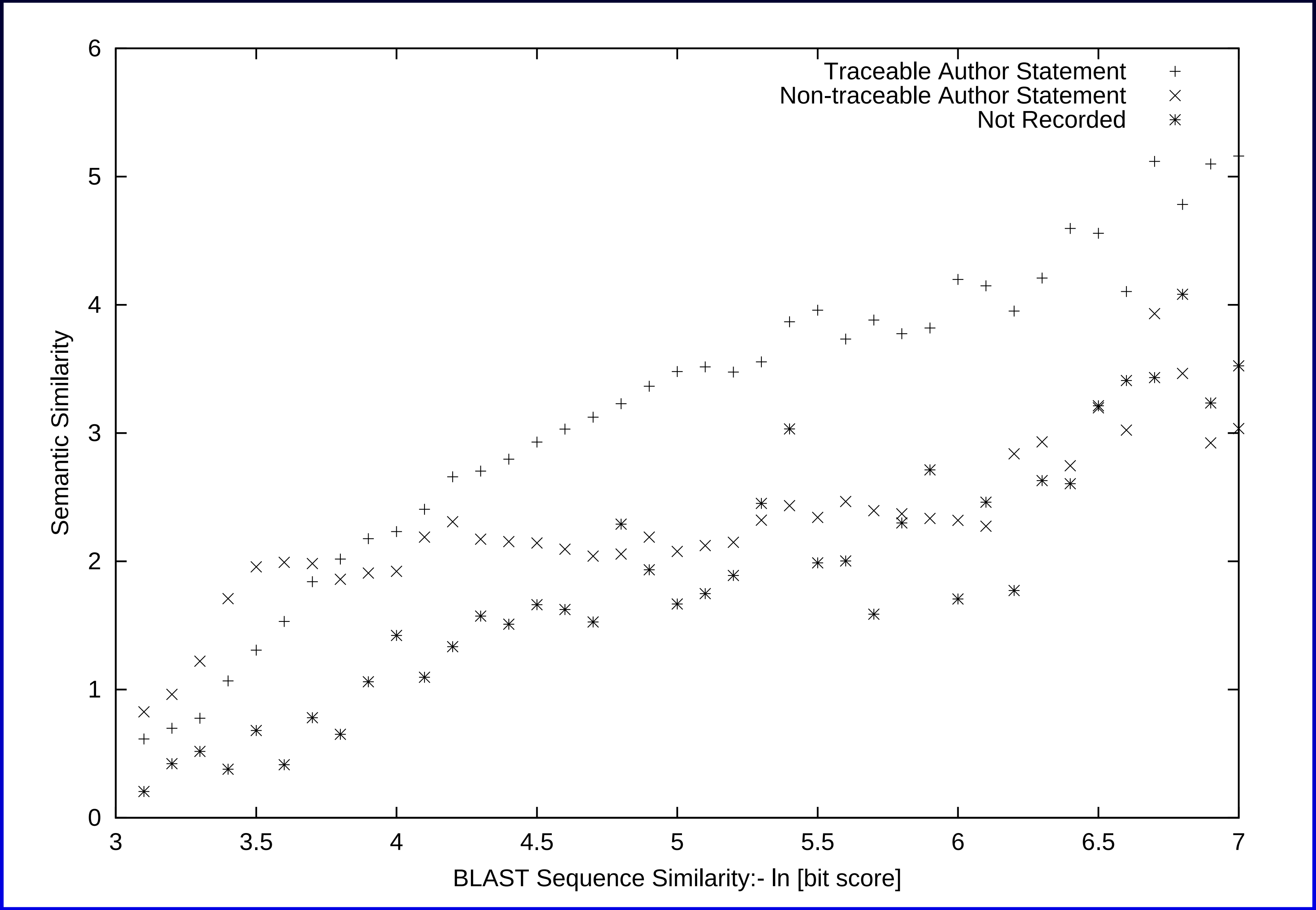
[Lin, 1998] Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

[Resnik, 1995] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453.

Irrelevant Cartoon

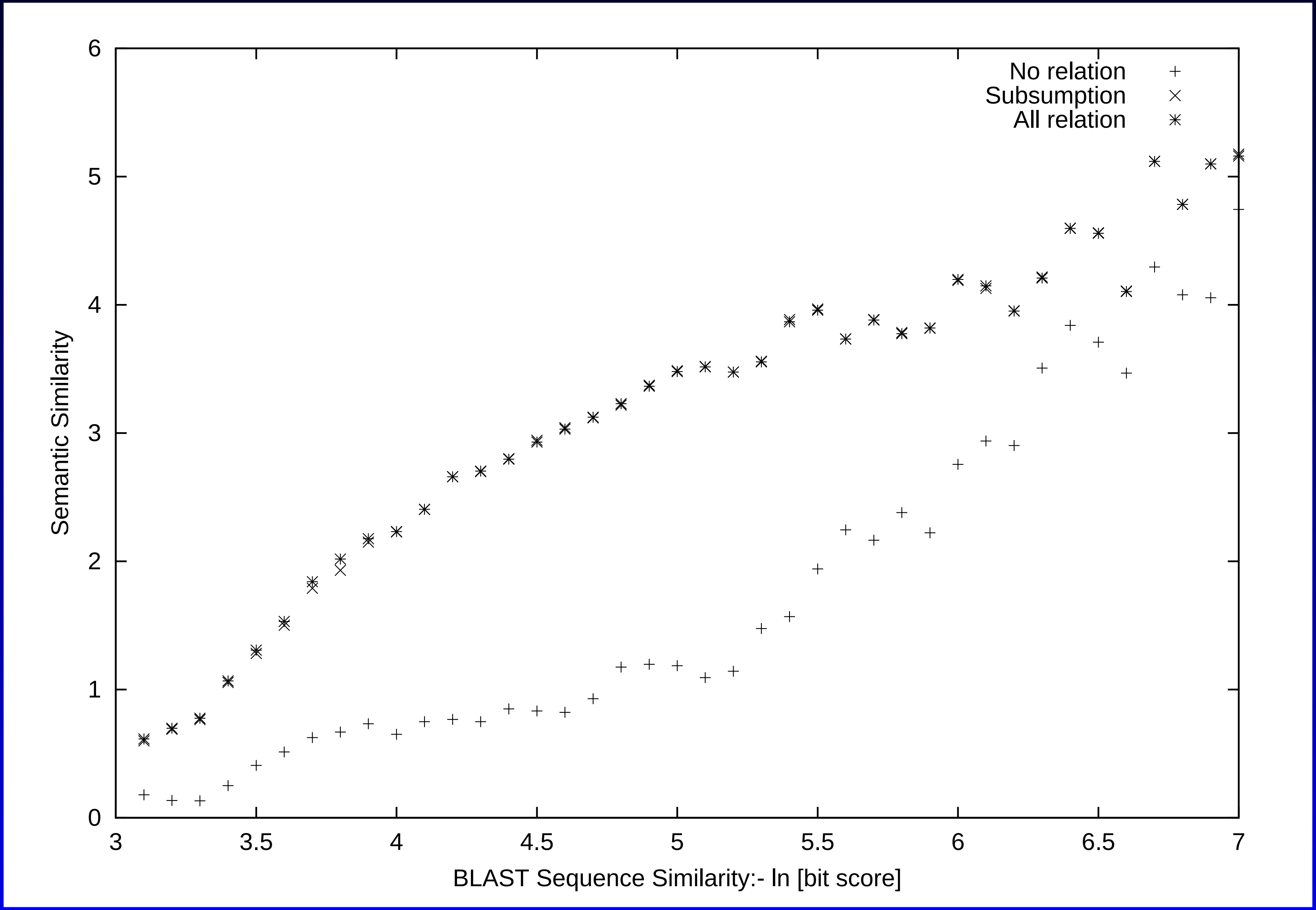
"Of course, the army have always denied that the experiments took place"

Other Data



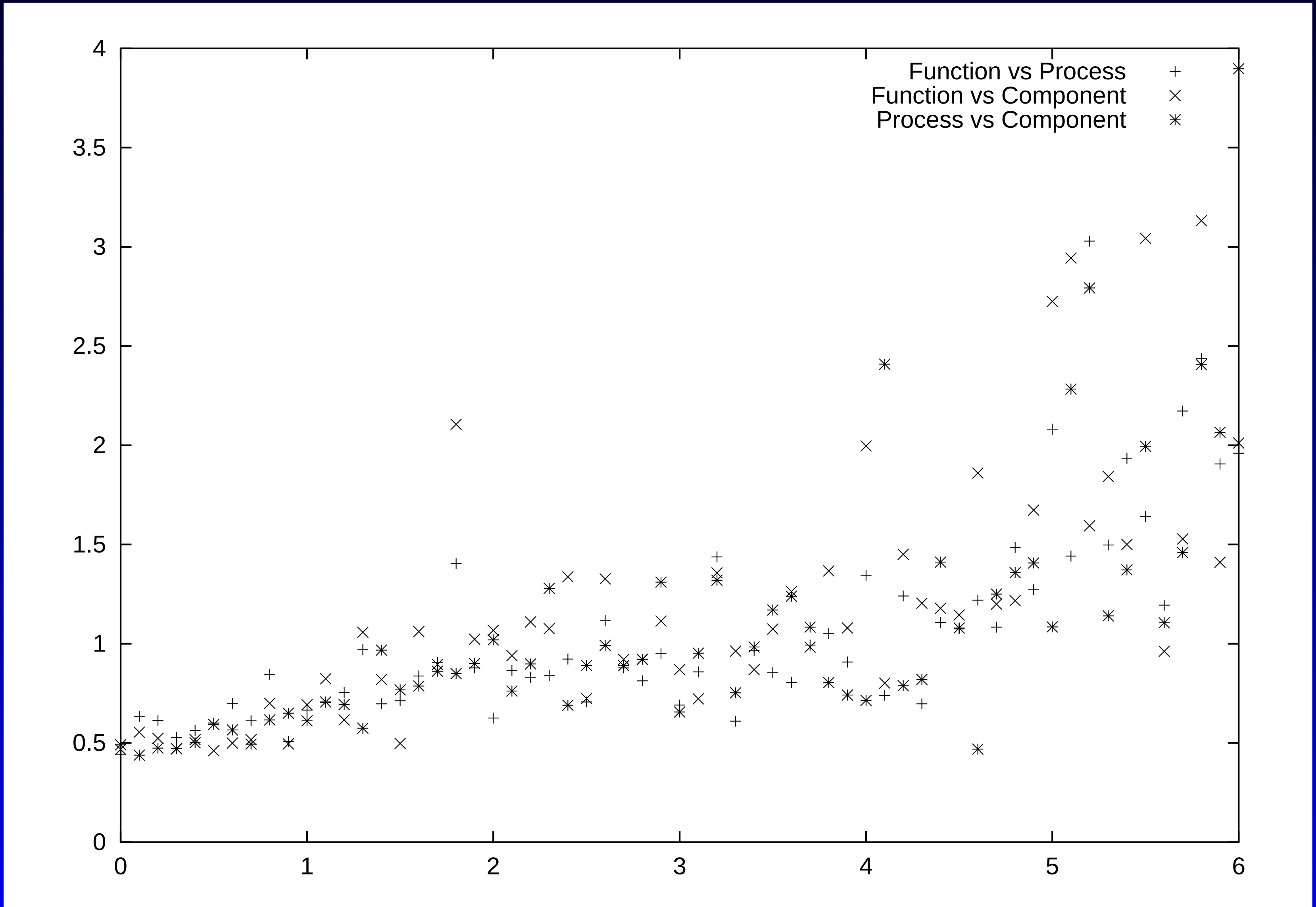
Traceable Author Statement +
Non-traceable Author Statement ×
Not Recorded ✳
Semantic Similarity
0
1
2
3
4
5
6
3
3.5
4
4.5
5
5.5
6
6.5
7
BLAST Sequence Similarity:- ln [bit score]

Other Data



No relation +
Subsumption ×
All relation ✳
Semantic Similarity
0
1
2
3
4
5
6
3
3.5
4
4.5
5
5.5
6
6.5
7
BLAST Sequence Similarity:- ln [bit score]

Other Data



Function vs Process +
Function vs Component ×
Process vs Component ✳
4
3.5
3
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6

Other Data

Aspect	Resnik
Molecular Function	0.577
Biological Process	0.280
Cellular Component	0.368