Assessing the correlation between contextual patterns and biological entity tagging.

M.KRALLINGER, M.PADRÓN, C.BLASCHKE, A.VALENCIA

Protein Design Group,
National Center of Biotechnology (CNB-CSIC),
Cantoblanco,
E-28049 Madrid,
martink, mpadron, blaschke, valencia@cnb.uam.es

Abstract

The tagging of biological entities, and in particular gene and protein names, is an essential step in the analysis of textual information in Molecular Biology and Biomedicine. The problem is harder than was originally thought because of the highly dynamic nature of the research area, in which new genes and their functions are constantly being discovered, and because of the lack of commonly accepted standards. An impressive collection of techniques has been used to detect protein and gene names in the last four-five years, ranging from typical NLP to purely bioinformatics approaches. We explore here the relationship between protein/gene names and expressions used to characterize protein/gene function. These expressions are captured in a collection of patterns derived from an original set of manually derived expressions, extended to cover lexical variants and filtered with known cases of association patterns/ names. Applying these patterns to a large collection of curated sentences, we found a significant number of patterns with a very strong tendency to appear only in sentences in which a protein/gene name is simultaneously present. This approach is part of a larger effort to incorporate contextual information so as to make biological information less ambiguous.

1 Introduction

Molecular Biology and biomedical research covers a broad variety of research topics, connected to the function of genes and proteins. The information on the experimental characterization of essential functional aspects of these genes and proteins is manually extracted from primary scientific publications by field-specific databases. This process requires highly specialist personnel, and is costly and time-consuming. Indeed, only a small number of genes and proteins have been annotated with information directly related to experiments, whereas in the immense majority of cases the annotations are transferred from other similar entries. The annotations provided by the databases are a valuable source for large-scale analysis, but are inevitably incomplete at the level of detailed function and experimental results.

It is in the context of fast-growing bibliographic information (over 12 million references are collected in the PubMed database, with an average of 500,000 new references added every year) and annotation of the function of genes and proteins that Text Mining and Information Extraction systems become important tools for biological research (Blaschke and Valencia, 2001).

Since the first papers were published in this field in the late 90's, it has become clear that the detection of gene and protein names (gene tagging) is a key first step towards Text Mining systems becoming really useful.

The detection of names is particularly complex in the domain of Molecular Biology, for a number of reasons:
(1) Sociological, since names are perceived as associated with the recognition of the groups that first discovered them.
(2) As biologists tend not to adopt available naming standards, often the disease related to a gene disorder has the same name as the gene itself (homonyms). This can be only be addressed using context based sense disambiguation procedures.
(3) Gene names or symbols are often the same as common English terms. For instance, many *D. melanogaster* gene names, such as 'hedgehog', lead to lexical ambiguity. (4) Symbols and abbreviations are commonly used without any control. This gives rise to the problems of acronym disambiguation and expansion. There is no high-quality gene acronym dictionary.
(5) Proteins are related by a process of evaluation, which creates ontological associations that

are mixed with the various levels of knowledge for different members of the protein families.
(6) The field itself is still evolving, and the catalogue of genes even for the genomes already sequenced, such as the Human one, is still incomplete.

Our own assessment of the evolution of gene names shows that names evolve over time into a complex system with scale-free behavior, with the presence of a few very oft-quoted names (attractors) and many very seldom quoted ones. The system itself is in a critical state and the fate of current names is unpredictable (Hoffmann and Valencia, 2003).

A significant number of applications have been developed to identify gene names and symbols in the biomedical literature (see (Tanabe and Wilbur, 2002; Yu et al., 2002; Proux et al., 1998; Krauthammer et al., 2000) for four different methodological approaches). In order to assess the performance of different approaches the BioCreative challenge was carried out. The recent BioCreative challenge showed that gene and protein names can be detected by several techniques, with a significant success that can be as high as 80% for the best-performing systems (Blaschke et al. in preparation; and special issue of BMC Bioinformatics on the BioCreative challenge cup, in preparation). However, detection of the remaining 20% of names is really important for many operations. Therefore, there is significant room for improvement, and a clear need for new approaches able to use alternative sources of information.
We explore here a new avenue for the detection of gene and protein names by using contextual information, since in many cases gene tagging requires knowledge of context (context-based approach for disambiguation). We previously explored relevant information by creating context-based sentence sliding windows for entity-relation extraction (Krallinger and Padron, 2004).

We propose here to detect those sentences describing the function of genes and proteins in the literature that are good candidates for containing unambiguous information about corresponding gene and protein names.

To detect these sentences, we relied on the identification of typical expressions (patterns) associated with the description of protein function in the literature. Context information in the form of heuristically extracted sentence patterns, known as frames, proved useful in the past for deriving protein interactions automatically (Blaschke et al., 1999; Blaschke and Valencia, 2001) from protein co-occurrences.

The approach proposed here is based on the extension of heuristically derived *trigger words* (Riloff, 1993; Agichtein and Gravano, 2000) and the filtering of patterns using previously gene-indexed sentences. The extraction patterns obtained were then ranked, using a validation set of gene-indexed sentences and sentences lacking the gene symbols. Precision-ranked extraction patterns and indexing of sentences using those patterns allowed ranking of these sentences according to whether they contained relevant information for protein indexing and annotation.

2 Methods

In the case of complex domains, such as Molecular Biology and Biomedicine, a prohibitively large training set is generally required in order to mine scientific literature. Often inter-domain portable methods do not perform well enough. Nevertheless, within relevant sentences containing protein or gene names, commonly used patterns often describe or define relevant aspects of those entities.

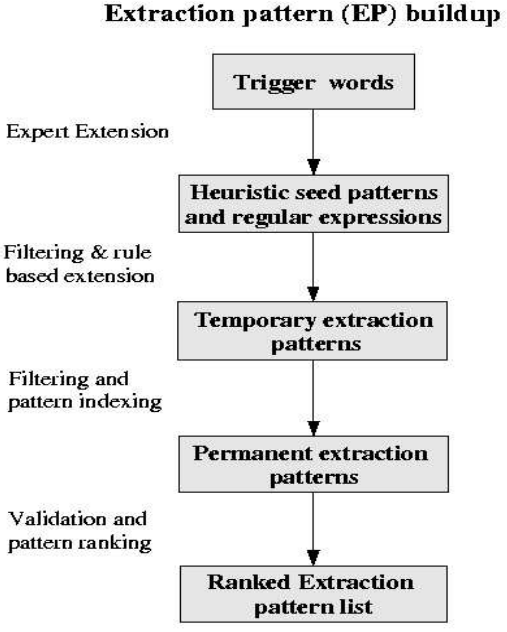

Figure 1: Flow chart of the main steps for constructing the extraction pattern set.

Therefore, a list or dictionary of such extraction patterns was developed, starting with a small list of trigger words which, after sev-

eral processing and filtering steps, resulted in a ranked list of protein-specific extraction patterns (see Figure 1).

2.1 Set of trigger words

First, a domain expert manually analyzed gene-indexed sentences to extract key words that could trigger potential extraction patterns. The expert used heuristics based on background knowledge of the domain. These trigger words (Riloff, 1993; Agichtein and Gravano, 2000) constituted frequent word types which, in the context of other word types (often prepositions or articles), displayed a strong association with the given gene or protein entity. Trigger words thus made up a sort of concept node [1] by scanning through gene-indexed sentences. Most of these trigger words were in fact verbal phrases (e.g. transitive verbs), which were often encountered in sentences defining or describing relevant features of genes and gene products. Therefore, only trigger words which helped describe or define relevant aspects of the protein were extracted. These trigger words are also useful for computerized annotation of extraction of proteins. Among the trigger words were 507 verbs, 127 adjectives and 265 nouns.

2.2 Heuristic trigger word extension

The trigger words were then extended and combined by the domain expert using context-based heuristics to extract a seed set of initial extraction patterns and a set of regular extraction expressions. An example (1) of the heuristic trigger words used was *'encoding'*. Among the resulting expert-derived extraction patterns were: *', encoding a', ', encoding the', 'encoding a','encoding the',"encoding a <PROT>','gene , encoding'* and *'protein, encoding'*. Here <PROT> represents a previously gene tagged word type.

2.3 Automatic extension of seed extraction patterns

To extend the set of extraction patterns and to expand the regular expressions to obtain defined patterns, a rule-based system was used. Among the extension rules for these seed patterns were preposition substitutions, comma addition before verbs, article insertion before certain nouns and pattern fusions. Some of the patterns generated were revised manually and inconsistencies were removed. Examples of the extraction patterns based on the seed patterns provided in example (1) were *'the gene encoding the' , ', a gene encoding', ', a gene encoding the',', gene encoding a',', the gene encoding',', the gene encoding a','the gene encoding the'*. Some of the extensions did not correspond to natural language and some were too long. Thus, in a second step, those patterns not encountered in free text, namely the initial set of gene-indexed sentences, were removed.

[1] Concept nodes are essentially case frames which are triggered through a lexical item and its corresponding linguistic context (Riloff, 1993)

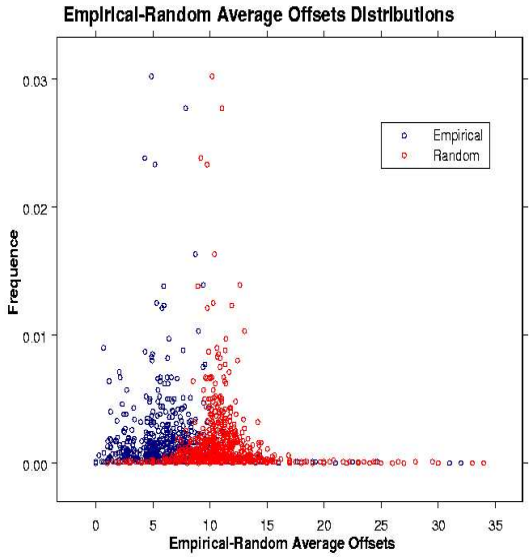

Figure 2: Emprirical and random distribution of the pattern to gene name average offsets.

2.4 Temporary extraction pattern filtering.

The extended set of extraction patterns had to be analyzed as to whether they really corresponded to patterns encountered in sentences in which gene and protein names or symbols were found. Therefore we tagged, using exact pattern matching, the temporary set of extraction patterns to a set of previously gene-indexed sentences.

These sentences contained gene symbols of the yeast *S. cerevisiae* provided by the SGD database. A total of 36,543 sentences were generated with the use of a refined gene tagger. In general, as yeast genes are easier to tag and on the whole do not correspond to common English word types, they became a high-quality gene-indexed data set for further analysis using offset statistics.

A total of 769 patterns were matched to these sentences and the rest of the patterns were dis-

carded. To determine whether those matched patterns had a distance association to the gene names, we calculated the empirical average offset of each pattern. The distances used for the offset calculation were measured in word tokens.

Thus average empirical offset $\bar{d}_e$ was calculated by

$$\bar{d}_e = \frac{\sum_{i=1}^{n} d_i}{n} \quad (1)$$

where n is the number of occurrences of the given pattern in the gene indexed sentences and d_i is the observed offset.

Taking into account the sentence length, the individual pattern length and the gene position within the pattern matching sentences also a random offset was calculated for each pattern occurrence and the average random offset for each pattern, $\bar{d}_r$ was calculated (see Figure 2). In order to determine whether the average empirical distance of the patterns were significantly different from the corresponding random offsets, the distributions were further analyzed. A chi-square test was applied to verify that both, the empirical and the random offset distributions were normally distributed. Then we used the Kullback-Leibler divergence to measure how different the two probability distributions were :

$$D(p\|q) = \sum_i p_i \log_2(p_i/q_i) \quad (2)$$

where p corresponds to the normal distribution of the empirical average offset and q to the normal distribution of the random average offset. In our case the distributions showed a large KL divergence. This means that $\bar{d}_e$ is significantly smaller when compared to $\bar{d}_r$ (i.e. the patterns are closer to the gene names).

To be able to use the average offset differences of the patterns as a filtering criterion we calculated the distribution of the differences δ_i between $\bar{d}_r$ and the corresponding $\bar{d}_e$ (see Figure 3). Only patterns with $\delta_i > 0$ passed the selection filter.

2.5 Permanent extraction pattern ranking.

After the filtering of the temporary extraction patterns using gene/protein indexed sentences, it was important to determine the precision of the extraction patterns for gene-indexed sentences, compared with sentences without mentions of genes. Therefore, two validation sets were constructed: one containing a set of gene indexed sentences using gene names contained in the SwissProt database, and the other consisting of the remaining sentences, without those symbols. The sets were used to calculate recall and precision:

$$R = \frac{TP}{TP + FN} \quad (3)$$

$$P = \frac{TP}{TP + FP} \quad (4)$$

The corresponding f-score is given by

$$F - score = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha)\frac{1}{P}} \quad (5)$$

Where P is precision, R is recall and α, which consists in a weighting factor for precision and

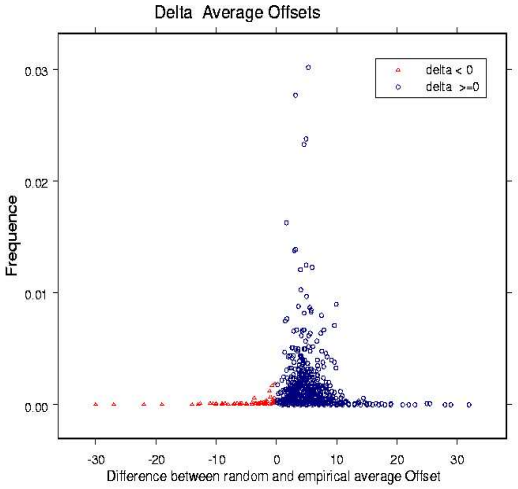

Figure 3: Difference between the average random and average empirical offset. The extraction patterns which did not pass the filtering step (difference > 0 are displayed in red. The remaining pattern set (blue) constituted the permanent pattern set which was used for the f-score ranking.

Data set	Total
Initial trigger words	899
Seed heuristic patterns	472,427
Extended heuristic patterns	525,408
Filtered heuristic patterns	53,185
Temporary patterns	769
Permanent patterns	655
Gene indexed sentences	36,543
Validation sentences (+))	45,119
Validation sentences (-)	45,119

Table 1: Overview of the used dataset of patterns and sentences.

recall, here $\alpha = 0.5$, were both precision and recall had the same weight. Regarding the obtained f-score or the precision we could rank the permanent extraction patterns.

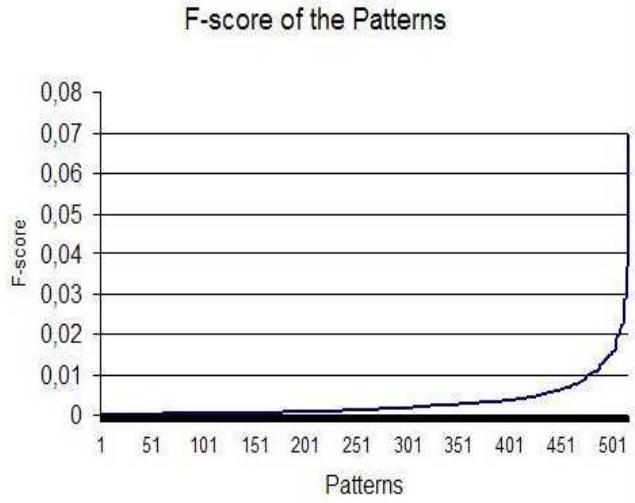

Figure 4: F-score plot for the extraction patterns in the validation set: f-score for each pattern (numbered)

There were a total of 59.82% of high precision patterns, i.e. with a precision greater then 0.8, of which 64.86% had a precision of 1. Thus the patterns are in general very specific for gene containing sentences.

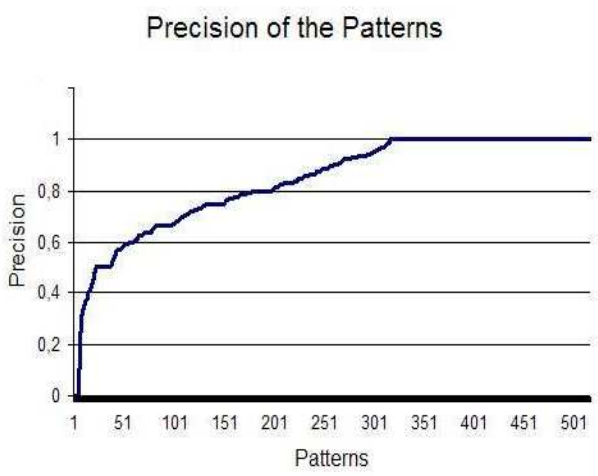

Figure 5: Precision plot for the extraction patterns in the validation set: PRECISION for each pattern (numbered)

3 Results

The total number of initial patterns was 472,427, the extended version included more than 525,408 patterns and the reduced filtered final set included 655 patterns. Even though these patterns clearly do not include all possible mentions of functions in texts, it is also clear that they provide a good statistical base (32,641 sentences detected in a corpus of 36,543 sentences) for screening the sentences to search for protein names.

To assess the relationship between patterns and names, we compared the frequency of the patterns in two sets of sentences, one containing and one not containing gene names. A highly relevant number of patterns appears more frequently in sentences containing names (324 of the 518 patterns). A subset of these, 202, appears only in sentences where a gene or protein name is present. This subset is an ideal candidate for enhancing the discriminative power of gene/protein detection systems.

The permanent extraction patterns displayed in general a very high precision (see Figure 5), but the recall for an individual pattern was relatively low. Nevertheless, most of the patterns were matched to the validation sentences (79.08%) and 13,799 sentences of the gene indexed validation set had at least one pattern.

Patterns	P	R	F-score
protein is required	1	5.73E-05	1.15E-04
was localized on	1	5.73E-05	1.15E-04
gene is essential for	1	1.15E-04	2.29E-04
located in the	0.93	3.26E-03	6.51E-03

Table 2: Sample of high precision patterns.

In table 2 some of the top scoring precision patterns can be seen. Most of them contained a verb as trigger word, in contrast to the lower scoring precision patterns (table 3), which often corresponded to patterns were the trigger word corresponded to a noun.

Patterns	P	R	F-score
the human	0.544	0.015	0.029
, acts as	0.5	5.72E-05	0.0001
is associated with	0.494	0.005	0.010
role for	0.463	0.003	0.006

Table 3: Sample of low precision patterns.

Moreover most of the high scoring patterns had a difference between the random and the empirical average offset greater then 8, while in cases of low scoring precision if was mainly be-

low 2.5. Therefore, the use of offset calculation as a filtering step to extract co-occurrences of gene-indexed sentences is seen as promising. For sample sentences of true pattern-matching cases, see:

(a) *Although generally involved with detoxification, overexpression of one family member, cytochrome P450 1B1 (***CYP1B1***), has been associated with human epithelial tumors* [PMID:12813131]

(b) *For example, we have identified a novel gene called mta1 (rat) or* **MTA1** *(human) that appears to be involved in mammary cell motility and growth regulation* [PMID:9891220]

(c) **PEX13** *protein has an SH3 docking site that binds to the PTS-1 receptor* [PMID:11405337]

(d) *We have previously reported the identification of human* **PEX13**, *the gene encoding the docking factor for the PTS1 receptor, or PEX5 protein* [PMID:9878256]

The above examples show correctly identified gene containing sentences using extraction patterns. The extraction patterns are underlined, while the relevant gene symbols are displayed in bold.

After detailed analysis of the positive matched patterns, we found that certain patterns were more suited to annotating functional implications of disease-related features of genes or proteins (see example a). Other patterns were more suited to extracting descriptions of the participation of proteins in distinct biological processes (example b). In addition, functional descriptions and protein-protein interaction information, useful for deriving functional annotation data and protein definitions, were associated with certain patterns (see c and d).

4 Conclusions

We have described here a new approach for the identification of sentences containing information relating to gene and protein names in biological literature. Our proposal is based on the detection of sentences that contain information relating to protein (or gene) function as an indicator of the presence of protein/gene names.

To identify these sentences, we used a pattern-based approach that encapsulates the characteristic ways in which function is described in text. To generate the set of patterns describing functions, we started with an initial set of manually derived patterns, which was extended to cover a number of lexical variations. This larger set was filtered by matching of the patterns using previously gene-indexed sentences. The trigger word extension idea is based on the proposal by (Riloff, 1993; Agichtein and Gravano, 2000). Among the extraction patterns with high precision, a significant number contained verbs as trigger words. This corroborates previous studies that used verbs to extract biological interactions (Hatzivassiloglou and Weng, 2002; Sekimizu et al., 1998).

We plan to analyze further the patterns used in the study in order to explore the differential behavior of verb-containing patterns and noun-containing patterns for protein annotation extractions. The use of verbs to trigger extraction patterns for biological interactions have already been explored (Hatzivassiloglou and Weng, 2002; Sekimizu et al., 1998), but their use for protein indexing and annotation extraction was not previously studied in detail. The overall performance of extraction patterns for interactions and for annotation extraction seems to be similar.

Most of the extraction patterns used showed no dependency on the organism that was the source of the genes, with the exceptions of the patterns containing the trigger words *human*, *yeast*, *mammalian* and *mouse*. Therefore, the majority of the extraction patterns could be used for extraction of genes from a broad range of organisms, and especially aid in disambiguation of fly genes. As the extraction patterns can be applied without prior gene indexing, they could be used to enhance compound gene-name indexing, to extract rare typographical variants of existing gene names (not deposited in annotation databases) or even to mine the literature to discover new genes not yet described in current annotation databases.

The main focus in extraction patterns was precision, which was attained through a pipeline of filtering steps. The use of a larger set of initial trigger words might further increase recall in some cases.

We also plan to explore the use of the information of the patterns to improve the capacity of our current entity recognition systems. In particular, we would like to do this in the context of our system for detecting associations

between proteins and their functions. In the recent BioCreative challenge, it was clear that our system could be substantially improved by enhancing its name recognition capacity. This could be done by incorporating the frames as additional context information into the previously developed subset strategy (Krallinger and Padron, 2004).
Finally, we also plan to compare extraction patterns with automatically derived n-grams from previously gene-indexed sentences, in order to find which features are best suited for iterative bootstrapping to create new extraction patterns.

5 Acknowledgements

This research was sponsored by DOC, doctoral scholarship of the Austrian Academy of Sciences and the ORIEL (IST-2001-32688) and TEMBLOR (QLRT-2001-00015) projects. We are grateful to R. Hoffmann for providing the filtering set of gene-indexed sentences.

References

E. Agichtein and L. Gravano. 2000. Snowball: Extracting relations from large plain-text collections. *Proc. 5th ACM International Conference on Digital Libraries.*, pages 85–94.

C. Blaschke and A. Valencia. 2001. The potential use of SUISEKI as a protein interaction discovery tool. *Genome Inform Ser Workshop Genome Inform.*, 12:123–134.

C. Blaschke, A. Andrade, M, C. Ouzounis, and A. Valencia. 1999. Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol.*, pages 60–67.

V. Hatzivassiloglou and W. Weng. 2002. Learning anchor verbs for biological interaction patterns from published text articles. *Int J Med Inf.*, 67:19–32.

R. Hoffmann and A. Valencia. 2003. Life cycles of successful genes. *Trends Genet*, 19:79–81.

M. Krallinger and M. Padron. 2004. Prediction of GO annotation by combining entity specific sentence sliding window profiles. *Proc. BioCreative Challenge Evaluation Workshop 2004.*

M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman. 2000. Using BLAST for identifying gene and protein names in journal articles. *Gene*, 259:245–252.

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

E. Riloff. 1993. Automatically Constructing a Dictionary for Information Extraction Tasks. *Proceedings of the Eleventh National Conference on Artificial Intelligence.*, pages 811–816.

T. Sekimizu, H.S. Park, and J. Tsujii. 1998. Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts. *Genome Inform Ser Workshop Genome Inform.*, 9:62–71.

L. Tanabe and W.J. Wilbur. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18:1124–1132.

H. Yu, V. Hatzivassiloglou, C. Friedman, A. Rzhetsky, and W.J. Wilbur. 2002. Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles. *Proc AMIA Symp.*, pages 919–23.