USING NATURAL LANGUAGE PROCESSING TO FIND RELATIONSHIPS BETWEEN GENES, DRUGS AND CANCER

Rusaw S, Bajdik C, Varhol R, Zuyderduyn S, Astakhov V, Jones S

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

Genome Sciences Centre

www.bcgsc.ca
info@bcgsc.ca

1. Introduction

This poster presents our work in using basic Natural Language Processing (NLP) to extract meaning from MEDLINE abstracts. In conjunction with the Genome Sciences Centre's DISCOVERYspace application, this work will allow researchers to take a gene expression profile and relate the up/down-regulated genes to possible drug targets, similar cancer types and other biological concepts.

2. Natural Language Processing (NLP)

The two basic NLP techniques used to extract information from MEDLINE are part-of-speech (POS) tagging and noun phrase chunking.

Part-of-Speech Tagging

A POS tagger takes arbitrary text and tags each word with an identifier to indicate if it is a noun, verb, adjective etc. We used the Brill Tagger to tag MEDLINE text using the default lexical and contextual rules that come with the source code. This has the positive effect of tagging unknown words (most genes, drugs, proteins, diseases etc.) as nouns. Here is an example (/XX is a tag):

Before Tagging

The genome of the mouse is widely regarded as one of the keys to understanding the human genome sequence .

After Tagging

The/DT genome/NN of/IN the/DT mouse/NN is/VBZ widely/RB regarded/VBN as/IN one/CD of/IN the/DT keys/NNS to/TO understanding/VBG the/DT human/JJ genome/NN sequence/NN ./.

Noun Phrase Chunking

A noun phrase chunker is used to extract 'concepts' from tagged text based on a set of patterns of POS tags that comprise a noun phrase. These noun phrase concepts are the key component of our analysis, so the fact that the POS tagger marks all unknown words as nouns in effect ensures that most genes, drugs and cancers will find their way into a noun phrase. Here is the above example further decomposed into noun phrases (noun phrases are between '[' and ']'):

After Chunking

[The/DT genome/NN] of/IN [the/DT mouse/NN] is/VBZ widely/RB regarded/VBN as/IN [one/CD] of/IN [the/DT keys/NNS] to/TO understanding/VBG [the/DT human/JJ genome/NN sequence/NN] ./.

3. Noun Phrase Repository

For this work we selected a small subset of MEDLINE abstracts (about 1.2 million sentences from 2000/01) and extracted all the noun phrases using the techniques described above. The resulting noun phrases are stored in a relational database along with information about which MEDLINE abstract and sentence the noun phrase occurred in.

The Genome Sciences Centre has a computing resource of approximately 80 dual processor servers configured as a Linux cluster. Processing the MEDLINE text takes a great deal of computing resources, and generating the noun phrases for <5% of MEDLINE took about 2 weeks. Here are some interesting statistics:

- 445,053 words were extracted ('the', 'The' and 'THE' are different words).
- 2,264,940 noun phrase were extracted.
- The most common word was 'the' (not a big surprise).
- The most common 2, 3 and 4 word noun phrases were 'this study', 'the present study' and 'the central nervous system'.

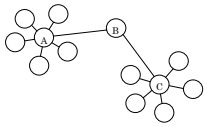
The MEDLINE noun phrase repository is currently an internal resource that hopefully will find other uses than simply linking drug, genes and cancer.

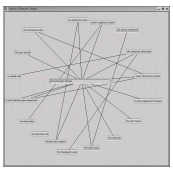
4. Linking Arbitrary Noun Phrases

During the noun phrase extraction phase, relationships between noun phrases are determined based on their co-occurrence in MEDLINE abstracts and sentences. In this work, two noun phrases are said to be *related* if both occur in the same abstract, and *strongly related* if both occur in the same sentence.

These relationships are stored in a number of relational database tables that can be queried based on various filtering criteria. Queries to find all the nouns related to lung cancer are easily constructed, but really provide no more novel biological information than a clever keyword search on PUBMED. The novel information comes from a higher level analysis of the *structure* of the noun phrase linking network. Consider the following figures:

This graph shows a noun phrase 'A' with a number of direct neighbours and a noun phrase 'C' with a number of direct neighbours. 'A' and 'C' share a common neighbour 'B', which acts to indirectly link the concepts 'A' and 'C'.





This graph shows a number of noun phrases related to noun phrases containing the words 'gene' and 'therapy'. The connecting lines show the inter-relationships between the phrases.

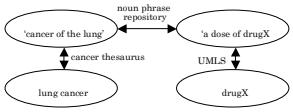
Now, if 'A' was a cancer, 'B' was a gene and 'C' was a drug, then the NLP analysis would have found an new indirect relationship between the drug and the cancer.

5. Linking Drugs, Genes and Cancer

Linking biological concepts like genes, drugs and cancer (or proteins and pathways for that matter) is a two step process. The first step is to link noun phrases in the noun phrase repository with data sources that supply information (names, synonyms, alternate spellings etc.) about gene, drugs and cancer. We used the following data sources:

- Gene information was supplied by The Weizmann Institute's GeneCards database.
- Drug information was supplied by NLM's Unified Medical Language System.
- Cancer info was supplied by an in-house thesaurus of cancer synonyms.

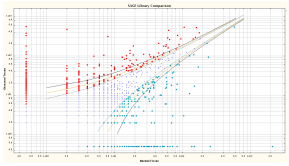
The second step of linking genes, drugs and cancer occurs through the noun phrase repository, where information about noun phrases co-occurrence in MEDLINE abstracts and sentences. This information is used to determine if particular drugs, genes and cancer are related or strongly related. The following graph shows how a cancer is related to a drug:



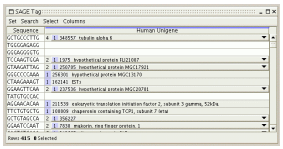
6. Processing SAGE Data With DISCOVERYspace

When incorporated with the The Genome Sciences Centre's DISCOVERYspace application and associated SAGE plugin, this NLP work allows the researcher to analyze SAGE data in novel ways. Here we present a representative workflow

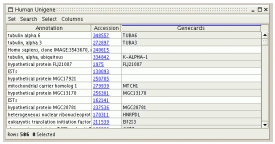
1. Two SAGE libraries (normal and diseased tissue) are compared on a 2D scatterplot and all up-regulated tags (red) are selected.



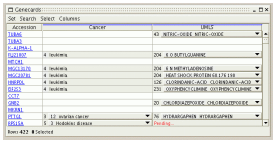
2. The up-regulated tags are mapped to genes (Unigene) via DISCOVERYspace's database backend DISCOVERYdb.



3. The Unigene entries are mapped to GeneCards (the entry point for our NLP work) via DISCOVERYdb.



4. The GeneCards entries are mapped to related cancer types and UMLS drug entries via the noun phrase repository.



An NLP plugin for the DISCOVERYspace application is currently in preparation, and will allow the users of DISCOVERYspace to begin exploring these relationships.

Acknowledgements and References

This research was funded by the BC Cancer Foundation and Genome Canada

Brill, E. 1992. A Simple rule-based part of speech tagger. *Proceedings of the Third Annual Conference on Applied Natural Language Processing*, ACL.

Lindberg, D, Humphreys, B. and McCray, A.. 1993. The Unified Medical Language System. *Methods of Information in Medicine* 32:281-291.

Noun Phrase Chunker - ftp://ftp.cis.upenn.edu/pub/chunker/