Bio-Ontologies 02. Script for “Measuring Similarity across the Gene Ontology”

Phillip Lord
Department of Computing Science,
University of Manchester p.lord@russet.org.uk

January 1, 2003

1 Intro

- **Name rank and serial number.**
- Talk about semantic similarity and GO.

2 What is GO for?

- What is GO intended for.
- Will talk here about using GO for querying within a database.

3 What do we want to ask

What do want from GO?

Read the Slide

What sort of queries do we want to perform. One query familiar to most biologists is “what proteins are similar to this one?”. GO equivalent might be “what proteins have similar annotation to this one?”.

For this we need to have a notion of *semantic similarity* between two terms in an ontology.

4 Judging Semantic Distance

- Direct matches. Simple and Straightforward.
- But two examples shown are clearly semantically similar.
- Probability of match depends on size. The larger the ontology gets the lower the probability. So this measure gets worse as the ontology gets bigger.
- GO curators are showing no sign of getting bored yet.

5 Edge Distance

Read the slide

6 How is GO used?

Read the Slide

7 Information Content

- **Read the Slide**
- Familiar from search engines.
- **Slide transistion**
- Search from search engine for “alpha mating factor”
- For those not familiar with sex life of yeast, alpha mating factor is yeasty equivalent for after shave.
- “Mating factor” also know colloquially as “sex pheromone”.
- **Slide transistion**
- Searching reveals a very different sort of biology.
- “sex” occurs so frequently, it has almost no information content.

8 Information Content and GO

- **read the slide**
- **Slide transistion**
- Part of GO DAG annotated with the number of occurrences in SWISS-PROT.
- 1/5 of uses are “signal transducers”
- Because occurrence depends on term, or any children, probability increases, as we move up the tree, and gets to 1 at the root node
- **Slide transistion**
- **Read the Slide**

9 Probabilities to Similarity

- **Read the Slide**
- To get from this probability to a similarity, simple take -ln.
- Varies from 0 (un-related, or share only the root node as a parent) to infinity.
- **Slide transistion**
- Also have another similarity score, and one distance score. Will not mention further here, but we have been experimenting with these also.

10 Measures

Read the Slide

11 Searching SWISS-PROT

- Original question was about querying databases.
- Can we build a search tool? Yes. Perform exhaustive search of SWISS-PROT for each protein, and rank results
- Shows results for search with "OPSR_HUMAN" against molecular function aspect. All GPCR's
- **Slide transistion**
- Search with biological process. All proteins involved with vision.
- **Slide transistion**
- With Cellular Component. All membrane proteins
- GO does not (or did not!) differentiate between membranes, hence get both internal and cellular membrane proteins.

12 Selecting a measure

Read the SlideSlide transistion

- Does it work?
- **Slide transistion**
- Took all proteins in SWISS-PROT and blasted them. Took the top 100 or so matches (which normally extends from very good matches, to complete rubbish). For each match compared semantic similarity to $\ln[bitscore]$, which is a similarity measure independent of size. Averaged semantic similarity for intervals down x axis.
- **Slide transistion**

- In this case we have calculated similarity for each aspect independently.
- Statistically significant correlation for all three aspects of GO. Correlation is much higher for “function” aspect, and it also has numerically greater value. Which is what we would expect from the biology.
- Also from Lin, and Jiang measures.

13 Orthogonal

Read the Slide
Slide transistion
One example of this data set, we have done the same with Lin and Jiang measure.
Read the Slide

14 Conclusions

Read the Slide

15 Future Work

Read the Slide

16 Acknowledgements

- Work was done by me, with valuable input from Robert, Andy, Carole
- David and Paul for my dodgy stats
- GO and SWISS-PROT people for helping with questions
- The work makes heavy use of GO database, and API, and bioperl, thanks for making freely available.

17 Irrelevant Cartoon

Read the Slide

18 Other data

A random collection of other data.

- Evidence:- TAS gives best correlation
- Relationships:- DAG better than CV
- Different aspects are correlated in usage.
- Same again.