Exact versus approximate string matching for protein name identification

Katrin Fundel, Daniel Güttler, Ralf Zimmer, Joannis Apostolakis
Institut für Informatik, Ludwig-Maximilians-Universität München,
Amalienstr. 17, 80333 München, Germany

Abstract

We present a simple and efficient tool for exact matching of terms in a synonymlist against medline abstracts. It does not recognize spellings of a synonym which are not in the synonymlist and does not consider context for matching. Its main application is to test different synonymlists and to evaluate different kinds of expansions of synonymlists performed during curation. This tool allows us to rapidly evaluate modifications of synonyms and enables us to build high-quality synonymlists. These can then also be used as a prerequisite for text-mining with other text-mining programs. Additionally, we used a simple post filter in order to improve specificity of our results.
Our goal in participating at the BioCreative-contest was to assess the sensitivity and specificity that can be achieved with extensively curated synonymlists and basically naive exact string-matching, and to assess the difference to more sophisticated text-mining approaches.
We participated as group 24 in Task 1b for yeast and mouse. We did not submit results for fly because of the significant overlap of fly protein names with common english words for which our approach is not adapted. Our mouse synonymlist was also used by group 16 with a more sophisticated search algorithm implemented in the tool ProMiner[1, 2]. This contest allows us to compare the two approaches on a blind prediction basis and for an independent test set.

Methods

Generation and curation of synonymlists The performance of a textmining-procedure like ours depends heavily on the quality and completeness of the synonymlist used for searching. The synonymlists for mouse and yeast were created on the basis of the lists provided by the BioCreative organisers. We curated the provided lists to cover additional, frequently used synonyms and remove unspecific and inappropriate synonyms. In a first step synonyms consisting solely of digits and/or special characters and synonyms of length less than two are removed. Subtype specifiers are expanded to equivalent other specifiers (a⇔alpha). Special characters at the beginning or end of a synonym are removed and different spelling variants like the insertion of a hyphen or space between alphabetic characters and digits are added (Igf 1⇔Igf-1⇔Igf1). Eventually, organism specific synonyms are added (e.g. yeast synonyms are often mentioned with extension 'p': SOH6→SOH6p). In a second step, synonyms matching common english words are removed. Synonyms containing subtype specifiers are expanded by the synonym without subtype specifier if there is only one subtype mentioned in the synonymlist (aminoacylase 1 → aminoacylase).
The third curation step consists of several expansion and pruning steps. The tool used for this step was kindly provided by D. Hanisch, for a detailed description see [2]. In the expansion phase, new synonyms are added to the existing ones, e.g. common acronyms are expanded to long names and long names are reduced to acronyms (IL⇔interleukin). Inappropriate synonyms are detected and removed in the pruning phase by using token-class based regular expressions. A token can be any sequence of letters and/or numbers. We define token classes as groups of words which have a similar meaning

or usage. Examples of token classes are: measuring units (contains: kDa, Da, mg,...), common words (if, and, as, for, ...), descriptions (tRNA, Ser, Tyr,...), numbers, single letters. These token classes are combined in regular expressions which allow to filter out all synonyms of a certain composition, e.g. synonyms consisting of a number and a measuring unit like '22kDa' or synonyms consisting of a common word and a number like 'If 1'. The lists and rules used for curation were created by detailed analysis of synonyms provided in Swissprot[1] and HUGO[2].
The results on the provided hand-curated training set were analysed manually, prominent false synonyms were removed and missing synonyms were added. Ambiguous synonyms (i.e. synonyms belonging to more than one protein) generally need to be assigned or disambiguated to one of the corresponding proteins. Our approach does no disambiguation, therefore ambiguous synonyms are removed from the synonymlist. Objects which have no synonym left are removed from the synonymlist. The curated synonymlists are significantly larger than the original ones. The curated mouse synonymlist contains 51981 objects and 394455 synonyms (7.6 synonyms per object in average) compared to the uncurated list with 52594 objects and 130358 synonyms (2.5 synonyms per object). The curated yeast synonymlist contains 7906 objects and 40725 synonyms (5.2 synoyms per object) compared to 7928 objects and 14715 synonyms in the original list (1.86 synonyms per object).

Match detection Synonyms as defined in the synonymlist are searched within the texts by exact text matching. The search is case insensitive only if the synonym length is above a certain threshold (5 characters) or if the synonym contains numbers. When several synonyms of different length can be matched at a certain text position, only the longest match is reported.

Post filter We implemented a simple post-filter that checks occurrences of synonyms for nearby appearance of modifiers (e.g. 'cells', 'domains', 'cell type', 'DNA binding site') indicating that a protein is not meantioned. Short synonyms in parentheses often overlap with definitions of abbreviations differing from the assumed protein. We clarify the meaning of these occurrences by checking the words right ahead of the parentheses. If no significant overlap of these words with the alternative names of the assumed protein is found the match is discarded.

Results and Discussion

General Performance Our system achieved good performance in the BioCreative Assessment. The results (shown in table 1) were evaluated in terms of:

$$Precision = \frac{TP}{TP+FP} \quad Recall = \frac{TP}{TP+FN} \quad F-measure = \frac{2*Precision*Recall}{Precision+Recall}$$

For mouse, we submitted two runs: one without any post-filtering (run 1) and one with the post filter described above (run 2). For yeast we submitted one run without post-filtering. For both organisms our tool achieved results close to the best overall results. For mouse, the best result is a run done with ProMiner, the difference to our results in F-measure is 0.026/0.017. For yeast the difference to the best result is 0.024. This difference is mainly due to higher precision (0.033), but also recall is somewhat higher (0.016). The results show that a straightforward approach for protein name recognition can be successful. The analysis of the results shows that exact matching results in good sensitivity and specificity, although further improvements are possible. Simple matching leads to results that are only marginally worse than the best methods available.
BioCreative also shows the different levels of difficulty for protein name recognition for different organisms, ranging from yeast with a quite precise nomenclature consisting mainly of distinctive single word synonyms over mouse having multi word protein names to fly with a large number of synonyms overlapping with standard english words and anatomic descriptions. The examples listed below are taken from mouse, errors of the yeast results resemble and are not discussed in detail.

[1] http://www.expasy.org/sprot/sprot-top.html
[2] http://www.gene.ucl.ac.uk/nomenclature/

	Mouse (1)	Mouse (2)	Yeast	Mouse max.	Yeast max.
F-measure	0.764	0.773	0.897	0.79	0.921
Precision	0.735	0.764	0.917	0.766	0.95
Recall	0.796	0.781	0.878	0.814	0.894
TP	433	425	538	443	548
FP	156	131	49	135	29
FN	111	119	75	101	65

Table 1: Results in BioCreative Task 1b: Our results and results for mouse and yeast with highest overall F-measure. For mouse the highest F-measure is achieved by ProMiner. Mouse(1) is the exact search with the curated list. Mouse(2) was additionally post-filtered.

Curation of synonymlist Figure 1 shows the impact of the different curation steps. It shows that already an exact search with a list returned from step 2 of our curation procedure yields results which are comparable to those of other groups. Our submitted results were generated by applying all three curation steps. The additional execution of the third step of curation, namely the removal of synonyms based on regular expressions of inappropriate tokens and the expansion of acronyms and long names yields a further increase in recall and precision.

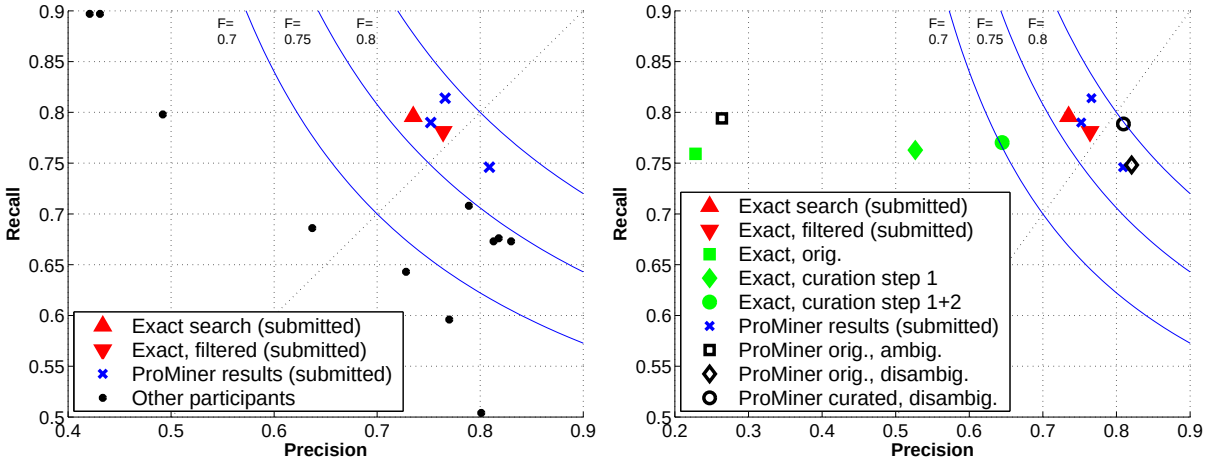

Figure 1: Left figure: BioCreative results for mouse. Right figure: Impact of curation on exact search and ProMiner approach. The submitted runs (Exact search (submitted)) were done with the fully curated synonym-list. (Exact, filtered(submitted)) was additionally post-filtered. (Exact orig.) is the exact search of the original, uncurated synonym list. For details on ProMiner results see section 'Comparison to ProMiner' and [1].

False positives The false positive matches can be classified in different categories. Some examples are listed in tables 2 and 3. All false positive matches are in principle correct matches of a valid synonym of the corresponding protein. They appear as false positives because the occurrency does not refer to the protein that was assumed to be mentioned. In all cases, the context reveals the intended meaning of the expression. The post filter removes several false positive matches and so slightly increases specificity.

Type of error	Examples
overlap with english words	*striated* muscle, *killer* cells, *Low* effectiveness ...
wrong organism	Mutations in the human *doublecortin*
no direct mention of protein	... inhibits *BMP2*-mediated induction of ...
description of different object	... with the *androgen receptor* antagonist cyproterone acetate ...
synonym has different meanings	... transgenic mice are *growth retarded*,
	... is required for normal *cardiac morphogenesis*

Table 2: Samples of false positive matches of exact search, synonyms are marked in *italics*.

Synonym	Context	Other synonym for wrongly identified object	Removed by post-filter
P21	chromosome 2p16-p21	cyclin-dependent kinase inhibitor 1A (P21)	no
FACS	fluorescence-activated cell sorter (FACS)	fatty acid Coenzyme A ligase, long chain 2	yes
PCR1	E. coli plasmid pCR1	mannosidase 1, alpha	no
CA1	area CA1 of the hippocampus	carbonic anhydrase 1	no
HEK	HEK cells	Eph receptor A3	yes
NT2	NTera 2(NT2) cell line	zinc finger protein 263	yes
Eph	Eph family of receptors	Eph receptor A1	no
PMN	polymorphonuclear (PMN) infiltration	progressive motor neuropathy	yes
all-trans	All-trans retinoic acid	retinol dehydrogenase 2	no
s1p	sphingosine 1-phosphate receptor genes	site-1 protease	no
Den	diethylnitrosamine (DEN)	denuded	yes

Table 3: Samples of false positive matches of exact search, mostly short names and abbreviations of protein names which have different meanings, and the effect of post filtering on these matches.

False negatives The false negative matches can be classified into three groups (see table 4 for some examples): missing synonyms, different spellings of synonyms, and ambiguous synonyms, which cannot be disambiguated. Sensitivity could be increased by covering more spelling variants. Some of the false negatives can be recovered by quite easy means such as equal treatment of space and hyphen. Another straightforward improvement would be a further extension of subtype descriptors (e.g. alpha, a, I, 1). Inversions are more difficult to deal with as they are not always allowed. In some cases proteins are mentioned by expressions which have no clear relation to any of the given synonyms. These cases are more difficult to handle. The analysis of the false negative matches of yeast showed that long names of several proteins were used in abstracts while the synonymlist contained only the corresponding short names. Some of these long names could have been extracted from Swissprot. By considering further sources (e.g. general databases like Swissprot or organism specific databases) for the generation of synonymlists it may be possible to cover further synonyms.

Synonym(s)	Occurrence in text	Type of error
Lpa1, Lpa2, Lpa3	lpa(1-3)	enumeration
Pkcb, Pkce	PKC beta, PCK-epsilon	different spelling
retinoic acid receptor, alpha	retinoic acid receptor-alpha	different spelling
interferon gamma	gamma-interferon	inversion
Braf2, Braf-rs1	Braf	ambiguity
peroxisome proliferator activated receptor gamma	peroxisome proliferating antigen receptor gamma	not evident

Table 4: Samples of false negative matches

Post filter The post filter increases precision by 2.9% and decreases recall by 1.5%. This shows that in principle the approach is correct but also shows its limits. Further enhancement is clearly possible. The rules applied for filtering out false positives can easily be extended and improved, this would cover some of the false positives not yet detected. There is no organism filter included yet which would be a further possibility of enhancement.

Comparison to ProMiner Our results show that especially for organisms having a stringent terminology, such as yeast, exact text matching is useful and reasonable for protein name recognition. For such organisms an approximate search like the algorithm applied in ProMiner[1, 2] does not improve the results significantly. The results for mouse show that for organisms with a more difficult terminology there is a slight difference in performance between exact text matching and approximate search. Considering the best submitted results of both approaches (those yielding highest F-measure) precision is similar but recall is higher for approximate search. Keeping in mind the approximate matching procedure of ProMiner, this is obvious. The difference is probably even more pronounced for organisms with a protein nomenclature consisting of many multi-word synonyms.

The result of the basic ProMiner approach with the uncurated synonym list and no disambiguation (Figure 1, ProMiner orig., ambig.)) is slightly better than the results of exact matching. This is due to approximate matching and the internal scoring function which is based on token classes and eliminates poor matches. The full ProMiner framework includes extensive filtering and disambiguation. With optimal parameter setting this system shows impressive results even when using the uncurated synonym list (F-measure 0.78, (ProMiner orig., disambig.)). The parameters used for this run were acquired during post evaluation and turned out to yield better results than the parameters used for BioCreative submissions. By using the curated synonym list with the same settings (ProMiner curated, disambig.) the F-measure increases further to 0.80. This shows that also for an approximate and advanced approach like ProMiner the curation of the synonymlist has a significant effect on the search result.
There are important advantages of the exact matching procedure: It is easy and fast to set up and run as it does not need any parameter optimisation. As the curation of the synonymlist is independent of the search, the curated list can be further manually curated after automatic curation. This is useful if the search result of a training set indicates bad synonyms in the synonymlist. The runtime of the curation procedure depends largely on the size and characteristics of the synonymlist, it is about 2 minutes for yeast and about 12 minutes for mouse. The exact search for the yeast training set of 5000 abstracts including analysis and report of results takes about 45 seconds on a standard machine. The exact search script is implemented in Perl, it has less than 750 lines of code and is easy to adapt to different input and output formats.
ProMiner is less dependent on the curation of the synonymlist and is capable of synonym disambiguation, but it is more difficult to set up and handle. The system needs adjustment of different matching parameters which have a significant effect on the results. ProMiner has a longer runtime for small data sets as it preprocesses the synonymlist. The preprocessing runtime depends largely on the size of the synonymlist. It needs about 90 seconds for preprocessing of the yeast synonymlist and 3.5 minutes for the search, filtering and report of results when applied on the corresponding training set.

Conclusions

With our system we showed that it is possible to achieve good performance in protein name recognition with exact text matching. Our system does not need to be adapted for a specific synonymlist in terms of parameter tuning or internal lists. This allows for straightforward application.
It is crucial for our attempt to use synonymlists which are as complete and correct as possible. Therefore, we used a system for the extensive curation of protein synonym lists. This curation is largely independent of the synonymlist to be curated as the curation steps are of general character. Nevertheless, the system can be adapted easily to cover specific problems of synonymlists, like missing synonyms which are frequently used in texts but are present in the synonymlist only with slight differences in spelling. One disadvantage of the extensive curation is the fact that the synonym lists become very large as they need to cover all possible different spellings of a protein name. In order to avoid this, one could consider making the text search more flexible, e.g. by including certain equivalent expressions directly in the search tool.

References

[1] D. Hanisch, K. Fundel, H.T. Mevissen, R. Zimmer, and J. Fluck. Prominer: Organism-specific protein name detection using approximate string matching. *Proceedings of the BioCreative Challenge Evaluation Workshop 2004*, 2004.

[2] 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*, 8:403–414, 2003.