ProMiner: Organism-specific protein name detection using approximate string matching

Daniel Hanisch[1], Katrin Fundel[2], Heinz-Theodor Mevissen[1],
Ralf Zimmer[2], Juliane Fluck[1]

[1]Fraunhofer Institute SCAI, Schloss Birlinghoven, D-53754 Sankt Augustin, Germany

[2]Institute for Informatics, Ludwig-Maximilians-Universität München,
Amalienstrasse 17, 80333 München, Germany

Recognition of gene and protein names in biomedical text owes its complexity to the inconsistent, highly variable and ambiguous nomenclature prevalent for many organisms. These problems become especially prominent when assignment of detected occurrences to major protein or gene databases is called for. The ProMiner system avoids prior tagging of potential name occurrences, but rather aims at direct detection of synonym occurrences in the text. It follows a rule-based approach and its search algorithm is geared towards recognition of multi-word names [3]. To account for peculiarities of the considered organisms, detection procedures for highly unspecific, ambiguous and case-sensitive synonyms have been implemented in the ProMiner system. In addition, further strategies for match disambiguation and organism specificity have been incorporated. The system has been applied to the test cases of the BioCreAtIvE competition with highly encouraging results.

The ProMiner system

The system consists of two essential steps, namely (1) generation and curation of a synonym dictionary and (2) accurate object detection in a given text corpus. Based on the procedure introduced in [3], customizations and extensions have been incorporated into the system.

Dictionary generation The quality of the dictionary is an important part of our approach. The dictionaries for *mouse* and *yeast* were created on the basis of the lists provided in the competition. The *fly* dictionary was obtained directly from the FlyBase database and entries were limited to *D. melanogaster* genes. All organism-specific dictionaries are processed in order to expand their scope and remove unspecific and inappropriate synonyms based on the curation procedure established in [3]. Manual corrections and external dictionaries for ambiguity detection and resolution were incorporated on the basis of the provided training sets.
In yeast, the only modification was the addition of the letter 'p' to each gene name. The mouse dictionary employed for the competition was cleaned stringently. As it was also used in conjunction with a simpler matching procedure without approximate search and no disambiguation capabilities [2], it was geared towards solving the problem of name variability via generation of a comprehensive dictionary including feasible spelling variants and removal of unspecific synonyms[1]. The curation of the fly dictionary was more lenient, as the incorporation of questionable synonyms into the search process for latter disambiguation is available in the ProMiner system.

[1]A more detailed discussion of the impact of curation is given in the workshop article of group 24 [2].

Rule-based classification of synonyms For each cleaned dictionary, a partition of the set of synonyms into three classes is computed. Class I contains all regular, i.e. probably specific, synonyms. Synonyms in this class are detected in the text through use of our approximate match procedure in a case-insensitive manner. Class II consists of case-sensitive synonyms. Case-sensitivity is needed if two unequal synonyms of different biological objects share the same normalized form. Class III consists of all questionable synonyms, i.e. synonyms which lead to highly unspecific matches when detected using standard string matching procedures.
To decide class III membership for each synonym, frequencies of all stemmed words in the Medline database are computed using the well-known Porter stemmer [4]. Frequently occurring synonyms are then assigned to class III. The thresholds were estimated based on the training set for the fly organism and remained unchanged for the other organisms. Rules detecting synonyms resembling sequence parts, numbers and subunit tags (e.g. *alpha 1*) are additionally applied. Synonyms in class III need to be augmented with specific context words (e.g. protein, gene, transcripts, etc.) in order to be detected accurately.

Match disambiguation For match disambiguation, we expand the dictionary temporarily using a controlled vocabulary of cellular processes, fly body parts, cell types[2] and an abbreviation dictionary. The abbreviation dictionary is compiled from two sources. First, the Biomedical Abbreviation Server [1] is queried for short uppercase synonyms present in the employed dictionaries. Furthermore, we extracted putative abbreviations from all test and training abstracts provided for task 1b. Secondly, we consider all short expressions in parentheses as possible abbreviations of long forms mentioned directly to the left of the occurrence. An abbreviation is accepted, if for each letter in the abbreviation a corresponding word in the preceding expression is found. Finally, the long forms of all abbreviations are checked against each dictionary in turn in order to prune long forms which correspond to protein names of the currently considered organism.
In case of conflicting matches, only the best scoring matches are retained. In conjunction with the controlled vocabulary, this leads to removal of unspecific synonyms[3]. In a next step, ambiguities are resolved in favor of objects for which most evidence, i.e. additionally detected synonyms, is present. As disambiguation to a unique object may not be possible in all cases, a threshold defining an acceptable size for the result set has been varied in the three submitted runs of the competition.
Additionally, to account for required organism specificity of matches, an organism taxonomy was obtained based on the NCBI taxonomy[4]. Organism name occurrences were detected using exact matching in the given text corpus. The basic idea is to reject matches in abstracts which only mention organisms not under investigation or their generalizations in the taxonomy.

Analysis of results

Based on this customized system, we computed three search runs for each organisms which differed with respect to the intended specificity-sensitivity tradeoff (cf. Table 1). The values of three parameters were changed:

- *Disambiguation threshold (D#):* Limits the number of objects relying on the same ambiguous synonyms occurrence. If the threshold is exceeded, none of the putative object occurrence will be reported.
- *Use of organism filter (O+/O-):* The use of the organism filter should increase specificity of matches. However, performance did not improve for the mouse organism based on the training set.

[2]For example, it may be desirable to distinguish the CD40 gene from the CD40+ cell type. This distinction is debatable, though, as the name of the cell type indeed stems from the CD40 gene. Distinction seemed to be required for BioCreative competition as judged from the training sets.
[3]For example, the gene *furrow* is often found as a substring match of the term *morphogenetic furrow* describing a fly body part.
[4]NCBI Taxonomy database, http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html

	Fly			Mouse			Yeast		
	1	2	3	1	2	3	1	2	3
Disambiguation (D#)	D3	D1	D3	D3	D1	D5	D3	D1	D3
Organism O(+/-)	O+	O+	O+	O-	O+	O-	O-	O-	O-
Dash significance S(+/-)	S-	S+	S+	S-	S+	S+	S-	S-	S-
F-measure	0,781	0,816	0,787	0,771	0,776	0,79	0,897	0,899	0,897
Specificity	0,728	0,831	0,744	0,752	0,809	0,766	0,951	0,966	0,951
Sensitivity	0,841	0,8	0,834	0,79	0,746	0,814	0,848	0,84	0,848

Table 1: Summary of submitted search runs. The table contains details of parameter sets (D# denotes disambiguation threshold, O(+/-) use of organism disambiguation, S(+/-) significance of dash at end of synonym) and resulting performance. Values where ProMiner achieved the best rank of all participants are marked.

- *Significance of '-' at the end of a synonym (S+/S-):* In the training set, erroneous matches were caused by matches such as '*protein* - induced'. While it is debateable whether such matches should be considered correct, we implemented a mechanism which disallows matches ending in a dash. Here, S+ implies that a dash is significant and S- denotes that a dash can be ignored.

The results obtained with the ProMiner approach are summarized in Table 1. Overall, our approach received highly encouraging results. For the mouse organism as well as for the fly organism, the best F-measures of all participants were obtained. The yeast results are also respectable given that no multi-word terms have been incorporated into the dictionary. The published gold standard allows to determine the impact of the various components of the ProMiner approach. A more detailed analysis of results with respect to each organism is presented in the following.

Fly For fly, all three submitted ProMiner runs outperform the results of the other participants. The best result was achieved setting the disambiguation threshold to one (D1), enabling the organism filter (O+) and treating a dash as significant (S+).
The fly organism probably poses the most severe problems to name detection as common English words occur frequently as parts of protein/gene names. The left part of Figure 1 shows results for different parameter settings which have been computed after the gold standard for the test set was released. Naïve string search with the provided dictionary is infeasible for fly and will result in unacceptable specificity (approximately 0.06, data not shown). The detection of questionable synonyms will alleviate this problem. A search based on the ProMiner core algorithm using only non-questionable synonyms and *no further disambiguation* leads to high sensitivity but low specificity ('approximate search, no questionables included'). Using the maximal disambiguation strategy with a threshold of one already provides acceptable performance (F-measure: 0.762, '+ disambiguation to 1'), though the sensitivity is decreased by 8%. The incorporation of questionable synonyms[5] improves sensitivity considerably (F-measure: 0.787, '+ add questionables'). The additional detection of case-sensitive genes improves sensitivity further, but impacts specificity (F-measure: 0.789, '+ case sensitivity'). An increase in sensitivity occurs as previously ambiguous occurrences can now be assigned to one synonym with the correct case. However, also false positives occur in this set which decreases specificity. The organism disambiguation for the fly organism seems to work well as only subspecies differentiation was needed in the test set (F-measure: 0.80, '+ organism disambiguation'). A final gain in specificity is obtained through use of the controlled vocabulary (e.g. acronym dictionary) for disambiguation purposes to result in the submitted run with maximum F-measure (F-measure: 0.816, '+ controlled vocabulary (best search, submitted)').

[5]Incorporation entails expansion of synonyms with high specificity words and using the questionable synonyms in their original form for disambiguation purposes only.

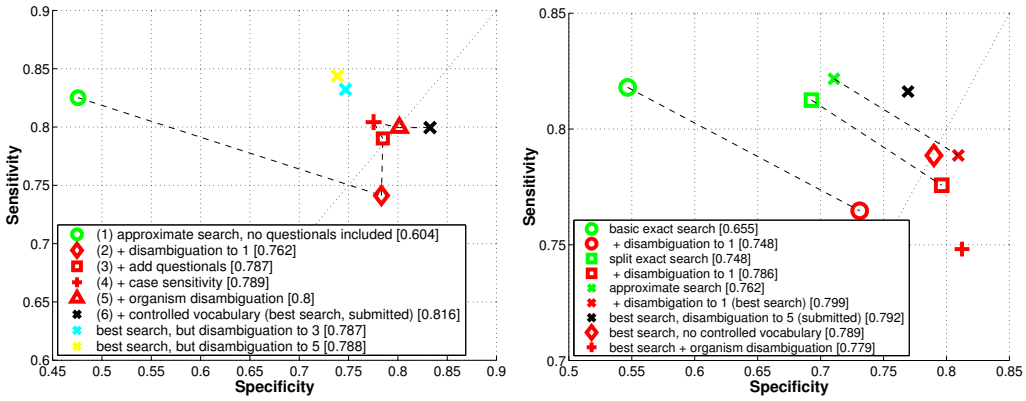

Figure 1: Impact of ProMiner components for fly (left figure) and mouse (right figure). For each result, the F-measure as determined from the published gold-standards is given in brackets.

For comparison, the disambiguation threshold was set to three and five allowed objects, respectively (F-measure: 0.787/0.788, 'best search, but disambiguation to 3/5'). Both settings negatively impact specificity and sensitivity. A-priori this was unclear as the training set contained examples, where (1) different allelic variations of the same gene or (2) members of a complex were annotated as true positives. These cases profit from a higher disambiguation threshold. In the test set, however, disambiguation to one provided the best results.
Concluding, all components of the ProMiner framework improve overall performance to various degrees. The single, most important step for the fly organism is the detection and sensible incorporation of questionable synonyms.

Yeast The results for yeast exhibit good performance for submitted runs of most participants reflecting the quite stringent terminology followed for that organism. In this setting, simpler name recognition approaches seem to suffice to obtain satisfactory results. Probably, the incorporation of multi-word protein names into the synonym list would improve sensitivity for our approach. Such improvements should be pursued in future work.

Mouse The terminology for mouse is not dominated by common word names as for fly and neither as stringent as in the case of yeast. Multi-word protein names are frequent, but may not have played a dominant role in the BioCreative test set. In general, overall results of all participants are better than for the fly organism but worse than for yeast. The ProMiner submissions achieve best F-measures.
As in the case of fly results, a more detailed analysis of the ProMiner approach is given in the right part of Figure 1. All runs were computed without organism disambiguation, if not stated otherwise. For mouse, an exact, case-insensitive search and no further disambiguation is feasible, though results are mediocre (F-measure: 0.655, 'basic exact search'). Disambiguation to one allowed object impacts sensitivity, but increases specificity to acceptable values (F-measure: 0.748, 'basic exact search + disambiguation to 1'). The incorporation of case-sensitive and questionable synonyms improves both, specificity and sensitivity (F-measure: 0.786, 'split exact search (+ disambiguation to 1)'). A final gain of overall performance stems from the approximate search (F-measure: 0.799, 'approximate search (+ disambiguation to 1)'). This optimal search run was not submitted, however, because the impact of the organism disambiguation procedure was unclear from the training set. Using the same search parameters including organism disambiguation leads to worse sensitivity (F-measure: 0.779, 'best search + organism disambiguation'). An ex-post analysis of the organism disambiguation performance revealed that (1) the disambiguation failed in more complicated cases where decisions for each gene instead of the complete

abstract were required, (2) disambiguation failed because of missing synonyms, e.g. "vertebrate" and (3) for several cases the provided gold standard might be incorrect as considered abstracts describe findings in *rat* or *human* instead of mouse.

Description	Examples
Unspecific synonym	growth retarded, perinatal lethality, long lived
Wrong context	TGF-beta superfamily, c-myc tumors
Unknown ambiguity	high dose set at MTD or MFD
Doubtful gold-standard	interleukin-2 , H-2 locus, c-Jun

Table 2: Sample of false positive matches in the best submitted mouse search run. Detected matches are marked .

The first thirty false positive matches of the best submitted results for mouse (run number 3) have been analyzed in more detail. Undetected ambiguities are the dominant reasons for false positive matches (60%). This number includes unspecific synonyms which have neither been detected during curation nor marked as questionable synonyms, detection of synonyms in wrong contexts, and unknown external ambiguities. Detection of genes from other organisms accounts for 13.3 % of false matches. In eight cases, the reason for exclusion from the gold-standard remained unclear.
Concluding, the ProMiner method provides high sensitivity using approximate matching and retains high specificity due to sensible incorporation of questionable synonyms.

Conclusions

The named entity recognition task of the BioCreative challenge was ideally suited for independent evaluation of the ProMiner method. The task required the customization of the framework to the organisms of fly, mouse and yeast, each of which exhibits specific naming characteristics. Using parameter settings and customized dictionary curation, the ProMiner method could be quickly adapted to the characteristics of each organism. The results of the competition were highly encouraging and approximate direct matching in conjunction with appropriate pre- and post-processing seems well suited for named entity recognition.
In the BioCreative workshop, other participants will present their approaches which may be based on other principles than ProMiner, e.g. prior tagging of putative name occurrences and subsequent assignment to a dictionary. It remains to be seen which approach will dominate in the long run. Currently, the direct approach as realized in the ProMiner framework works best.

References

[1] J.T. Chang, H. Schtze, and R.B. Altman. Creating an online dictionary of abbreviations from medline. *The Journal of the American Medical Informatics Association*, 9(6):612 – 620, 2002.

[2] K. Fundel, D. Guettler, R. Zimmer, and J. Apostolakis. Exact versus approximate string matching for protein name identification. In *Proceedings of the BioCreative Challenge Evaluation Workshop 2004*, 2004.

[3] D. Hanisch, J. Fluck, H. T. Mevissen, and R. Zimmer. Playing biology's name game: identifying protein names in scientific text. *Pacific Symposium on Biocomputing*, pages 403–14, 2003.

[4] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.