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

August 20, 2002

1 Intro

- **Name rank and serial number.**
- Talk about semantic similarity and GO.
- Already had introduction for GO, so don't have to do this.

2 Protein Databases?

- Most of biological databases are sequence based.
- Or are they?
- Slide shows SWISS-PROT entry.
- **Slide Transition**
- Only a small amount of SWISS-PROT is actually sequence.
- Searching annotation may be more important that searching sequence.

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 Edge Counting with Weighting

- based on depth?
- GO's to big to hand annotate
- **Slide Transition**
- Anyway on what basis would we hand annotate?
- Not easy to put numerical value on edges.

7 How is GO used?

Read the Slide

8 Information Content

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

9 Information Content and GO

- **read the slide**
- **Slide Transition**
- 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
- This is only true if all the nodes have a parent.
- In GO not all terms have explicit “is-a” parents, which we call “orphan terms”, which we have worked around, by providing explicit links to the root node.
- **Slide Transition**
- **Read the Slide**

10 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 Transition**
- Also have another similarity score, and one distance score. Will not mention further here, but we have been experimenting with these also.

11 Validation

- Does it work?
- **Slide Transition**
- Well yes, it does work. Stick two proteins in and you get a value out the other end. But is this biologically meaningful, or useful in any sense.
- **Slide Transition**
- Problem is, its not clear how we are supposed to test this measure. User studies perhaps?
- **Slide Transition**

- Well biology tells us if we have two similar protein sequences, then surely we should have similar annotation?

- We can test this.

- **Slide Transition**

- 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 Transition**

- 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.

- This serves as our primary validation of the measure.

- **Slide Transition**

- One of the features of GO, is that each association between a term, and a protein is annotated with evidence code, saying what evidence there is for the association.

- Only three are commonly used in SWISS-PROT. These three can be ordered into a rough hierarchy, TAS being best, followed by NAS, followed by NR.

- Same experiment as before, but using associations with different evidence, over the molecular function aspect. TAS data shows best correlation.

- This both validates the measure, and also provides large scale statistically evidence to support the assertion that TAS evidence is best.

- **Slide Transition**

- Also interested in what contribution the relationships are providing to the measure. Is possible for us to "flatten" GO into controlled vocabulary. Similarity measure then becomes direct match scaled by information content.

- Same experiment again, using molecular function aspect.

- Can ignore "part of" relationships. Makes no difference, as almost all relationships in molecular function are "is-a"

- If we ignore all relationships get a much poorer match. Match gets better as we move to high sequence similarities...chance of exact matches increase in this area.

- Again as we would expect, again validating usefulness of GO, and validating our measure.

12 Scatter

- But is this measure actually any practical use?
- This slide is same data as before (function against sequence similarity), but shown as a scatter (only showing about 1/5 of data points, or acrobat has fits).
- Very wide spread. But very few points in top left corner.
- Whats happening with protein pairs in this area.

13 Outliers

- Selected a whole bunch of these outliers.
- Examined them by hand. Somewhat laborious.
- Not go into detail here, as there are many pairs.
- Some perfectly sensible.
- Other like this example, where not. Spermine synthase annotated as "spermidine synthase". These enzymes act in the same pathway, one after the other, but are totally different enzymes (different E.C. numbers).
- Several examples like this. "insulin" and "insulin-like". Use of obsolete terms. One or two problems with GO.
- GO people say these have all been fixed now!

14 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 Transition**
- Search with biological process. All proteins involved with vision.
- **Slide Transition**
- With Cellular Component. All membrane proteins
- GO does not (or did not!) differentiate between membranes, hence get both internal and cellular membrane proteins.

15 Conclusions

- **Read the Slide**
- Can be applied to any database.
- Does not require expert curation beyond what is already available, or hand augmentation of GO.
- Should scale well.

16 Future Work

- **Read the Slide**
- link types...we have just conflated "is-a" and "part-of", which is not entirely satisfactory, but we are not sure what to do about it.
- Performance optimisation. Takes about 30 seconds to search over SWISS-PROT, but is a complete memory pig.

17 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.

18 Irrelevant Cartoon

Read the Slide