Classification methods for finding articles describing protein-protein interactions in PubMed

Sérgio Matos*, José Luis Oliveira

University of Aveiro, DETI/IEETA, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal, http://bioinformatics.ua.pt

Summary

With the rapid expansion in the number of published papers in the biomedical field, finding relevant articles has become a demanding task for researchers. This has led to increasing interest in the use of text mining tools that help search the literature and identify the most relevant documents or information. One specific topic of interest is related to the identification of articles that might be used for extracting protein-protein interactions. Using the BioCreative III Article Classification Task dataset, composed of PubMed abstracts classified as relevant or non-relevant for describing protein-protein interactions, we compare different classification methods with different sets of features. The best results – area under the interpolated precision-recall curve of 0.654 – indicate that the proposed classification strategy could be incorporated in the database curation workflows in order to prioritize articles for extraction of protein-protein interactions. Furthermore, we also analysed the use of this method for ranking documents resulting from general PubMed queries, and propose that this approach could be useful for general researchers looking for publications describing protein-protein interactions within a particular topic of interest.

1 Introduction

An important phase in any scientific work refers to literature review. Researchers have to search the literature when preparing an experiment or project, and again when analysing their results. However, due to the rapid increase of the number of scientific articles, researchers often have to scan through a large number of publications in order to select the most relevant ones. In fact, since much more publications are available, finding articles satisfying a specific information need has become a very demanding task. Of particular interest is the field of biomedical sciences, where the MEDLINE literature database alone includes over 18 million citations, with over 2000 being added daily [1]. Although several information retrieval and text mining tools have been proposed to deal with specific demands in this area, these are not yet part of the usual information search routines of researchers [2, 3].

Another way in which scientific knowledge and data is accessible to researchers is through structured information, annotated in various biomedical databases. The emergence of these validated resources has helped alleviate the data explosion problem. An example is the important topic of Protein-Protein Interactions (PPIs), where over the past few years several publicly available databases have been published to collect and store high-quality, manually curated protein-protein interaction data. However, the existence of several data sources means that data may be replicated or distributed along different places, with possible inconsistencies between

*To whom correspondence should be addressed. Email: aleixomatos@ua.pt

different instances of the same information. Keeping these databases up to date and managing the spreading and replication of data are difficult tasks. Lehne and Schlitt [4] compared six main PPI databases and highlighted several problems such as overlapping and terminology inconsistencies. It is suggested that the adoption of a common ontology, such as the ones already existent for sequence and microarray data, will definitely increase the quality of PPI data.

From an information extraction perspective, the problems are also far from being solved. Considering that the primary sources for PPI data are scientific publications – which may still contain much more evidence of PPIs than what has been annotated in databases – a significant amount of manual work is continually needed to identify these interactions from those documents. Therefore, finding articles with that information is clearly of major importance both for database annotation tasks and for general biologists [5]. These factors have led to increased interest in the application of text mining and information retrieval methods in biomedical literature [5, 6]. The relevance of such methods is well documented in the literature, as well as the importance of including them in automated tools that can be used by researchers for identifying relevant documents or to accelerate the process of database annotation.

In this paper, we compare different combinations of classification models and features for prioritizing documents according to their relevance for extracting PPIs. We also describe a method for ranking MEDLINE abstracts resulting from any particular search topic, according to their likelihood of describing PPIs. The proposed method works as follows: firstly, the National Center for Biotechnology Information (NCBI) Entrez Programming Utilities (eUtils) [7] are used to search and retrieve abstracts for any general PubMed query; the returned articles are then scored in terms of their similarity to a reference corpus of PPI-relevant abstracts. Documents classified as relevant are returned to the user, ordered by relevance score, while the others are rejected. Some examples of general PubMed queries ranked by this method are explored. These experimental results show that acceptable performance can be achieved with this approach.

This paper is organized as follows: Section 2 presents a summary of related work, Section 3 describes the methods and resources used and Section 4 presents the results. The results and methodology are discussed in Section 5, and final conclusions are presented in Section 6.

2 Related Work

The use of document classification and ranking techniques for scientific publications has been described in many works. Suomela and Andrade [8] proposed a classification method based on word frequencies, which, given any two articles, decides which one is more related to a topic. The extracted keywords were restricted to words that commonly convey meaning, that is, nouns, verbs, and adjectives. The authors propose a classification and ranking model to evaluate the entire MEDLINE database with respect to a topic of interest. The method, which presents an f-score of 65%, relies on the different frequencies of discriminating words between the training set and other non-relevant articles on a reference set. This approach is also behind the MedlineRanker web-service (Fontaine *et al.* [9]), which allows to retrieve a list of articles ranked by similarity to a training set defined by the user. This training set can be obtained from a PubMed search, or from PubMed document identifiers associated with a given indexing term from the Medical Subject Headings (MeSH) vocabulary[1], for example. Another possibility, as

[1]National Library of Medicine Medical Subject Headings, `http://www.ncbi.nlm.nih.gov/mesh`

referred by the authors, is to use a list of document identifiers obtained from a PPI database, therefore getting as the result articles related to that topic.

Several other authors have tackled the problem of PPI extraction from documents. Jang *et al*. [10] present a method based on co-occurrences of protein names in the same sentence to validate known protein-protein interactions. They use PubMed queries to firstly collect a set of abstracts where two given protein names or any of their synonyms are present. They use a procedure for sentence simplification that normalizes protein names and noun phrases in order to increase the parsing accuracy. The system was validated with the BioCreative II (BC-II) and the Database of Interacting Proteins (DIP) PPI corpora, achieving a precision of 81% and a recall of 43% for the identification of PPIs. Yin *et al*. [11] focused on the special case of identifying research articles describing Host Pathogen Protein-Protein Interactions (HP-PPIs). In the described approach, documents are normalized by substituting lexical variants by their base forms, and nouns and adjectives by the corresponding verbs. They trained an Support Vector Machine (SVM) classifier, with uni- and bigram features on a training set of 1360 abstracts, achieving a positive predicted value (PPV) of 50%, for a recall rate of 51%.

Marcotte *et al*. [12] used a log likelihood scoring function to identify articles discussing PPIs. They report an accuracy of 77% for articles with a log likelihood score of 5, corresponding to a recall around 55%. The feature set of 83 discriminating words was selected from a training set of 260 MEDLINE abstracts involving yeast proteins. Lan *et al*. [13] compared the use of Bag-of-Words (BoW), interaction trigger words and protein Named Entities (NEs) features in a SVM classifier for identifying articles discussing PPIs. They tested the classifiers using the BC-II PPI data set, and reported a precision of 70% and a recall of 87% when using the BoW features. Their best result, when using a single classifier, was obtained with a feature set containing BoW features and protein NEs co-occurring with interaction trigger words (f-score of 77%). Abi-Haidar *et al*. [14] tested three classifiers in the BC-II PPI data set: SVM, singular value decomposition (SVD) and Variable Trigonometric Threshold (VTT). They reported a top f-score of 78% using the VTT classifier and a feature set of 650 discriminating words.

Retrieval and extraction of PPI related information has been a major focus of recent shared evaluations in the biomedical domain. On the recent BioCreative III Challenge (BC-III), the PPI Article Classification Task (ACT) counted with 52 submissions from ten participating teams [15]. The best AUC iP/R (area under the interpolated precision-recall curve) was 0.680 and the highest MCC (Matthew's correlation coefficient) was 0.553, with an accuracy of 89.2% and an F-score of 61.3%. Most teams applied some sort of machine learning technique, the best results being obtained using Support Vector Machines, Maximum Entropy or Large Margin classifiers. The top performing teams used various levels of lexical analysis, including Part-of-Speech (PoS) tagging and Named Entity Recognition (NER), and the best team overall also used dependency parsing to extract the textual features used for classification.

Although several methods exist, protein-protein interactions are highly under-studied. PPI Finder [16] is a web-based tool that uses a two-fold approach for protein-protein interaction. For a given human gene it finds related genes based on their co-occurrences in PubMed abstracts and then extracts the semantic descriptions of interactions words. A case study is presented, showing that only 28% of the co-occurred pairs in PubMed abstracts appeared in any of the

commonly used human PPI databases (HPRD[2], BioGRID[3] and BIND[4]).

3 Methods

The classification methods investigated in this work allow ranking a set of documents, giving higher relevance to articles that most likely describe protein-protein interactions. We compared a vector-space classification approach, implemented through a simple indexing strategy, to different probabilistic and machine learning based classifiers: Naïve Bayes (NB), Maximum Entropy (MaxEnt), and Support Vector Machines.

We then analysed the use of these classification methods to rank the results of a general PubMed query. The proposed approach uses the PubMed web services, processing the results and returning only those results that are classified as relevant, ordered in terms of the calculated relevance.

3.1 Resources

In order to compare the different methods, we use the dataset of PPI-relevant and non-relevant documents from the BioCreative III ACT task [15]. This corpus is composed of manually annotated MEDLINE abstracts, containing 2280 documents in the training set, 4000 in the development set, and 6000 in the test set. The training set has the same number of positive and negative examples, while the development and test sets are unbalanced (15-17% of positive examples), reflecting the expected real scenario.

Our approaches make use of biologically relevant words, occurring in the documents, as discriminating features. For this, we built a lexicon composed of domain-specific verbs, names and adjectives, compiled from the BioLexicon resource [17]. The lexicon links each verb form, name or adjective, to the corresponding lemma, usually the base (infinitive) form of the verb. This way, the BioLexicon terms can be normalized to the base form of the associated verb (for example, “interacts”, “interacting” and “interaction” can all be normalized to “interact”). The lexicon also includes a list of interaction method names extracted from the Proteomics Standards Initiative Molecular Interactions Ontology (PSI-MI Ontology) [18].

3.2 Vector-space classification

The vector-space classifier is based on a Lucene [19] index. However, instead of using the basic bag-of-words approach, the documents are represented as vectors of domain specific terms which compose our lexicon. Since this subset of the language is more representative of the domain than the complete language used in the texts, it is expected that classification models based on these features will be more robust. Furthermore, this significantly reduces the number of possible words used in the document vectors therefore improving efficiency.

[2]Human Protein Reference Database, `http://www.hprd.org/`

[3]Biological General Repository for Interaction Datasets, `http://thebiogrid.org/`

[4]Biomolecular Interaction Network Database, `http://bond.unleashedinformatics.com/`

Table 1: Structure of the index used for vector-space classification.

ID	Class	Terms	Lemmas
...			
17503969	1	interacts required mediated transcription repression complex gene expression development ...	interact require mediate transcribe repress complex gene express develop ...
19393007	0	limiting transport sensing lies transporter control concentration import sensing add transport ...	limit transport sense lie transport control concentrate import sense add transport ...
...			

During the index creation, each training document is analysed using a dictionary-matching approach, in order to locate the lexicon terms appearing in that document. Simple exact matching is used, since we assume that the dictionary is complete. For each term of the lexicon appearing in the document, we add it to the index, together with the corresponding base form or lemma. The use of lemmas allows normalizing related terms to a single lexical entry. For classification, either the lemmas or the textual occurrence of terms can be used to represent the documents. The class of each training document (1 for relevant, or 0 for non-relevant) is also stored in the index, to be used in the classification stage. This way, documents in the training set are represented in this index as vectors of lexicon words, to which the unclassified documents are compared. Table 1 shows the structure of the index used.

During the classification of a new document, each occurrence of a lexicon term (or alternatively its corresponding lemma) is added to a query string, which is then used to search the index. From this search, the top M documents are retrieved, together with the corresponding classifications and the Lucene similarity scores. The class probability for the new document is then calculated as the sum of the similarity scores for each class, normalized by the sum of the scores for the M documents. A threshold, corresponding to the operating point of the classifier, is then used to select the class for that document. Fig. 1 illustrates this process.

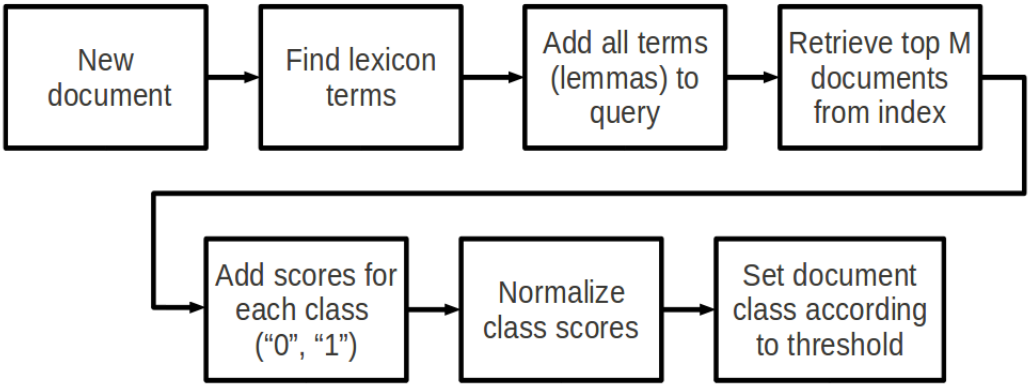

Figure 1: Vector-space classification

3.3 Probabilistic and Machine Learning methods

The Naïve Bayes and Maximum Entropy classifiers were implemented and trained using the Mallet toolkit [20]. The list of PubMed stopwords, available from the NCBI website, was added to Mallet's default list. The SVM classifier was implemented using the LIBSVM library [21]. We tested linear and radial basis function (RBF) kernels.

When training the different classifiers, we merged the BC-III ACT training and development sets and used cross-validation. The test set was used for final evaluation. Different strategies were tested, considering the use of different preprocessing of the corpus. Namely, we tested the use of the complete text of the documents (title plus abstract), the same preprocessing as done in the vector-space classification scheme, that is, using only the set of lexicon terms found in each document, and a third approach, using the complete text but substituting lexicon entries by their corresponding lemma. We also compared the use of unigrams and bigrams.

4 Results

For the vector-space classification approach, we first evaluated the number of documents to retrieve from the index for determining the class of the new document (M, see Fig. 1). The iP/R curves, based on cross-validation results using the merged training and development sets, for values of M between 50 and 500 are illustrated in Fig. 2. Although results are very similar, a value of M=500 returns higher precision for recall values above 50%, as can be observed in the graph1q. Using more documents did not significantly improve the classification results and, as expected, slows down the classification procedure.

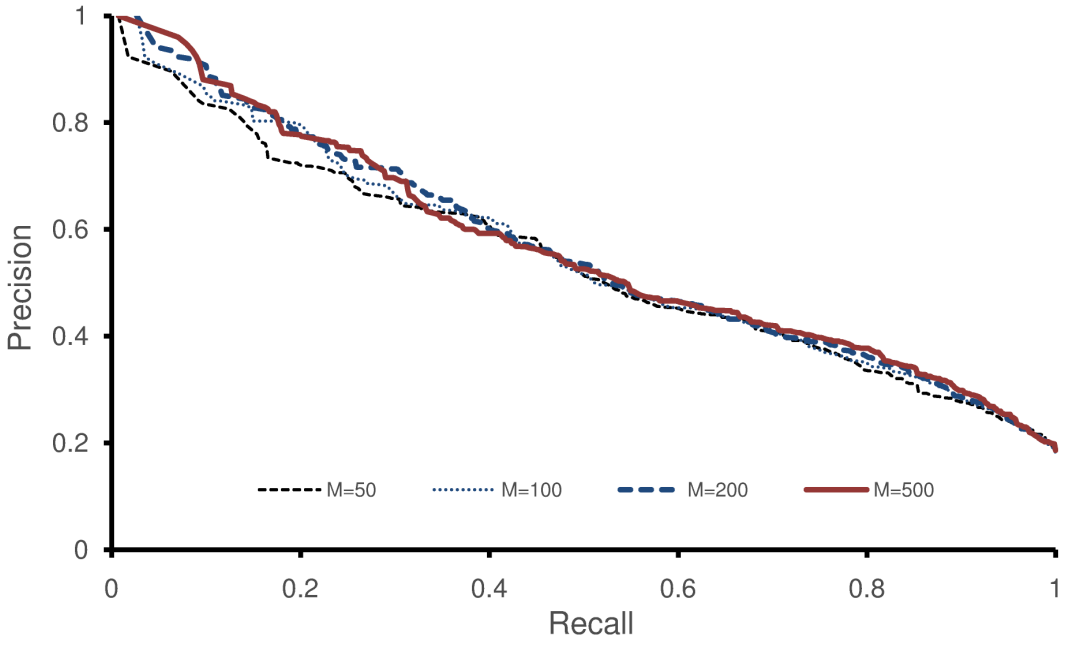

Figure 2: Vector-space classification: impact of the number of documents used for classification

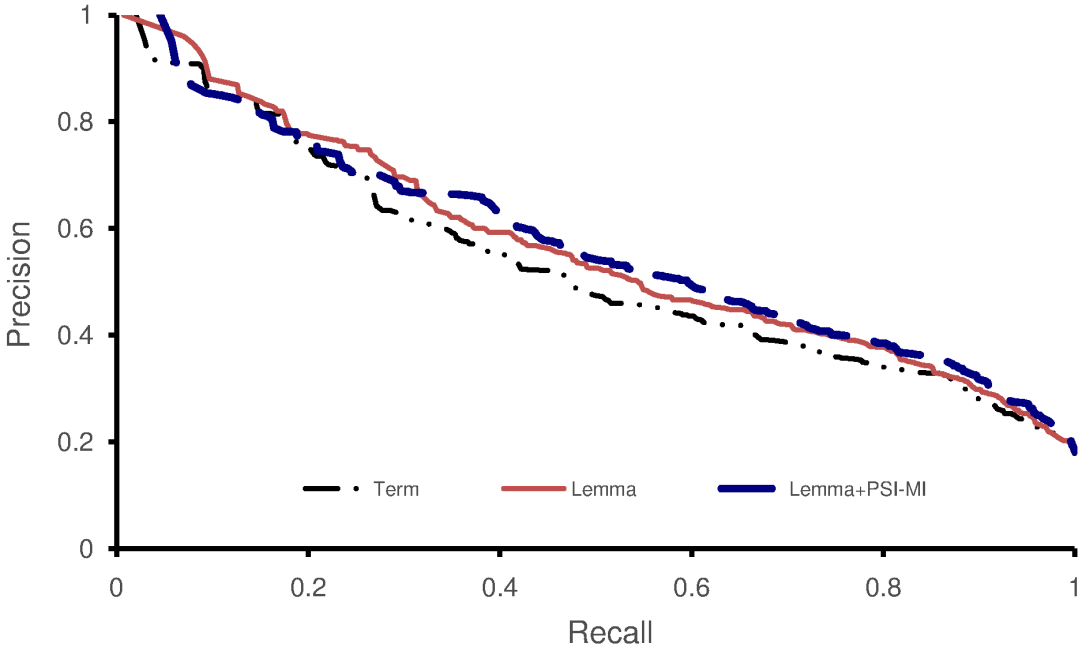

Figure 3: Vector-space classification: impact of lexicon on classification performance

We also compared the use of different lexicons to represent the documents (Fig. 3). Comparing the use of lemmas against the use of the textual occurrence of the lexicon terms, the overall improvements in AUC iP/R varied between 3% (for M=50) and 6% (for M=500). A further improvement was obtained when PSI-MI terms were used in conjunction of BioLexicon lemmas, as seen on the figure.

In order to compare the proposed ranking strategies, we used the official BC-III performance metrics: AUC iP/R, sensitivity, specificity, MCC and F-score. Table 2 shows the results, obtained on the BC-III ACT test set, for six combinations of classifiers and features. The corresponding iP/R curves are shown in Fig. 4. The best result in terms of AUC iP/R was obtained using a SVM classifier using unigrams, bigrams and lexicon lemmas as features (0.654). In terms of overall accuracy, the best results were obtained with a MaxEnt classifier using unigrams and bigrams, with lexicon terms substituted by their lemmas (0.879). Generally, the SVM classifier achieved slightly lower specificity (0.914 vs. 0.936) but significantly higher sensitivity (0.664 vs. 0.585). Interestingly, using a MaxEnt classifier with lemmas as features led to worst results than using the same classifier with unigram and bigram features extracted from the texts (AUC=0.578 vs. AUC=0.619).

We also explored the use of these classification models for ranking the results of general PubMed queries. For this, we used four queries: “Alzheimer Disease”, “Breast Cancer”, “Hypertrophic Cardiomyopathy” and “Chronic Myeloid Leukemia”. We ran each query using the Entrez e-utilities, and ranked the results using the vector-space classification method. This method was selected because of its simplicity in terms of implementation, but any of the classifiers could be used in the same approach. We ran each query as a MeSH term query, and limited the results to the years from 2006 to 2010. The total number of documents returned ranged from 1143, for “Hypertrophic Cardiomyopathy” to 32338, for “Breast Cancer”. Table 3 shows the top five ranked results for the query “Alzheimer Disease”.

Table 2: Classification results on the BioCreative III ACT test set (AUC iP/R: area under the interpolated precision recall curve; MCC: Matthew's correlation coefficient)

Classifier features	AUC iP/R	Sensitivity	Specificity	Accuracy	MCC	F-score
Vector-space lemmas	0,578	0,523	0,926	0,865	0,461	0,540
Naïve Bayes 1+2 grams	0,559	0,513	0,937	0,873	0,479	0,551
MaxEnt lemmas	0,578	0,573	0,918	0,866	0,485	0,564
MaxEnt 1+2 grams	0,619	0,580	0,932	0,879	0,520	0,592
MaxEnt 1+2 grams lemmas	0,629	0,585	0,936	0,878	0,520	0,592
SVM 1+2 grams lemmas	0,654	0,664	0,914	0,876	0,547	0,619

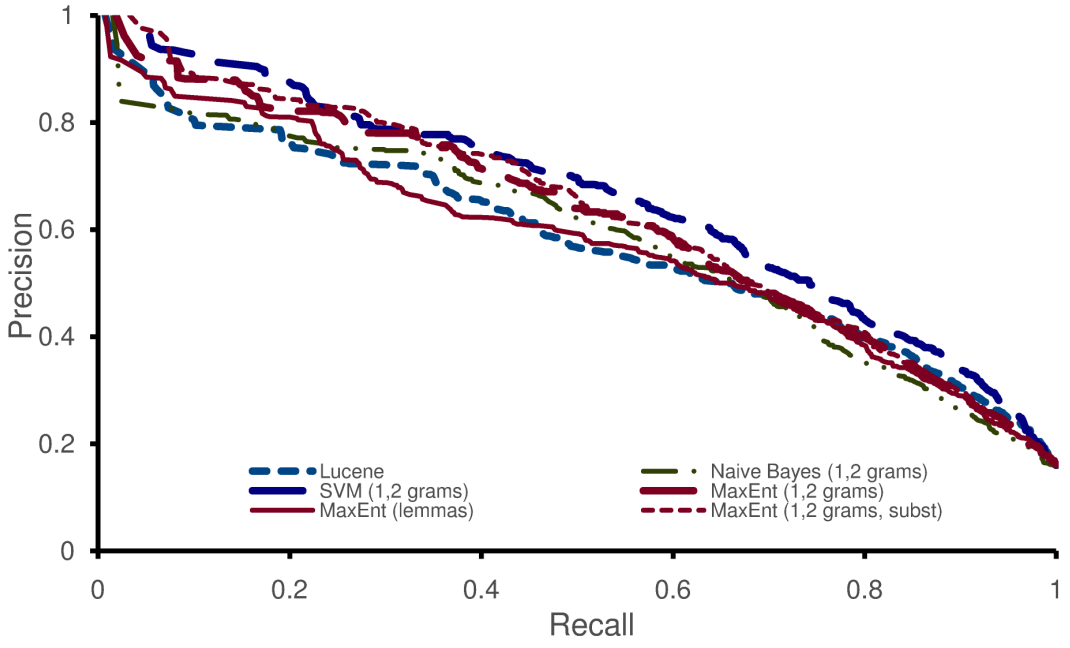

Figure 4: iP/R curves for the BioCreative III ACT test set

Table 3: Top five results for a PubMed search for "Alzheimer Disease"

Search term: Alzheimer Disease Total results: 14448
alpha-Synuclein and its disease-related mutants interact differentially with the microtubule protein tau and associate with the actin cytoskeleton. Neurobiol Dis. 2007 Jun;26(3):521-31. Epub 2007 Feb 16. PMID: 17408955
Phosphorylation regulates tau interactions with Src homology 3 domains of phosphatidylinositol 3-kinase, phospholipase Cgamma1, Grb2, and Src family kinases. J Biol Chem. 2008 Jun 27;283(26):18177-86. Epub 2008 May 8. PMID: 18467332
Interaction between presenilin 1 and ubiquilin 1 as detected by fluorescence lifetime imaging microscopy and a high-throughput fluorescent plate reader. J Biol Chem. 2006 Sep 8;281(36):26400-7. Epub 2006 Jun 30. PMID: 16815845
Histone deacetylase 6 interacts with the microtubule-associated protein tau. J Neurochem. 2008 Sep;106(5):2119-30. Epub 2008 Jul 12. MID: 18636984
RNA aptamers selectively modulate protein recruitment to the cytoplasmic domain of beta-secretase BACE1 in vitro. RNA. 2006 Sep;12(9):1650-60. Epub 2006 Aug 3. PMID: 16888322

An initial analysis of the ranked PubMed results, both in terms of title and abstract, suggests that the top documents contain some information related to protein-protein interactions. However, validating these results is not straightforward since there is no gold-standard to compare to. Therefore, we compared the results with a list of MEDLINE documents compiled from the publications used for annotation purposes by the following databases: BioGRID, IntAct[5], MINT[6], MIPS[7], NCBI. The complete "silver-standard" list contains 42890 MEDLINE document identifiers which are known to be relevant for PPI extraction. Using this list for validation, we obtained 100% recall for three of the four queries. For the "Alzheimer Disease" query, the recall was 82%, corresponding to 14 of 17 documents previously used for annotating PPIs being classified as relevant. Specificity varied between 52% and 73%. However, this value does not represent the real performance, since many positive documents will not be recorded on the "silver-standard" list.

5 Discussion

Biomedical literature retrieval is becoming a very demanding task for researchers, given the large amount of publications appearing every day. Tools to prioritize the existing publications given a specific topic of interest could help researchers find relevant documents more efficiently.

[5]IntAct molecular interaction database, `http://www.ebi.ac.uk/intact/`
[6]Molecular INTeraction database, `http://mint.bio.uniroma2.it/mint/`
[7]Mammalian Protein-Protein Database, `http://mips.helmholtz-muenchen.de/proj/ppi/`

An important topic in the biomedical domain is related to protein-protein interactions. Although many existing databases focus on this aspect, most information on PPIs is still only available in publications, where they are difficult to find by researchers.

This paper presents a methodology for ranking documents resulting from a PubMed query regarding the likelihood of containing information relevant for extracting protein-protein interactions. This ranking is performed according to previously classified documents in a training set. This can be seen as a classification problem, in which each document should be classified as belonging to the topic (positive) or otherwise (negative). The ranking can then be defined by the membership probability for the positive (relevant) class, as determined by the classifier. For evaluation purposes, it is expected that relevant documents be ranked higher than non-relevant documents. This can be measure by looking at precision values at different recall rates, or by a measure such as the area under the iP/R curve.

In order to evaluate the ranking method, we used the BC-III PPI-ACT dataset. We tested different classifiers and obtained a top result of AUC iP/R=0.654, using a SVM classifier. The classifiers evaluated allow setting different operating points, which could be more useful for different users. Database curators, for example, usually prefer higher precision, while general biologists may be more interested in obtaining a higher recall.

We also analysed the use of the method for ranking the result of general PubMed queries. The strategy proposed here can be used to identify articles describing PPIs within a particular area of interest, such as “Breast Cancer” or “Alzheimer’s Disease”. Testing the ranking method on such queries is not a simple procedure, since a gold standard cannot be established. We have therefore assessed the results by comparing to a list of true positive articles, compiled from five PPI databases. This gives a good indication of the method’s sensitivity (or recall) regarding the already known and annotated PPI articles. On the other hand, estimating the specificity (or precision) using this methodology is less accurate since it is expected that many articles describing PPIs have not yet been annotated.1 Nevertheless, our subjective evaluation of the results indicates that the large majority of the top ranked documents for each query are related to PPIs, although many of them are not included in any of the PPI resource datasets we used. This is in agreement with the expected result that many evidences for PPI are still only found in the literature.

6 Conclusions

We evaluated different methods and models for classifying scientific abstracts as to whether they are relevant for extracting protein-protein interactions. The best result (AUC iP/R=0.654; MCC=0.547; Accuracy=87.6%) was obtained using a Support Vector Machine classifier with unigrams, bigrams and lexicon lemmas as document features. These values are comparable to the best results from the BioCreative III ACT task.

Using this classification method, an approach for ranking the results from any general PubMed query in terms of their likelihood for describing PPIs is proposed. This approach makes use of the PubMed web services, allowing a simple and efficient way to provide researchers with prioritized results for their PubMed queries. The initial results indicate that this ranking may facilitate the identification of important interactions within a specific topic of interest.

Further research on the features used to represent the documents and on text preprocessing may

help improve the classification results. Also, it would be interesting to integrate the results obtained from this type of strategy within a graphical display, in order to give researchers a more interactive way to explore the possible interactions contained in the documents.

Acknowledgements

The research leading to these results has received funding from Fundação Para a Ciência e a Tecnologia (FCT) under the project number PTDC/EIA-CCO/100541/2008 (FCOMP-01-0124-FEDER-010029). S. Matos is funded by FCT under the Ciência2007 programme.

References

[1] National Library of Medicine. MEDLINE Fact Sheet. http://www.nlm.nih.gov/pubs/factsheets/medline.html. Accessed 3 December 2010.

[2] Altman R, Bergman C, Blake J, Blaschke C, Cohen A, Gannon F, Grivell L, Hahn U, Hersh W, Hirschman L *et al.* Text mining for biology - the way forward: opinions from leading scientists. Genome Biol 9 (Suppl 2):S7, 2008.

[3] Rebholz-Schuhmann D, Kirsch H, Couto F. Facts from text–is text mining ready to deliver?. PLoS Biol 3(2):e65, 2005.

[4] Lehne B, Schlitt T. Protein-protein interaction databases: Keeping up with growing interactomes. Hum Genomics 3(3):291-297, 2009.

[5] Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol 9:(Suppl 2):S8, 2008.

[6] Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 7:119-129, 2006.

[7] National Library of Medicine. Entrez Programming Utilities. `http://eutils.ncbi.nlm.nih.gov/corehtml/query/static/eutils_help.html`. Accessed 3 December 2010.

[8] Suomela BP, Andrade MA. Ranking the whole MEDLINE database according to a large training set using text indexing. BMC Bioinformatics 6:75, 2005.

[9] Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA. MedlineRanker: flexible ranking of biomedical literature. Nucleic Acids Res 37(Web Server issue):W141-W146, 2009.

[10] Jang H, Lim J, Lim JH, Park SJ, Lee KC, Park SH. Finding the evidence for protein-protein interactions from PubMed abstracts. Bioinformatics 22(14):e220-e226, 2006.

[11] Yin L, Xu G, Torii M, Niu Z, Maisog JM, Wu C, Hu Z, Liu H. Document classification for mining host pathogen protein-protein interactions. Artif Intell Med 49(3):155-160, 2010.

[12] Marcotte EM, Xenarios I, Eisenberg D. Mining literature for protein-protein interactions. Bioinformatics 17(4):359-363, 2001.

[13] Lan M, Tan CL, Su J. Feature generation and representations for protein-protein interaction classification. J Biomed Inform 42(5):866-872, 2009.

[14] Abi-Haidar A, Kaur J, Maguitman A, Radivojac P, Rechtsteiner A, Verspoor K, Wang Z, Rocha LM. Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks. Genome Biol 9 (Suppl 2):S11, 2008.

[15] Krallinger M, Vazquez M, Leitner F, Salgado D, Valencia A. Results of the BioCreative III (Interaction) Article Classification Task. In Proceedings of the Third BioCreative Workshop, Bethesda, USA, 13-15 September 2010, 2010.

[16] He M, Wang Y, Li W. PPI finder: a mining tool for human protein-protein interactions. PLoS One 4(2):e4554, 2009.

[17] Sasaki Y, Montemagni S, Pezik P, Rebholz-Schuhmann D, McNaught J, Ananiadou S. BioLexicon: A Lexical Resource for the Biology Domain. In Proceedings of the Third International Symposium on Semantic Mining in Biomedicine, Turku, Finland, 1-3 September 2008, 2008.

[18] HUPO Proteomics Standards Initiative. MI Ontology. `http://psidev.sourceforge.net/mi/rel25/data/psi-mi25.obo`. Accessed 3 December 2010.

[19] The Apache Software Foundation. Apache Lucene. `http://lucene.apache.org/`. Accessed 3 December 2010.

[20] McCallum AK. MALLET: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu`. Accessed 3 December 2010.

[21] Chang CC, Lin CJ. LIBSVM : a library for support vector machines. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`. Acessed 10 May 2011.