Using POS Tagging for the Identification of Gene and Protein Names in Biological Literature

Brian Tucker

Bioinformatics- Fall 2002

Introduction

Ever since the advent of large-scale data acquisition projects in biology, including genome sequencing and microarray data, the issue of information retrieval has become an escalating yet important problem. The time when every researcher working in biology knew everything about the genes/proteins that they work with is long gone. For example, someone who is doing a whole genome expression array with yeast cannot be familiar with every gene that may or may not be expressed on their array chip. The question then arises, how does one obtain information about certain genes or proteins from previously published literature/data? And how can one learn about those genes/proteins that interact with a gene/protein of interest? This would not be such a difficulty if there were a systematic organization of gene and protein nomenclature (for example, protein 1, 2, 3, etc). However, this is far from the actual case. Biological names are as diverse as language itself, incorporating not only nouns but also verbs, adjectives, and prepositions. Additionally, many contain numbers and symbols, refer to other biological entities (ex – HIV protease), and sometimes there are even several terms that refer to the same thing (ex – caspase-3, CASP3). However, there is an active area of research dedicated to constructing computer programs based on POS tagging, which are able to extract gene

and protein names from the literature, making it easier to compile data on certain genes/proteins and the molecules that interact with them.

Part-Of-Speech (POS)

There are many programs out there, such as PROPER (PROtein Proper-noun phrase Extracting Rules), which rely on simple rules to deduce if a word is a gene/protein name (Fukada et al, 1998). For example, words are tagged if they contain a symbol or number. However, one of the most fundamental and basic approaches used in identifying gene/protein names is the use of a part-of-speech (POS) tag. Developed in the early 1970s at Brown University, this system simply assigns each word in a sentence its part of speech (Greene et al, 1971). A computer does this relatively easily, deriving its information from some sort of compiled lexicon, which has words with their part of speech.

The POS tag is really just a base for further analysis since, because of the complicated nomenclature of biological names (as described above), a POS tag would not necessarily distinguish between a gene/protein name and another part of speech if the name contained verbs, for example. Because of that, many programs incorporate POS tagging into systems with the addition of rules and error-correction for false positives and negatives. Additionally, POS tagging can be trained by checking against some sort of manually confirmed database.

Increasing Precision with Rules (The Brill Tagger)

The fundamental goal of identifying gene/protein names is to accurately tag a word or group of words as a "gene" (for simplicity, I will refer now to just gene names, although the same case would apply to protein names also). When the tagging program encounters a new word, which may be a gene, it needs rules in order to 'disambigulate' that word. This can be done by having a list of manually derived rules that the program follows (Proux et al, 1998). However, a more effective method at constructing rules is through

error-driven learning. A wonderful example of this method is the Brill tagger (Brill, 1992).

The Brill tagger, which is based upon POS, does not require a manually constructed list of rules. However, an initial 'training corpus' is needed in order to teach the program. For example, in a program devised by Tanabe et al, a list of 7000 hand-tagged sentences were provided, which allowed the Brill tagger to deduce rules from the context of the gene names (Tanabe et al, 2002). This allowed the tagger to produce 78 new lexical rules and 81 new contextual rules. An example of a lexical rule could be "tag a word as 'gene' if it contains a greek symbol" whereas a contextual rule might be "tag a word as 'gene' if it is proceeded by the word 'gene'". From these rules, the Brill tagger can go on to annotate words not listed in the lexicon.

Post-Processing of Tagging Data: False Positives and Negatives

Although the Brill tagger is an expeditious way of identifying gene names, it is, of course, subject to error. There is then a need to filter out false positives while, at the same time, finding false-negatives.

False positives can be filtered by checking tagged words against a compiled list of words known not to be genes, such as general biological terms like 'operon' or 'antigen'. As for false negatives, a number of approaches can be taken. The easiest is to check it against a large database of known gene names. A more sophisticated approach taken by Tanabe et al is to check against a list of contextual words, like 'inducing' or 'truncated' (Tanabe et al, 2002).

Conclusions

The use of POS-based tagging in identifying gene and protein names has had relative success, although it does suffer from complications, especially when dealing with complicated names like compound names (Tanabe et al, 2002). However, this system is flexible to refinement, since not only can you train it in a certain way but you can also post-process the results. One of the best ways to refine identification might be to further

contextualize the putative positives. For example, a gene or protein name would have a greater tendency to appear in certain areas of the literature, as in the abstract of the article or in a table of results. Additionally, the POS tag is simplistic in that it refers to basic parts of speech (verb, noun, preposition), without relating the parts of speech in a sentence. However, a program could be devised that understands grammar and can ascribe certain words as the 'subject' of a sentence, and then deduce from that gene or protein names, which will often appear as the subject (ex- "IDH has a Rossmann fold", where IDH is the subject).

References

Brill, E. *A simple rule-based part of speech tagger*. Proceedings of the DARPA Speech and Natural Language Workshop. pp. 112-116. Morgan Kauffman. San Mateo, CA.

Fukada, K., Tsunoda, T., Tamura, A. and Takagi, T. (1998). *Toward information extraction: identifying protein names from biological papers*. Proceedings of the Pacific Symposium on Biocomputing (*PSB98*). Pp. 705-716.

Greene, B.B. and Rubin, G.M. (1971). *Automatic Grammatical Tagging of English*. Providence RI: Department of Linguistics, Brown University.

Proux D, Rechenmann F, Julliard L, Pillet V V, Jacq B. Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. Genome Inform Ser Workshop Genome Inform. 1998;9:72-80.

Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics. 2002 Aug;18(8):1124-32.