1 Intro

- Work done when.
- PRECIS is producing automated annotation for the PRINTS database.

2 RoadMap

What is... First going to cover precisely what annotation is

Where does... Where does annotation come from both in general, and specifically for the PRINTS database.

What does... Where does PRECIS fit into all of this

Results Show some of the results of PRECIS

Conclusions And some conclusions and future work.

3 Sequence data

- Lots of biological data of many sorts
- Large range of data types.
  - Sequence data
  - Micro array data
  - Genetic data
  - Protein interaction data
  - Expression data
  - And many more....
- But most of data in molecular biology is built on top of sequence data. Will explain what is meant by "on top of" later.

Reasons for this.

- We have lots of sequence data, particularly since the genome projects, when the amount of sequence data has expanded beyond the point of all sanity.
- Simple data type. Everybody agrees what a "C" means. Compare to "positive genetic interaction".
- Sequence data is relatively hard, so the data does not need to represent multiple opinions, or multiple experiments. For instance biological sequence is usually presented as a string, whilst most other forms of data are supported by the primary data.

4 Sequence data is Opaque data

- Sequence data is hard to interpret.
- Rather is actually easy to interpret as a sequence
- But what does it mean, or imply.
- Often associate other data with sequence data
- This is what I mean by built “on top of”
- The data usually relates in some way with the sequence.

5 SWISS-PROT entry

- SwissProt entry fairly large.
- *Can you read this at the back? Good*
- Only a small amount of this is sequence.

Move Transition

- In this case I’ve chopped out some of the annotation.

In other words whilst SwissProt is normally described as a “protein sequence database”, its actually mostly an annotation database.

6 Annotation Pipeline

- There are a large number of databases in the world.
- Primary and secondary. All seem to have annotation though
- Good illustration of one annotation pipeline.
- EMBL, DNA sequence, SW protein, PRINTS protein families.
- As we move along the pipeline the nature of the annotation changes. More on this later.

PRINTS is therefore a data on multiple SWISS-PROT entries. It involves collation of data.

7 Problem

What it says on the tin.

8 PRECIS Objectives

- PRECIS is an annotation to annotation transformation tool.
- Would like PRECIS to operate over word based data, but this is much harder.
- How far can we get using simple techniques.

9 High or Low Level

- Annotation covers a multitude of sins
- Low level data is structured.
- Often relates to other databases
- Although not always, for example species data, is structured, but not a database. “animal, chordate, mammal, rodent, mus musculus”. Systematic.
- Low level database relates to a wide range of data. For instance sequence data bases, literature cross reference, taxonomy.
- High Level is often unstructured, or free text.

10 Other systems

There are other systems available for annotation generation. Not going into these in detail

Move Transition

They all generally focus on lower level annotation, and large scale analysis (genomic).

11 Knowledge

There are many ways in which we can extract knowledge from existing database entries.

11.1 Database Structure

- Structured information in the database.
- Usually the easiest method.
- Multiple syntaxes is a bit of a pain

11.2 Words

- Failing that we can analyse the words
- Multiple systems which already do this
  - Easy
  - Protein Annotators Assistant
  - AbXtract
- These tend not to provide contextual information, but keywords.
- This is nice as far as it goes.

11.3 Domain Knowledge

- The standard annotation pipe line uses large amounts of domain knowledge, as its done by humans.
- This is problematic to replicate computationally.
- Mostly knowledge is built implicitly within the application. PRECIS makes extensive use of this.
- Can also build knowledge explicitly as an ontology. At the moment PRECIS does not do this, but it would be nice if it did.
- Can combine human and machine annotation.

12 PRECIS phases

- How does PRECIS work?
- Can be split into multiple distinct phases
  - Fingerprint formation, ID gathering. Don't care about this here.
  - Annotation gathering. Get lots of stuff from lots of places. Mostly SWISS-PROT, but also medline.
  - Information culling and report generation phase.
  - Report formatting.
- At the moment this phase and the last phase are rather more intertwined than they should be, and we would like to separate them out more.

13 Harvesting and Generation

There are lots of filters, and transformations used. Going to cover the various types here.

13.1 Ranking

- Too much data to present.
- Need to rank the data and only present the “best”.
- “Show me everything you know…” is not an option
- We have too major mechanisms for Ranking

  Majority Voting Extract the information which occurs most frequently in the input SWISS-PROT entries.

  Golden Voting Some data we hold to be of particular importance, therefore even a single occurrence is enough.
- Various examples of this ranking. Will show in more detail later.

13.2 Redundancy Checks

Redundancy checks. Vitally important. Much of SWISS-PROT is repetitious, particularly looking at multiple entries of related sequences.

- Vitally Important. SWISS-PROT is repetitious anyway
- This is made worse as we are looking at multiple related entries.
- In some cases its very easy. When looking at database ID’s for instance
- **Move transition**
- Its also relatively straight forwards for some “free text”, which in reality is a lot less “free” than it appears to be at first site. For example....
- **Move transition**
- However in some cases its taxing. Last two are obvious word order alterations, but the first sentence? Is “pi turnover” a synonym, or what?

13.3 Heuristics

- Heuristics. There are quite a few of these.
- One example though is the SWISS-PROT ID heuristic.
- Explain ID’s
- Do we have a family or a super family?
- If 75% of identifiers are the same, we have a family.

Are Heuristics the way to go

One problem is that they are clearly not generalisable, so you have to re-write them for each application. On the flip side it does produce useful output. Its possible that many of the heuristics will be reusable between databases. And this is partly how annotation is generated at the moment.

14 PRECIS Output:-Databases

First identifier line. Direct from swissprot using ranking (majority voting).

Database Cross links. Also from SW, ranking (majority voting, and preferred databases).

15 PRECIS Output:-References

Again based on SW references.

Using date criterion, mostly commonly occurring reference. And keywords.

One problem with commonly occurring references is that this tends to select heavily "genome" references which are fairly uninteresting.

Like to investigate various clustering technologies, based on term recognition, to perhaps remove this problem.

16 PRECIS Output:- Description

Extraction based on commonly occurring sentences.

17 PRECIS OutputDisease

In this case a single disease association is enough for appearance in the end report, because this information is particularly important. Includes provenance.

18 The rest

- As with disease structure information is useful.
- Family designation extracted on a majority vote
- Keywords by majority.

19 Strengths

The reports are easy to read which is always nice.

Retains context information, which is good. Its unlikely that any extraction system is going to be perfect. Must allow the biologist to make their own decisions.

Provenance. Again its hard to trace data back. Can only do this poorly in swissprot, and PRINTS. Wanted to improve this situation.

Results are updateable. Standard problem with all databases. Data goes out of date straight after you have got it. PRECIS can update the data automatically.

20 Future Direction

Implementation. Have already started on a Java implementation, using XML transformations on the data. This also provides us with a syntax to enable a structured meta data layer. Pluggability should allow easier reuse of heuristics than at the moment.

Ultimately this should make the lives of secondary users of this data much easier. Should make it possible to combine human and machine generated annotation, to provide a half way house between these two.

21 Acknowledgements

- Jacqueline Reich:- Implementation, and user analysis.
- Alex Mitchell:- Integration of PRECIS into existing pipelines
- Robert Stevens:- Paper writing, and initial inspiration
- Terri Attwood:- Expert annotation, and data.
- Carole Goble:- General lab head stuff...