Text Detective: BioAlma's gene annotation tool

Javier Tamames
BioAlma SL, 28750 Tres Cantos (Madrid)-SPAIN

Introduction

Our system for annotation of articles is named "Text Detective". It is capable of annotating a wide range of biological entities, such as genes, proteins, chemical compounds, drugs, diseases, biological processes, etc. For the purposes of BioCreative, this document only refers to the gene annotation machinery.

Common steps for task 1

It is important to notice that gene names can be found in two different ways: as **full names** (a functional description of the gene, such as "tumor necrosis factor" or "janus kinase"), and as **gene symbols** (an abbreviation or acronym, such as "TNF" or "JNK"). Our approach is different for detecting both instances, since the associated problems are different as well.

The first step in the process is to parse (split in sentences, remove punctuation, etc) and tokenize the document. Then, every sentence in the document is processed independently.

Text Detective is then able to tag every word in the sentence according to biological relevant categories. For instance, chemical compounds are recognized and labelled. The identification of "central words" (also known as "core terms") is a key step in this process (words such as "receptor", "kinase", "transporter", etc). For this purpose, we have built a lexicon and used some carefully curated rules.
Also "types" are tagged (words such as "alpha", "a1", "c", "12", "TNF"), since they may define the exact identity of the gene (distinguishing between "interferon alpha" and "interferon gamma", for instance). These "type" words are recognized after a set of carefully designed rules (presence of capital letters, numbers, Greek letters, etc.). Notice that gene symbols (such as "TNFalpha") are also tagged as "type".

The full list of categories that apply in this tagging process is: "CENTRAL", "TYPE", "LOCATION", "BIOWORD", "CHEMICAL", "VERB" and "OTHER". Notice again that this is NOT a part-of-speech tagging, rather we try to recognize the role of the word in a possible gene mention.

An example of a tagged sentence follows:

Decay -> BIOWORD
accelerating -> BIOWORD
factor -> CENTRAL
(DAF) -> TYPE
is -> VERB
a -> OTHER
complement -> BIOWORD

regulator -> CENTRAL
that -> OTHER
dissociates -> VERB
autologous -> BIOWORD
C3 -> TYPE
convertases -> CENTRAL
which -> OTHER
assemble -> VERB
on -> OTHER
self -> OTHER
cell -> LOCATION
surfaces -> BIOWORD

This procedure is common both for task 1A and task 1B.

Task 1A

Once all words are tagged, Text Detective attempts to discover gene mentions both as **full names** ("tumor necrosis factor alpha", "interleukin 1") or **gene symbols** ("TNFalpha", "IL 1"). The procedure is different for both instances.

In order to discover full names, the system extracts chains of words that can represent a gene mention. In order to select these chains, they must fulfil several criteria. For instance, No "OTHER" or "VERB" words are allowed, and the presence of a "CENTRAL" word is needed. These resulting chains of words are possible gene mentions.

Gene symbols (TNF, EGR, p53) identification follows a different procedure, since we consider it a different problem. A symbol of this type in the article could refer to a real gene or often to something different (SCT could stand for "secretin" gene, but also for at least 20 more meanings, being "Stem Cell Transplant" the most used).

As we said above, gene symbols are recognized as chains of "TYPE" words (one or more). Then the context (the surrounding words) for each of these chains is evaluated, and scored according a scoring matrix, a list of words that are statistically very frequent in the neighbourhood of a gene symbol (computed using proprietary algorithms). An example of some entries in the scoring matrix follows:

WORD	POSITION					
	-3	-2	-1	+1	+2	+3
gene	0	0.5	5.0	5.0	0.5	0
function	0	1.8	0	2.1	0	0
cell	0	0	-2.5	-5.0	0	0

Then, for the sentence:

"The function of c-fos gene in CC2 cells is partially inhibited"

The possible gene symbol "c-fos" scores 1.8 (word "function" at position –2) plus 5.0 (word "gene" at position +1), total 6.8, while the possible symbol "CC2" scores 0.5 (word "gene" at position –2) plus –5.0 (word "cell" at position –1), total –4.5.

The score for the symbol must exceed a minimum score in order to be declared a valid gene mention. This minimum score can be set up independently for each possible symbol, taking into account factors such as surface clues (presence of capital letters, numbers, greek letters, etc.). In other words, each possible symbol has a different minimum score, so that “AD” is recognized as a unlikely gene name, while “ftsZ” is seen as a very likely gene name.

After this step, we will have a set of possible gene names mentioned in the article, both as full names or as gene symbols. This is the result for task 1A.

Task 1B

In order to identify the exact gene reference (task 1B), we try to match the gene mention we have found (full names or gene symbols) with a list (lexicon) of possible genes (full names or gene symbols), either provided by BioCreative organizers or extracted from relevant databases (HUGO, MGI, SGD, SwissProt, etc.). Again, the procedure is different for full names and for gene symbols.

For full names, we have tagged the lexicon, the list of full names, using the same procedures described above. For instance, the lexicon entry “gamma-aminobutiric acid receptor delta” is tagged as:

gamma-aminobutiric -> CHEMICAL
acid -> CHEMICAL
receptor -> CENTRAL
delta -> TYPE

This is done for all full names in the lexicon.
A match between the full name found in the article and an entry in the lexicon is only scored if several criteria are fulfilled. For instance, all central terms and chemical compounds must be present both in the full name found in the article and the full name in the lexicon. All types must be also present. This matching procedure is very flexible, so that word ordering, dashes, slashes, brackets, etc, do not influence the result. A match is only scored if just one gene in the list fulfil all criteria, and the “official” reference is returned. In case that several genes in the list could match the gene mention, ambiguity is detected and no identification is provided.

For gene symbols, we try to match **all** possible gene symbols found in the article with a lexicon of allowed gene symbols, either provided by BioCreative organizers or extracted from different databases (HUGO, LocusLink, MGI, SGD, SwissProt, etc.). As before, the matching procedure is flexible. But ambiguity can be present and a given symbol can stand for different genes (for instance, gene symbol “mcad” in mouse can stand for the official references MGI:106672 and MGI:87867)

To overcome the ambiguity, we have compiled additional information for every gene in the list. This is usually done extracting annotations and comments from different databases (SwissProt, LocusLink, HUGO, Gene Ontology, etc.), by means of fully-automatic, carefully designed procedures . Using statistical algorithms, we can extract several “key words” from these annotations (words that are relevant for some features of the gene, such as “cell cycle”, “apoptosis”,

“microtubule”, etc.). Therefore, we have a list of “key words” for every possible gene (linked to official references) in the lexicon.

The final step is to look for these key words in the article, to resolve the ambiguity. For instance, if we have found a gene symbol that can correspond to two different genes in the list, we look for the key words of these two genes in the article. The one having more key words is the winner.

Finally, the system assembles the information about full names and gene symbols in one single result, and returns it as the final output.

Discussion of the results- task 1A

The system achieves 84.2% precision, 71.7% recall, performing as the one with highest precision in this task. The recall, however, is on average, and the system will benefit of a improvement in this point.
Nevertheless, we think that in many cases the system identified correctly the presence of a gene/protein mention in the text, but it was not scored correctly since it included (or lacked) one extra word. Some examples follow:

14756 […] cdc42, a **conserved morphogenetic g protein**, […]

In this case, the system matched “conserved morphogenetic g protein”, while only “morphogenetic g protein” was accepted in the Gold Standard (GS).

12722 […] transcription by **mammalian RNA polymerase II** […]

In this case, the system matched “mammalian RNA polymerase II”, while only “RNA polymerase II” was accepted in GS.

14561 […] coordinately regulated opaque-phase-specific **gene PEP1** […]

In this case, only “opaque-phase-specific gene PEP1” was accepted in GS, while our system tagged only “gene PEP1”.

We consider that the system performed correctly in these instances, since it accurately identifies the gene/protein mention. Mismatches of this kind introduce a big penalty in the automatic scoring of the results, since they count as both a “false positive” (an annotation has been made that does not correspond to what was expected), and a “false negative” (we missed one annotation that was expected), therefore lowering both precision and recall.

These imperfect matches illustrate the difficulties of creating a Golden Standard for automatic annotation. We estimate that, in our case, this can introduce a 2-3% bias in precision and recall.

An additional obstacle for our system was the fact that the provided texts consisted in just one sentence per article. In order to evaluate a gene/protein mention, our system takes into account all mentions of the possible gene/protein in the text. The context of all the instances is evaluated, and global features are extracted. That means that the results improve as the text grows,

since we can use more information. We obtain better results when using a complete PubMed abstract (usually between 10-15 sentences).

Also we found difficulties in dealing with the definition of “what-is-a-gene” that was used. For task 1, entities like domains, regions, mutants, mutations, sequences, etc, were considered “gene names”. In the development of our system we did not consider that these should correspond to gene names. Therefore, the system was not able to recover many of these instances.

We estimate that the joint influence of these factors could cause at least 5% less recall and precision.

Discussion of the results- task 1B

As before, the system achieves high precision (80% for mouse, 91% for yeast), with good but lower recall (70% for mouse, 81% for yeast). We have performed a careful check of our results for mouse, and extracted the following conclusions:

There are two factors that influence greatly the recall:

-First is the annotation of full names that do not match exactly the lexicon. For instance, in the article mouse_00001, the human experts annotated “fibronectin” (MGI:95566). But in the lexicon, the related entry refers to “fibronectin 1”. In such cases, our system is tuned to “think” in the following way: “If I find ‘fibronectin 1’ in the lexicon, it is likely that ‘fibronectin 2’ also exists, even if it is not present in my lexicon. Therefore, I must consider that ‘fibronectin’ is ambiguous, since it probably refers to different fibronectins”. In our experience, this feature increases the precision by lowering the recall. This can be the case for 10% of our false negatives.

-There are also several instances of difficult cases, in which the lexicon is not enough to annotate the gene. It is the case for article mouse_00006, in which the notation “lpa(1-3)” is used to refer to lpa1, lpa2 and lpa3. Or in article mouse_00021, where “PKC beta” is cited in the article, while in the lexicon we find “protein kinase C beta”. Or also in article mouse_00024, where we find “pol gamma B”, while the equivalence in the lexicon is “pol gamma 2”. These instances present a real challenge for an automatic system, and very sophisticated rules must be devised to deal with them. We find that this is the case for 20-25% of our false negatives.

For the precision, we find that at least 20% of our false positives are due to instances of a mouse gene in other organisms, mainly human. It is the case for “*dcx*” in the article mouse_00033, talking about the human ortholog of the mouse gene *dcx* . We did not tune the system to deal with this fact.

Another important source of false positives is related to the lexicon: it is the case in which a single gene is annotated in the lexicon with a name that actually designates the whole family it belongs. An instance is “eph”, that a lexicon entry links only to MGI:107381 (Eph receptor A1). Indeed, “eph” refers to a wide family of receptors. The effect of this is that the system will assign to MGI:107381 any mention to “eph receptors”. This can explain at least 25 % of our false positives.