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."

A SWISS-PROT entry

```
ID   PRIO_HUMAN     STANDARD;      PRT;   253 AA.
AC   P04156;
DT   01-NOV-1986 (Rel. 03, Created)
DT   01-NOV-1986 (Rel. 03, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   Major prion protein precursor (PrP) (PrP27-30) (PrP33-35C) (ASCR).
GN   PRNP.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=86300093; PubMed=3755672;
RA   Kretzschmar H.A., Stowring L.E., Westaway D., Stubblebine W.H.,
RA   Prusiner S.B., Dearmond S.J.;
RT   "Molecular cloning of a human prion protein cDNA.";
RL   DNA 5:315-324(1986).
RN   [2]
RP   SEQUENCE OF 8-253 FROM N.A.
RX   MEDLINE=86261778; PubMed=3014653;
RA   Liao Y.-C.J., Lebo R.V., Clawson G.A., Smuckler E.A.;
RT   "Human prion protein cDNA: molecular cloning, chromosomal mapping,
RT   and biological implications.";
RL   Science 233:364-367(1986).
RN   [3]
RP   SEQUENCE OF 58-85 AND 111-150 (VARIANT AMYLOID GSS).
RX   MEDLINE=91160504; PubMed=1672107;
RA   Tagliavini F., Prelli F., Ghiso J., Bugiani O., Serban D.,
RA   Prusiner S.B., Farlow M.R., Ghetti B., Frangione B.;
RT   "Amyloid protein of Gerstmann-Straussler-Scheinker disease (Indiana
RT   kindred) is an 11 kd fragment of prion protein with an N-terminal
RT   glycine at codon 58.";
RL   EMBO J. 10:513-519(1991).
RN   [4]
RP   STRUCTURE BY NMR OF 118-221.
RX   MEDLINE=20359708; PubMed=10900000;
RA   Calzolai L., Lysek D.A., Guntert P., von Schroetter C., Riek R.,
RA   Zahn R., Wuethrich K.;
RT   "NMR structures of three single-residue variants of the human prion
RT   protein.";
RL   Proc. Natl. Acad. Sci. U.S.A. 97:8340-8345(2000).
CC   -!- FUNCTION: THE FUNCTION OF PRP IS NOT KNOWN. PRP IS ENCODED IN THE
CC       HOST GENOME AND IS EXPRESSED BOTH IN NORMAL AND INFECTED CELLS.
CC   -!- SUBUNIT: PRP HAS A TENDENCY TO AGGREGATE YIELDING POLYMERS CALLED
CC       "RODS".
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR.
CC   -!- POLYMORPHISM: THE FIVE TANDEM OCTAPEPTIDE REPEATS REGION IS HIGHLY
CC       UNSTABLE. INSERTIONS OR DELETIONS OF OCTAPEPTIDE REPEAT UNITS ARE
CC       ASSOCIATED TO PRION DISEASE.
CC   -!- DISEASE: PRP IS FOUND IN HIGH QUANTITY IN THE BRAIN OF HUMANS AND
CC       ANIMALS INFECTED WITH NEURODEGENERATIVE DISEASES KNOWN AS
CC       TRANSMISSIBLE SPONGIFORM ENCEPHALOPATHIES OR PRION DISEASES, LIKE:
CC       CREUTZFELDT-JAKOB DISEASE (CJD), GERSTMANN-STRAUSSLER SYNDROME
CC       (GSS), FATAL FAMILIAL INSOMNIA (FFI) AND KURU IN HUMANS; SCRAPIE
CC       IN SHEEP AND GOAT; BOVINE SPONGIFORM ENCEPHALOPATHY (BSE) IN
CC       CATTLE; TRANSMISSIBLE MINK ENCEPHALOPATHY (TME); CHRONIC WASTING
CC       DISEASE (CWD) OF MULE DEER AND ELK; FELINE SPONGIFORM
CC       ENCEPHALOPATHY (FSE) IN CATS AND EXOTIC UNGULATE ENCEPHALOPATHY
CC       (EUE) IN NYALA AND GREATER KUDU. THE PRION DISEASES ILLUSTRATE
CC       THREE MANIFESTATIONS OF CNS DEGENERATION: (1) INFECTIOUS (2)
CC       SPORADIC AND (3) DOMINANTLY INHERITED FORMS. TME, CWD, BSE, FSE,
CC       EUE ARE ALL THOUGHT TO OCCUR AFTER CONSUMPTION OF PRION-INFECTED
CC       FOODSTUFFS.
DR   EMBL; M13667; AAA19664.1; -.
DR   EMBL; M13899; AAA60182.1; -.
DR   EMBL; D00015; BAA00011.1; -.
DR   PIR; A05017; A05017.
DR   PIR; A24173; A24173.
DR   PIR; S14078; S14078.
DR   PDB; 1E1G; 20-JUL-00.
DR   PDB; 1E1J; 20-JUL-00.
DR   PDB; 1E1P; 20-JUL-00.
DR   PDB; 1E1S; 21-JUL-00.
DR   PDB; 1E1U; 20-JUL-00.
DR   PDB; 1E1W; 20-JUL-00.
DR   MIM; 176640; -.
DR   MIM; 123400; -.
DR   MIM; 137440; -.
DR   MIM; 245300; -.
DR   MIM; 600072; -.
DR   MIM; 604920; -.
DR   InterPro; IPR000817; Prion.
DR   Pfam; PF00377; prion; 1.
DR   PRINTS; PR00341; PRION.
DR   SMART; SM00157; PRP; 1.
DR   PROSITE; PS00291; PRION_1; 1.
DR   PROSITE; PS00706; PRION_2; 1.
KW   Prion; Brain; Glycoprotein; GPI-anchor; Repeat; Signal;
KW   3D-structure; Polymorphism; Disease mutation.
FT   SIGNAL        1     22
FT   CHAIN        23    230       MAJOR PRION PROTEIN.
FT   PROPEP      231    253       REMOVED IN MATURE FORM (BY SIMILARITY).
FT   LIPID       230    230       GPI-ANCHOR (BY SIMILARITY).
FT   CARBOHYD    181    181       N-LINKED (GLCNAC...) (PROBABLE).
FT   DISULFID    179    214       BY SIMILARITY.
FT   DOMAIN       51     91       5 X 8 AA TANDEM REPEATS OF P-H-G-G-G-W-G-
FT                                Q.
FT   REPEAT       51     59       1.
FT   REPEAT       60     67       2.
FT   REPEAT       68     75       3.
FT   REPEAT       76     83       4.
FT   REPEAT       84     91       5.
FT   VARIANT     102    102       P -> L (IN GSS AND EOAD).
FT                                /FTId=VAR_006464.
FT   VARIANT     105    105       P -> L (IN GSS).
FT                                /FTId=VAR_006465.
FT   VARIANT     117    117       A -> V (LINKED TO DEVELOPMENT OF
FT                                DEMENTING GSS).
FT                                /FTId=VAR_006466.
FT   VARIANT     129    129       M -> V (DETERMINES THE DISEASE PHENOTYPE
FT                                IN PATIENTS WHO HAVE A PRP MUTATION AT
FT                                CODON 178: PATIENTS WITH MET DEVELOP FFI,
FT                                THOSE WITH VAL DEVELOP CJD).
FT                                /FTId=VAR_006467.
FT   VARIANT     171    171       N -> S (IN SCHIZOAFFECTIVE DISORDER).
FT                                /FTId=VAR_006468.
FT   VARIANT     178    178       D -> N (IN FFI AND CJD).
FT                                /FTId=VAR_006469.
FT   VARIANT     180    180       V -> I (IN CJD).
FT                                /FTId=VAR_006470.
FT   VARIANT     183    183       T -> A (IN FAMILIAL SPONGIFORM
FT                                ENCEPHALOPATHY).
FT                                /FTId=VAR_006471.
FT   VARIANT     187    187       H -> R (IN GSS).
FT                                /FTId=VAR_008746.
FT   VARIANT     188    188       T -> K (IN EOAD; DEMENTIA ASSOCIATED TO
FT                                PRION DISEASES).
FT                                /FTId=VAR_008748.
FT   VARIANT     188    188       T -> R.
FT                                /FTId=VAR_008747.
FT   VARIANT     196    196       E -> K (IN CJD).
FT                                /FTId=VAR_008749.
FT                                /FTId=VAR_006472.
SQ   SEQUENCE   253 AA;  27661 MW;  43DB596BAAA66484 CRC64;

     MANLGCWMLV LFVATWSDLG LCKKRPKPGG WNTGGSRYPG QGSPGGNRYP PQGGGGWGQP
     HGGGWGQPHG GGWGQPHGGG WGQPHGGGWG QGGGTHSQWN KPSKPKTNMK HMAGAAAAGA
     VVGGLGGYML GSAMSRPIIH FGSDYEDRYY RENMHRYPNQ VYYRPMDEYS NQNNFVHDCV
     NITIKQHTVT TTTKGENFTE TDVKMMERVV EQMCITQYER ESQAYYQRGS SMVLFSSPPV
     ILLISFLIFL IVG

//
```

A SWISS-PROT entry

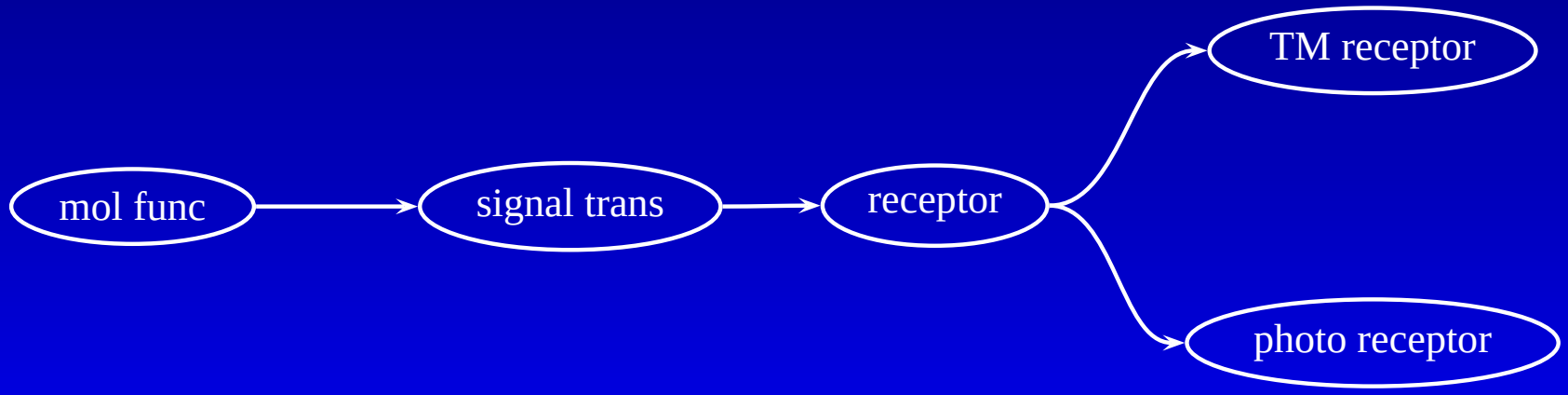
```
ID   PRIO_HUMAN     STANDARD;      PRT;   253 AA.
AC   P04156;
DT   01-NOV-1986 (Rel. 03, Created)
DT   01-NOV-1986 (Rel. 03, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   Major prion protein precursor (PrP) (PrP27-30) (PrP33-35C) (ASCR).
GN   PRNP.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=86300093; PubMed=3755672;
RA   Kretzschmar H.A., Stowring L.E., Westaway D., Stubblebine W.H.,
RA   Prusiner S.B., Dearmond S.J.;
RT   "Molecular cloning of a human prion protein cDNA.";
RL   DNA 5:315-324(1986).
RN   [2]
RP   SEQUENCE OF 8-253 FROM N.A.
RX   MEDLINE=86261778; PubMed=3014653;
RA   Liao Y.-C.J., Lebo R.V., Clawson G.A., Smuckler E.A.;
RT   "Human prion protein cDNA: molecular cloning, chromosomal mapping,
RT   and biological implications.";
RL   Science 233:364-367(1986).
RN   [3]
RP   SEQUENCE OF 58-85 AND 111-150 (VARIANT AMYLOID GSS).
RX   MEDLINE=91160504; PubMed=1672107;
RA   Tagliavini F., Prelli F., Ghiso J., Bugiani O., Serban D.,
RA   Prusiner S.B., Farlow M.R., Ghetti B., Frangione B.;
RT   "Amyloid protein of Gerstmann-Straussler-Scheinker disease (Indiana
RT   kindred) is an 11 kd fragment of prion protein with an N-terminal
RT   glycine at codon 58.";
RL   EMBO J. 10:513-519(1991).
RN   [4]
RP   STRUCTURE BY NMR OF 118-221.
RX   MEDLINE=20359708; PubMed=10900000;
RA   Calzolai L., Lysek D.A., Guntert P., von Schroetter C., Riek R.,
RA   Zahn R., Wuethrich K.;
RT   "NMR structures of three single-residue variants of the human prion
RT   protein.";
RL   Proc. Natl. Acad. Sci. U.S.A. 97:8340-8345(2000).
CC   -!- FUNCTION: THE FUNCTION OF PRP IS NOT KNOWN. PRP IS ENCODED IN THE
CC       HOST GENOME AND IS EXPRESSED BOTH IN NORMAL AND INFECTED CELLS.
CC   -!- SUBUNIT: PRP HAS A TENDENCY TO AGGREGATE YIELDING POLYMERS CALLED
CC       "RODS".
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR.
CC   -!- POLYMORPHISM: THE FIVE TANDEM OCTAPEPTIDE REPEATS REGION IS HIGHLY
CC       UNSTABLE. INSERTIONS OR DELETIONS OF OCTAPEPTIDE REPEAT UNITS ARE
CC       ASSOCIATED TO PRION DISEASE.
CC   -!- DISEASE: PRP IS FOUND IN HIGH QUANTITY IN THE BRAIN OF HUMANS AND
CC       ANIMALS INFECTED WITH NEURODEGENERATIVE DISEASES KNOWN AS
CC       TRANSMISSIBLE SPONGIFORM ENCEPHALOPATHIES OR PRION DISEASES, LIKE:
CC       CREUTZFELDT-JAKOB DISEASE (CJD), GERSTMANN-STRAUSSLER SYNDROME
CC       (GSS), FATAL FAMILIAL INSOMNIA (FFI) AND KURU IN HUMANS; SCRAPIE
CC       IN SHEEP AND GOAT; BOVINE SPONGIFORM ENCEPHALOPATHY (BSE) IN
CC       CATTLE; TRANSMISSIBLE MINK ENCEPHALOPATHY (TME); CHRONIC WASTING
CC       DISEASE (CWD) OF MULE DEER AND ELK; FELINE SPONGIFORM
CC       ENCEPHALOPATHY (FSE) IN CATS AND EXOTIC UNGULATE ENCEPHALOPATHY
CC       (EUE) IN NYALA AND GREATER KUDU. THE PRION DISEASES ILLUSTRATE
CC       THREE MANIFESTATIONS OF CNS DEGENERATION: (1) INFECTIOUS (2)
CC       SPORADIC AND (3) DOMINANTLY INHERITED FORMS. TME, CWD, BSE, FSE,
CC       EUE ARE ALL THOUGHT TO OCCUR AFTER CONSUMPTION OF PRION-INFECTED
CC       FOODSTUFFS.
DR   EMBL; M13667; AAA19664.1; -.
DR   EMBL; M13899; AAA60182.1; -.
DR   EMBL; D00015; BAA00011.1; -.
DR   PIR; A05017; A05017.
DR   PIR; A24173; A24173.
DR   PIR; S14078; S14078.
DR   PDB; 1E1G; 20-JUL-00.
DR   PDB; 1E1J; 20-JUL-00.
DR   PDB; 1E1P; 20-JUL-00.
DR   PDB; 1E1S; 21-JUL-00.
DR   PDB; 1E1U; 20-JUL-00.
DR   PDB; 1E1W; 20-JUL-00.
DR   MIM; 176640; -.
DR   MIM; 123400; -.
DR   MIM; 137440; -.
DR   MIM; 245300; -.
DR   MIM; 600072; -.
DR   MIM; 604920; -.
DR   InterPro; IPR000817; Prion.
DR   Pfam; PF00377; prion; 1.
DR   PRINTS; PR00341; PRION.
DR   SMART; SM00157; PRP; 1.
DR   PROSITE; PS00291; PRION_1; 1.
DR   PROSITE; PS00706; PRION_2; 1.
KW   Prion; Brain; Glycoprotein; GPI-anchor; Repeat; Signal;
KW   3D-structure; Polymorphism; Disease mutation.
FT   SIGNAL        1     22
FT   CHAIN        23    230       MAJOR PRION PROTEIN.
FT   PROPEP      231    253       REMOVED IN MATURE FORM (BY SIMILARITY).
FT   LIPID       230    230       GPI-ANCHOR (BY SIMILARITY).
FT   CARBOHYD    181    181       N-LINKED (GLCNAC...) (PROBABLE).
FT   DISULFID    179    214       BY SIMILARITY.
FT   DOMAIN       51     91       5 X 8 AA TANDEM REPEATS OF P-H-G-G-G-W-G-
FT                                Q.
FT   REPEAT       51     59       1.
FT   REPEAT       60     67       2.
FT   REPEAT       68     75       3.
FT   REPEAT       76     83       4.
FT   REPEAT       84     91       5.
FT   VARIANT     102    102       P -> L (IN GSS AND EOAD).
FT                                /FTId=VAR_006464.
FT   VARIANT     105    105       P -> L (IN GSS).
FT                                /FTId=VAR_006465.
FT   VARIANT     117    117       A -> V (LINKED TO DEVELOPMENT OF
FT                                DEMENTING GSS).
FT                                /FTId=VAR_006466.
FT   VARIANT     129    129       M -> V (DETERMINES THE DISEASE PHENOTYPE
FT                                IN PATIENTS WHO HAVE A PRP MUTATION AT
FT                                CODON 178: PATIENTS WITH MET DEVELOP FFI,
FT                                THOSE WITH VAL DEVELOP CJD).
FT                                /FTId=VAR_006467.
FT   VARIANT     171    171       N -> S (IN SCHIZOAFFECTIVE DISORDER).
FT                                /FTId=VAR_006468.
FT   VARIANT     178    178       D -> N (IN FFI AND CJD).
FT                                /FTId=VAR_006469.
FT   VARIANT     180    180       V -> I (IN CJD).
FT                                /FTId=VAR_006470.
FT   VARIANT     183    183       T -> A (IN FAMILIAL SPONGIFORM
FT                                ENCEPHALOPATHY).
FT                                /FTId=VAR_006471.
FT   VARIANT     187    187       H -> R (IN GSS).
FT                                /FTId=VAR_008746.
FT   VARIANT     188    188       T -> K (IN EOAD; DEMENTIA ASSOCIATED TO
FT                                PRION DISEASES).
FT                                /FTId=VAR_008748.
FT   VARIANT     188    188       T -> R.
FT                                /FTId=VAR_008747.
FT   VARIANT     196    196       E -> K (IN CJD).
FT                                /FTId=VAR_008749.
FT                                /FTId=VAR_006472.
SQ   SEQUENCE   253 AA;  27661 MW;  43DB596BAAA66484 CRC64;

     MANLGCWMLV LFVATWSDLG LCKKRPKPGG WNTGGSRYPG QGSPGGNRYP PQGGGGWGQP
     HGGGWGQPHG GGWGQPHGGG WGQPHGGGWG QGGGTHSQWN KPSKPKTNMK HMAGAAAAGA
     VVGGLGGYML GSAMSRPIIH FGSDYEDRYY RENMHRYPNQ VYYRPMDEYS NQNNFVHDCV
     NITIKQHTVT TTTKGENFTE TDVKMMERVV EQMCITQYER ESQAYYQRGS SMVLFSSPPV
     ILLISFLIFL IVG

//
```

What do want to ask?

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

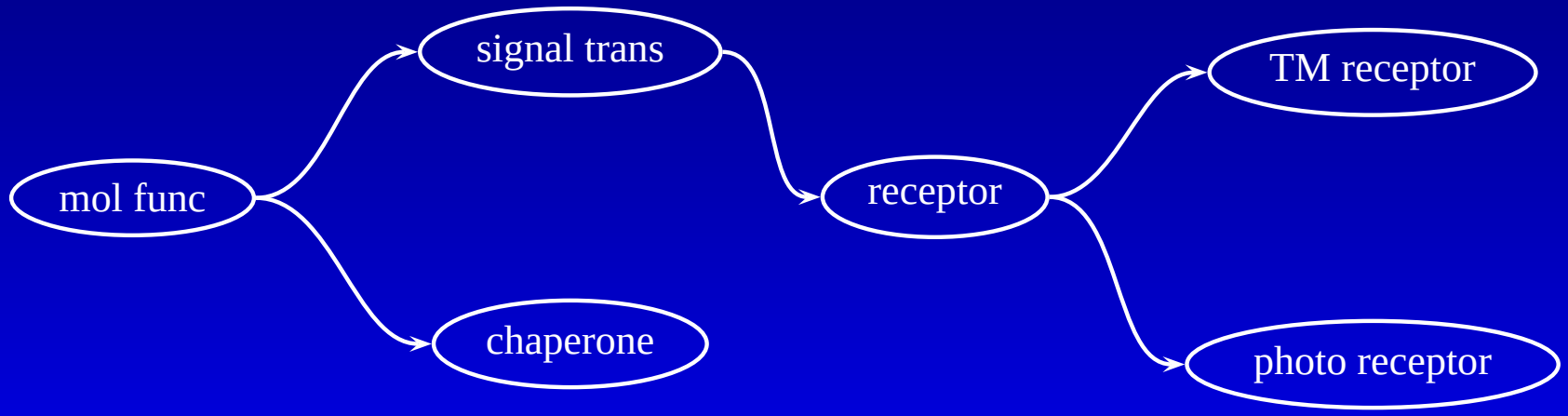
Judging Semantic Distance

- 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) share a common parent.



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 !!

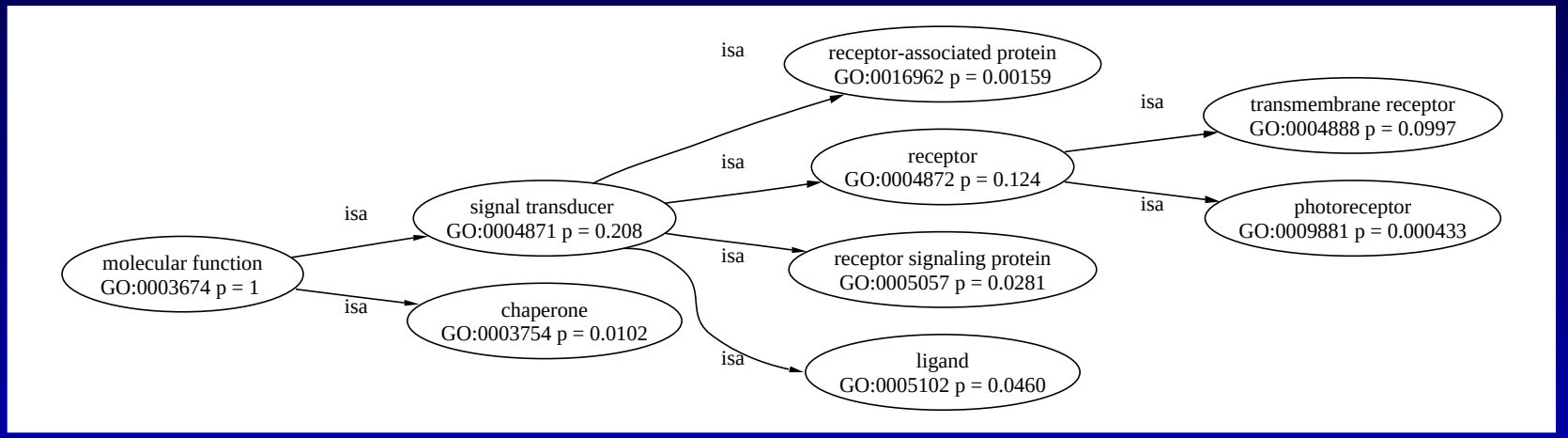
THE UNIVERSITY of MANCHESTER

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.

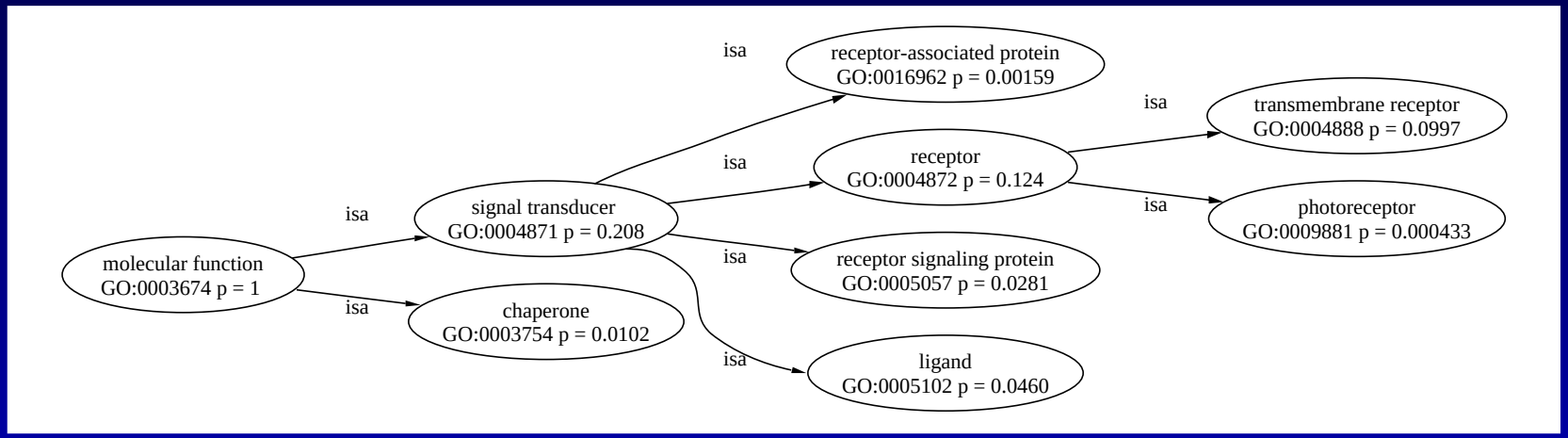
Information Content and GO

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.

THE UNIVERSITY of MANCHESTER

Probabilities to Similarity

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

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

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

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

after Jiang and Conrath, 1998

Measures

We can use these metrics to provide a “semantic similarity search”, similar to a sequence similarity search.

Here we have performed a pairwise comparison between the search protein, and each protein in SWISS-PROT, and ranked the results, using the Resnik measure.

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 phosphodi-esterase	2.89
CNRC_HUMAN	Cone cGMP-specific 3',5'-cyclic phosphodi-esterase	2.89
CNRD_HUMAN	Retinal rod rhodopsin-sensitive cGMP	2.89
CRB1_HUMAN	Beta crystallin B1.	2.89

Searching with OPSR_HUMAN

THE UNIVERSITY of MANCHESTER

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

THE UNIVERSITY of MANCHESTER

Selecting a measure

- but which measure is best?

Selecting a measure

- but which measure is best?

Previously we have used comparison with sequence similarity to test the Resnik measure.

Selecting a measure

- but which measure is best?

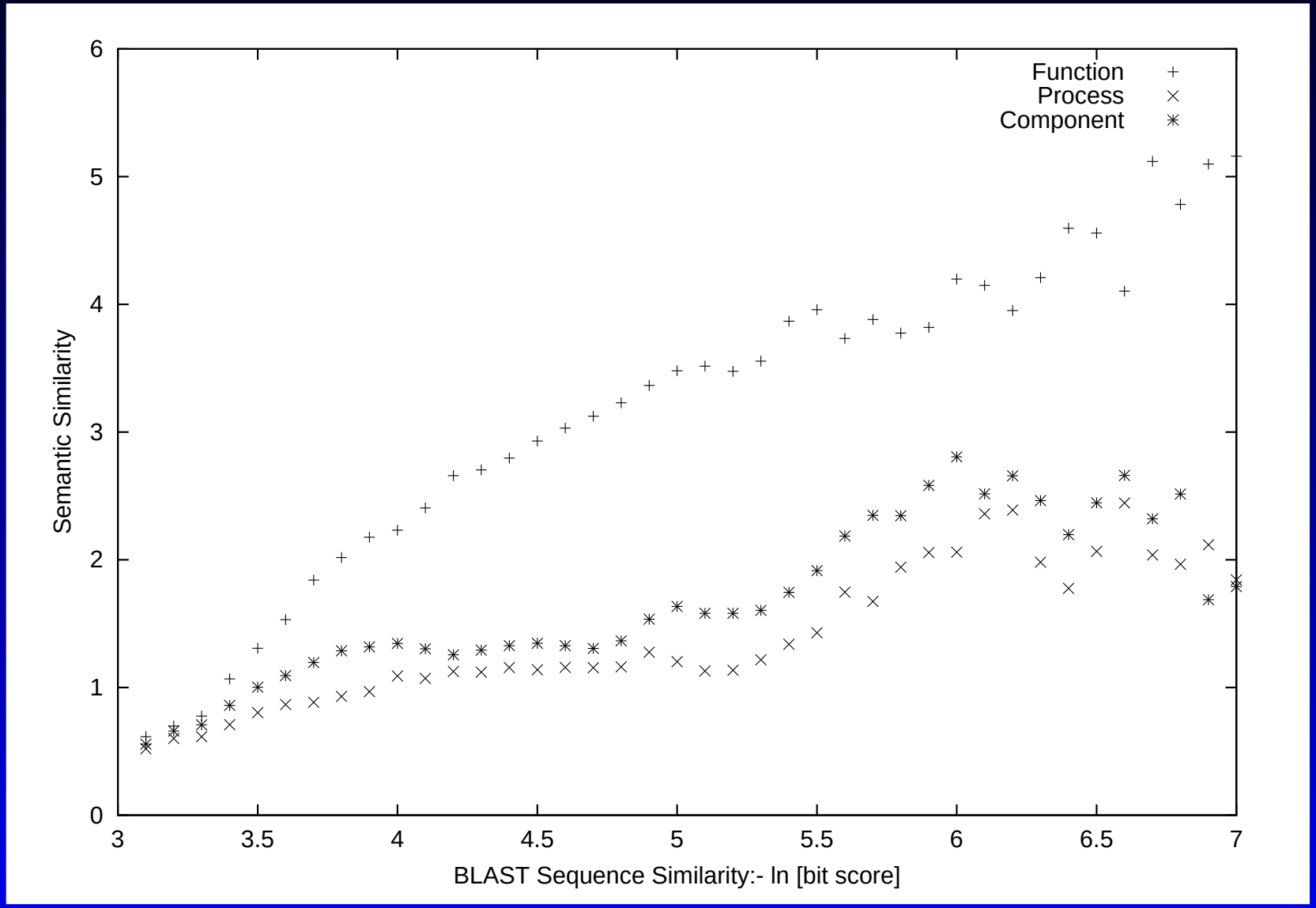
Previously we have used comparison with sequence similarity to test the Resnik measure.

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

Selecting a measure

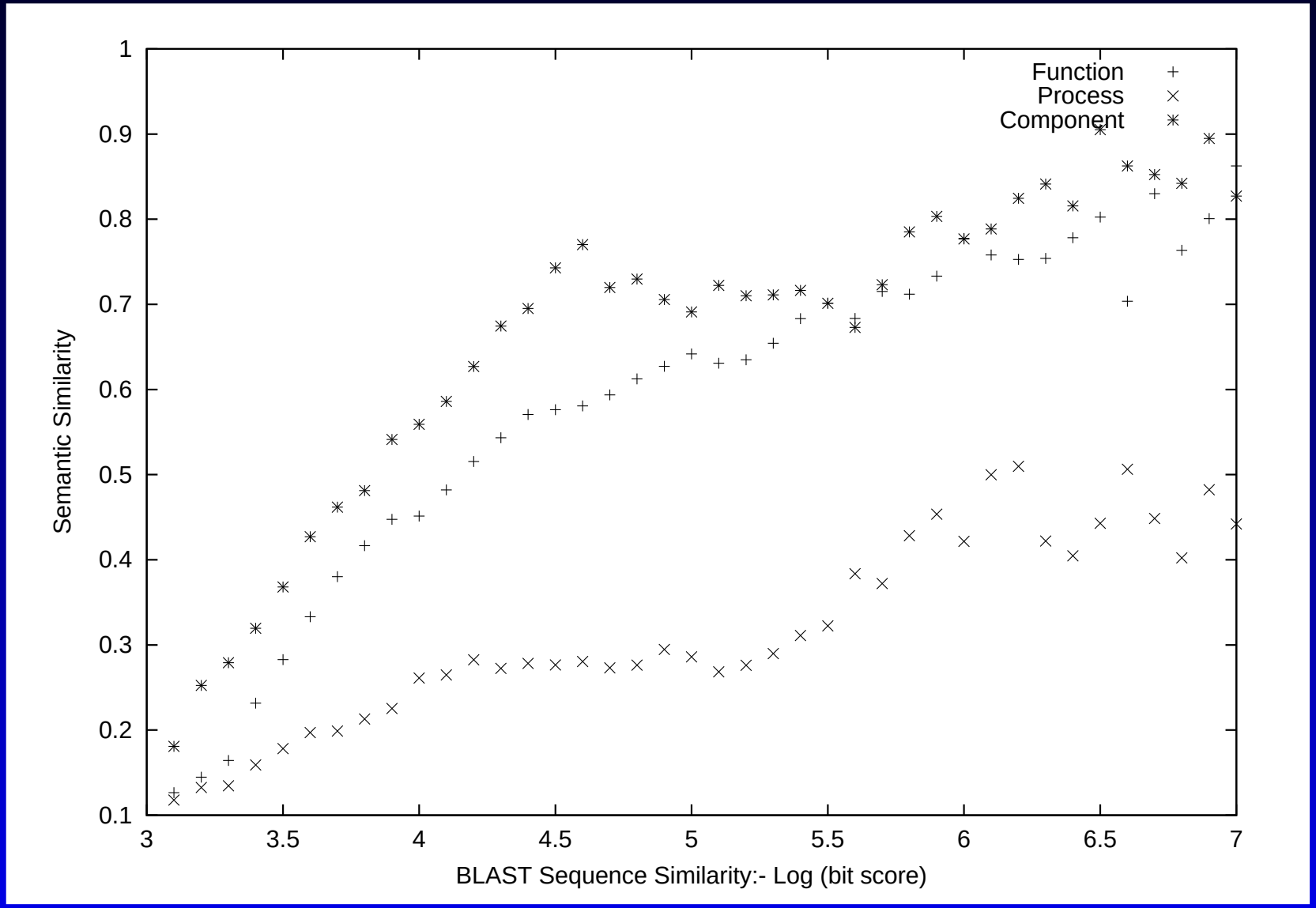
- BLAST all SWISS-PROT sequences.
- For each, take all pairs (query and hit).
- Compare semantic similarity, with $\ln[bitscore]$.
- Average semantic similarity for intervals of $ln[bitscore]$

Selecting a measure



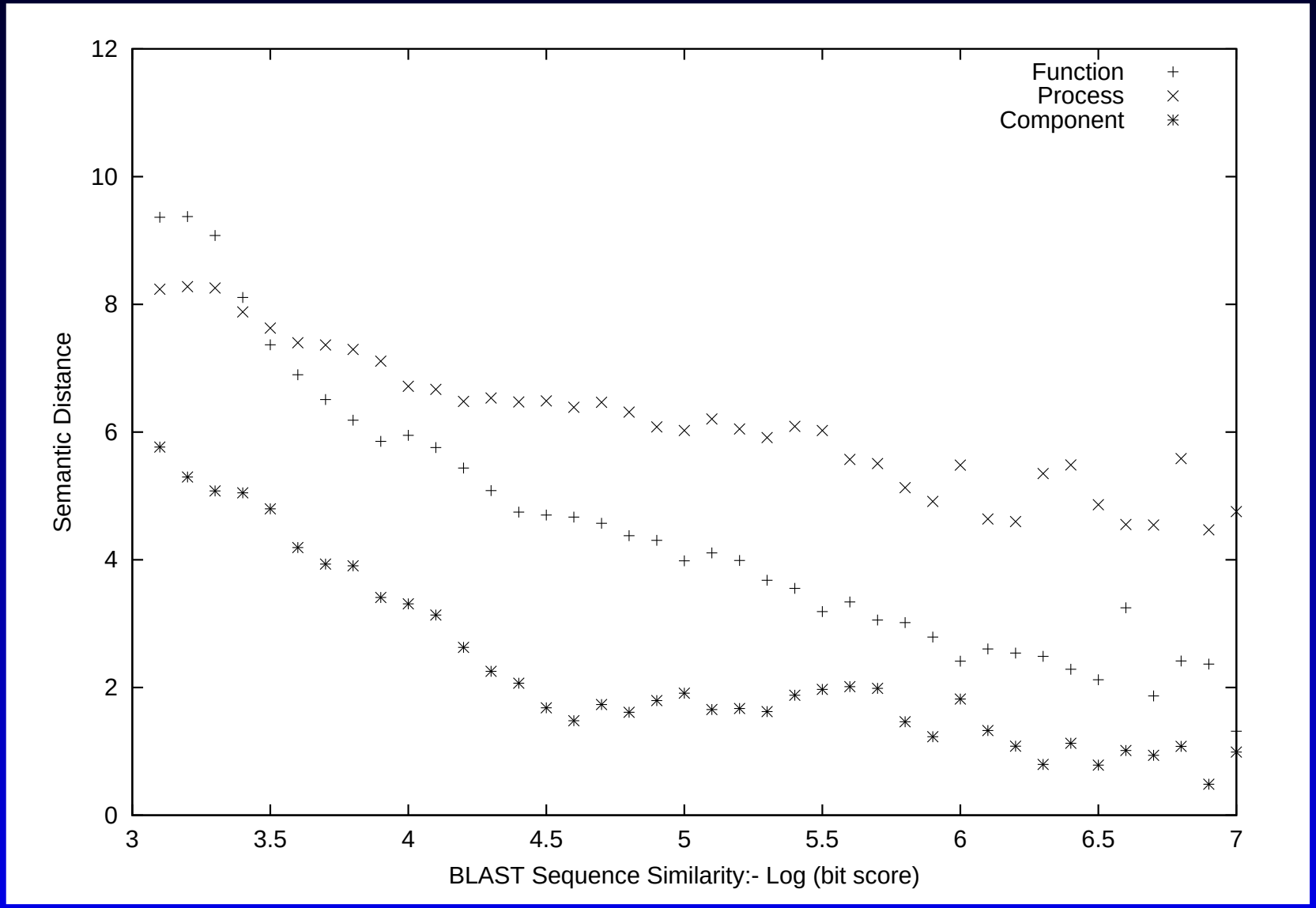
THE UNIVERSITY of MANCHESTER

Selecting a measure



THE UNIVERSITY of MANCHESTER

Selecting a measure

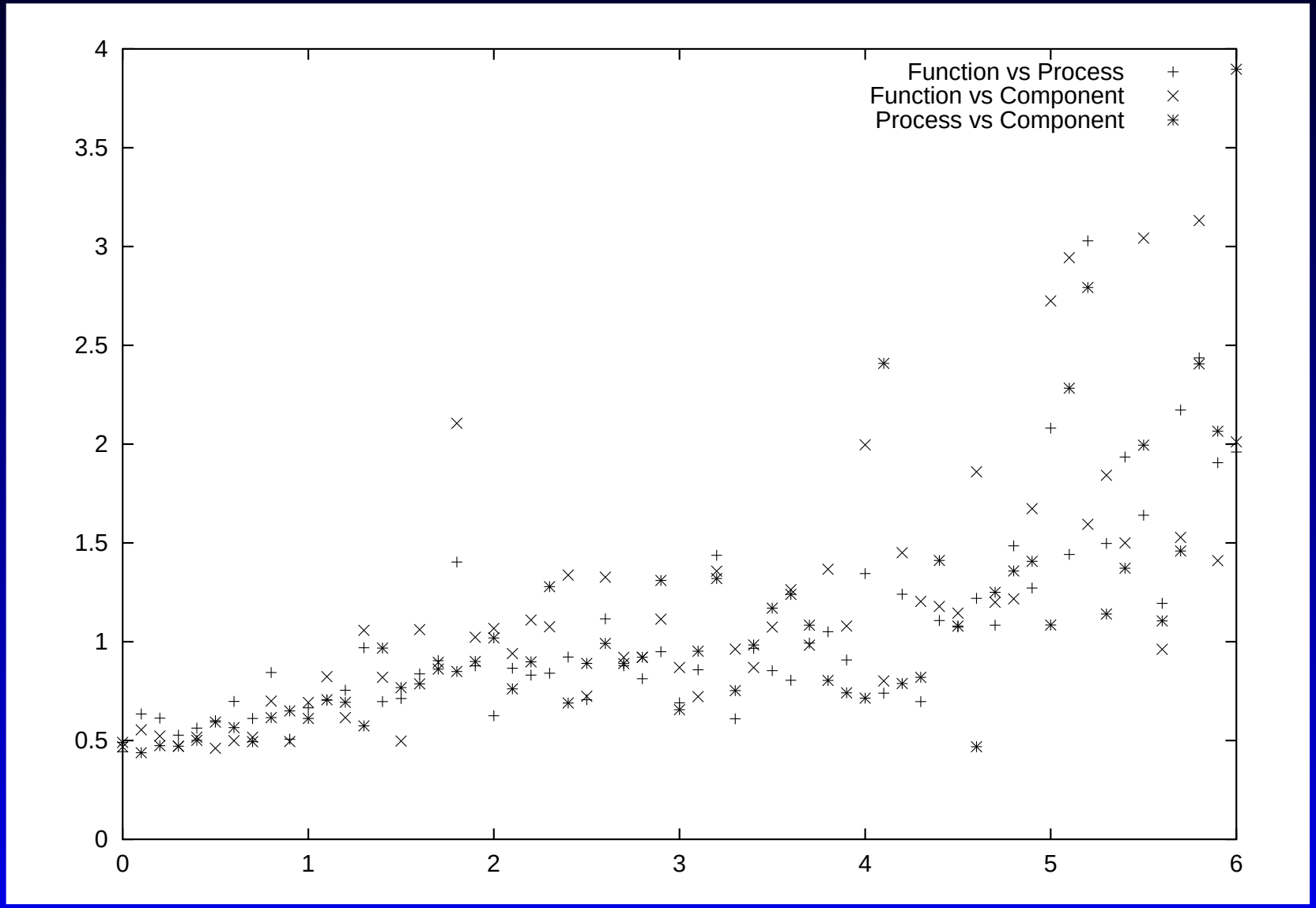


THE UNIVERSITY of MANCHESTER

Are the aspects orthogonal?

The semantic similarity measures can be based on the aspects or GO as a whole. Which is best for a search?

Are the aspects orthogonal?



Are the aspects orthogonal?

In all cases there is a weak but significant correlation between all the semantic similarity of all three aspects. The large degree of independence between the aspects therefore suggests that searching over GO as a whole is inappropriate, unless looking for exact matches.

Conclusions

- Information content based measures can usefully be applied to GO.
- Of the various information based content measures no stand out as having a clear advantage.
- The aspects are largely independent, and searching is probably best done over each individually.

Future Work

- User studies.
- We have a web search engine in place, and a downloadable library. (http://www.gosst.man.ac.uk).
- Applying these measures to outlier analysis, which has previously been used to identify mis-annotations.

Acknowledgements

Robert Stevens, Andy Brass, Carole Goble
David Hoyle, Paul Kirby
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.