BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 7 2006, pages 857–865
doi:10.1093/bioinformatics/btk044

Databases and ontologies

LSAT: learning about alternative transcripts in MEDLINE

Parantu K. Shah[1,2,*] and Peer Bork[1,2]

[1]European Molecular Biology Laboratory, Heidelberg, Germany and [2]Max Delbrück Centre for Molecular Medicine, Berlin-Buch, Germany

Received on October 20, 2005; revised on December 9, 2005; accepted on January 5, 2006
Advance Access publication January 12, 2006
Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Generation of alternative transcripts from the same gene is an important biological event due to their contribution in creating functional diversity in eukaryotes. In this work, we choose the task of extracting information around this complex topic using a two-step procedure involving machine learning and information extraction.

Results: In the first step, we trained a classifier that inductively learns to identify sentences about physiological transcript diversity from the MEDLINE abstracts. Using a large hand-built corpus, we compared the sentence classification performance of various text categorization methods. Support vector machines (SVMs) followed by the maximum entropy classifier outperformed other methods for the sentence classification task. The SVM with the radial basis function kernel and optimized parameters achieved F_β-measure of 91% during the 4-fold cross validation and of 74% when applied to all sentences in more than 12 million abstracts of MEDLINE. In the second step, we identified eight frequently present semantic categories in the sentences and performed a limited amount of semantic role labeling. The role labeling step also achieved very high F_β-measure for all eight categories.

Availability: The results of our two-step procedure are summarized in the LSAT database of alternative transcripts. LSAT is available at http://www.bork.embl.de/LSAT

Contact: shah@embl.de

Supplementary information: Supplementary data are available at *Bioinformatics* online

INTRODUCTION

Published literature is the largest repository of biological information and this information is generally curated into community knowledge-bases by human experts. Explosive growth of publications is making it harder for human experts to keep track of the state-of-the-art knowledge and quickly update the knowledge-bases. Thus, text-mining methods are becoming increasingly important in molecular biology to handle collections of biological texts automatically. Such methods include systems that efficiently classify and retrieve documents in response to complex user queries, and beyond this, systems that carry out a deeper analysis of the literature to extract specific events or relationships, such as tissue-specific splicing or protein–protein interactions and fill database entries with information about the participating gene products and circumstances of the event (Krallinger and Valencia, 2005).

*To whom correspondence should be addressed at European Molecular Biology Laboratory, Meyerhofstraße 1, Heidelberg 69117, Germany

Generation of alternative transcripts in different cells or tissues are contributing events for the functional complexity and evolution of eukaryotes (Boue *et al*., 2003). Alternative transcripts generated with alternative splicing (AS) allow eukaryotes to generate different proteome from a limited amount of gene pool. Differential promoter usage and alternative polyadenylation in synergy with AS may change terminal exons or in general regulate expression of mRNA transcripts (Black, 2000; Edwalds-Gilbert *et al*., 1997; Zavolan *et al*., 2003). Several instances of these mechanisms are scattered in the literature and it is important to have a curated list of genes that utilizes the above mentioned mechanisms to express alternative transcripts in various tissues and across species to facilitate annotation of transcriptome. Moreover, knowledge about differences in structure/function of alternative transcripts is also important for function annotation. Therefore, an information extraction tool is much required by the community working on elucidating the extent of usage of these mechanisms and their functional implications. It will also provide experimentally verified training sets to develop computational methods for predicting such events. Thus, we aim to identify descriptions of alternative transcripts from abstracts in MEDLINE. Furthermore, we concentrate on finding out information about alternative transcripts expressed only in natural (non-disease) states.

A number of efforts for event/relationship extraction that label constituents of sentences with appropriate roles are already underway (Daraselia *et al*., 2004; Novichkova *et al*., 2003; Yakushiji *et al*., 2001). High performance event/relationship extraction usually requires full-parsing of sentences and a reliable database of predicate argument structures (Pradhan *et al*., 2004). Efficient and accurate parsing of biomedical texts is not within the reach of current parsers. Standard methods are computationally expensive to use and are trained on English texts from the newswire domain (Shatkay and Feldman, 2003). Thus, full parsing of all sentences could be impractical when applied to a large database like MEDLINE or full-text articles. A database of predicate argument structures for biomedical domain is still under development (Wattarujeekrit *et al*., 2004). Hence, any practical event extraction task should be preceded by the identification/retrieval of the event-containing sentences that extraction systems can handle. This binary classification step would constrain the number of predicates, giving a better idea of the semantic roles of sentence constituents and reduce computational demands. It would also help to prioritize the predicates for PAS analysis in the PASBio database (Wattarujeekrit *et al*., 2004) for biomedical event extraction.

In this work we show feasibility of the sentence classification task with inductive learning for obtaining sentences about alternative

transcripts. It has been already suggested that classifiers at the sentence level have the potential to improve precision of information extraction (Craven and Kumlien, 1999). Retrieval at the sentence level was followed by a high precision semantic role labeling step to generate a database of experimentally verified alternative transcripts and associated information such as gene name, species, tissue, mechanism, expression-specificity, difference in structure/function of the alternative transcripts, etc.

Inductive learning methods learn patterns from the features extracted from the training set and generalize. Generalization performance of many methods degrades when dealing with large amounts of rarely occurring features. Text data are a typical example of this situation. Moreover, the process of preparing a reliable training set is expensive and time-consuming. Hence, a good learning method should be able to learn from a small amount of training examples and should be able to handle large number of features. We compared the performance of well-known text categorization methods: (1) naïve Bayes, (2) maximum entropy, (3) Expectation Maximization (EM), (4) variants of the term frequency-inverse document frequency (tf*idf) methods, (5) K-nearest neighbor (KNN) algorithm and (6) support vector machines (SVM) for the sentence classification task (Mitchell, 1997).

The classification performance was compared for inductive learning with different fractions of the training set in order to choose the best performing method. Then, we carried out parameter optimization of the SVM classifier, the best performing method. Four different feature sets differing in richness of features were generated and tested for the generalization performance of the SVM. All sentences in MEDLINE abstracts were classified using the trained classifier. Eight semantic categories were identified from the extracted sentences and sentence constituents were labeled with the appropriate category. We show benchmarks for the sentence classification as well as for the tagging step.

METHODS

Sentence classification by inductive learning

We sought to perform the sentence classification task using inductive machine learning. During inductive learning the learner $\boldsymbol{L}$ is given a training set $\boldsymbol{S}$ containing $\boldsymbol{n}$ examples $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$, $y \in \{-1, +1\}$. Each example consists of a text feature vector $\vec{x}$ and its class label y. The learning task involves the maximization of correct class labels. In such a setup, the classification performance of any method depends upon the quality of features in the training examples and various learning parameters. Also, methods good at selecting useful features and rejecting others will outperform those that cannot. The learning performance of a trained classifier is assessed on a set of previously unseen examples. The learning process is repeated until the classifier achieves a satisfactory performance.

We used the Bow toolkit (http://www-2.cs.cmu.edu/~mccallum/bow/) for experiments with naïve Bayes, EM, maximum entropy (Nigam *et al*., 1999), tf * idf and its variants. SVM implementations from the package SVM^{light} was used (http://svmlight.joachims.org). Please see Joachims (2001), Mitchell (1997) and Ribeiro-Neto (1999) for a detailed discussion of the methods used here. Also, see Supplementary material for a short introduction to SVM and various kernels.

Generation of the training corpus

Generating a reliable training set is a slow and a labor-intensive task. There are no publicly available datasets for carrying out information extraction about alternative transcripts. Thus, generating training set for the sentence classification was a limiting step. For this, we manually chose appropriate sentences from article titles and abstracts (Fig. 1a; Supplementary Figure 1). The training set was generated iteratively using three rounds of inductive learning with various classifiers until satisfactory classification performance was achieved. The final set contained 4240 positive sentences and 13 520 negative sentences.

The positive sentences described instances of generation of physiologically relevant (natural) transcript diversity (TD) from a single gene. They contained descriptions of various ways of AS (e.g. usage of alternative first exons, exon skipping, intron retention and usage of cryptic splice sites), differential promoter usage, alternative polyadenylation or mentioning of multiple transcripts form the same gene. Sentences describing transcripts that differed at the protein level (absence/presence of protein domains or motifs) and those containing descriptions of tissue and developmental stage specific transcripts were also considered positive sentences.

The negative training set included sentences that did not provide information about alternative transcripts. In particular, negative training set was enriched with sentences describing aberrant transcripts generated owing to splice site mutations and transcripts generated in diseased tissues, splicing mechanisms, and ordinary gene expression events and exon/intron structure of genes among others. These sentences utilize similar vocabulary to those that describe alternative transcripts as they all belong to the domain of gene expression (Supplementary Figure 1a). This is also the case with sentences describing protein isoforms that may be generated by different gene paralogs. These kinds of sentences pose challenges to the sentence classification process. The kappa score for inter-annotator agreement on the final training set was 0.98 (Cohen, 1960).

Pre-processing

The Oak system (http://nlp.cs.nyu.edu/oak/; S. Sekine, unpublished data) was used to split abstract text into sentences. Sentences were broken into words, assigned part of speech tags and stemmed using the Tree-tagger (Schmid, 1994). For generating the input feature set, stop words (see Supplementary material) and words occurring with very low frequencies (<5) were removed from the list of words. The pre-processing module is shown in Figure 1b.

Feature enrichment

The process of extracting a rich feature set from the training examples is the most important step in machine learning because methods provided with rich feature require fewer training examples and provide better generalization. Feature enrichment was achieved as follows (Fig. 2). Many phrases (word bi-grams and tri-grams) frequently present in the positive training examples were incorporated as additional input features. In most cases the phrases reflected text patterns specific to description of alternative transcripts. For example, they reflect existence of alternative transcripts (additional transcript), mechanisms ('AS or alternative first exon or second promoter'), specificity (e.g. 'brain-specific'), etc. Such word usage distinctly specifies addition of 'domain knowledge' to learning features. Cardinals were summarized as a single feature. In addition, synonyms were defined for the sparsely occurring features (e.g. long transcript, larger transcript and elongated transcript). The process of feature enrichment added additional 900 learning features in terms of word bi-grams, tri-grams and synonyms.

Generation of four feature sets

The resultant of pre-processing is a feature set containing all the words (bag of words) occurring in the corpus with a total of 23 742 features. We also generated a second feature set termed 'vocabulary' by manually inspecting and removing non-essential words from the first file to result in 9590 features for this set. We combined 900 features resultant of 'feature enrichment' procedure to 'bag of words' and 'vocabulary' to generate two additional feature sets. Sentences regenerated as feature vectors were used as input to various learning methods. We compared classification performance of various methods using 4-fold cross validation on the training set.

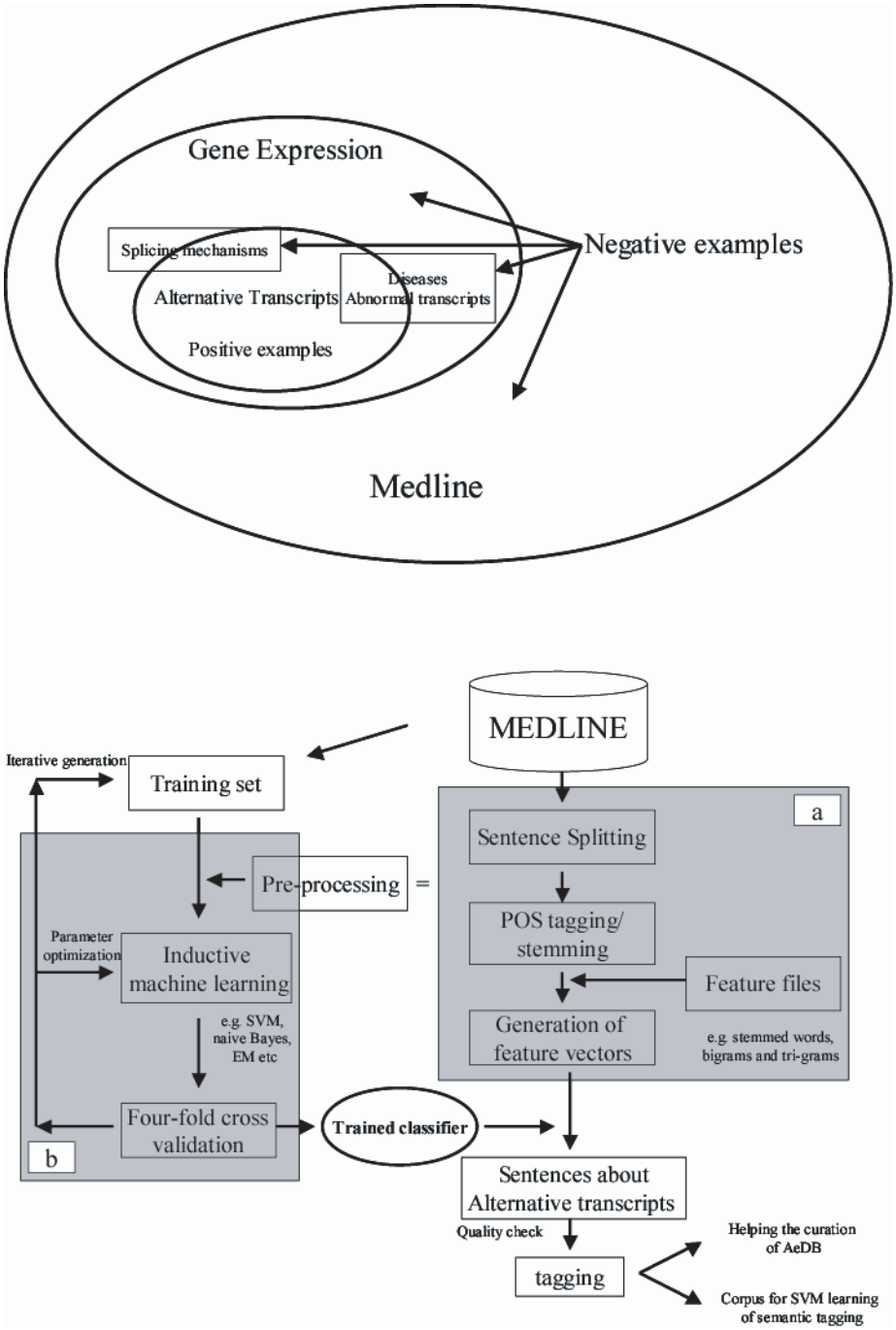

Fig. 1. The procedures for generating learning corpus (top panel) and learning procedure are shown. Generation of alternative transcripts is a part of gene expression process. Thus, sentences describing gene expression have similar contents. Moreover, alternative transcripts generated because of disease conditions and other reasons (see Methods) poses major challenges to the learner due to presence of many common word patterns. The bottom panel shows the pre-processing (a) and training (b) modules.

Sets for benchmarking the performance of the best SVM classifier

In order to prove the strength of our machine learning approach, we benchmarked the classification performance of the best SVM classifier for extracting sentences describing only the natural TD. Annotators at the National Library of Medicine provide the MeSH term alternative splicing (one of the mechanisms for generating TD) to appropriate MEDLINE abstracts. Hence, a limited determination of recall was possible by comparing the abstracts (sentences) identified by the classifier with the MeSH term annotation for AS.

MEDLINE 2004 contained 8133 records (abstracts) with the MeSH term 'alternative splicing' assigned to them. But only a subset of records contains information about physiologically relevant (natural) TD. For example,

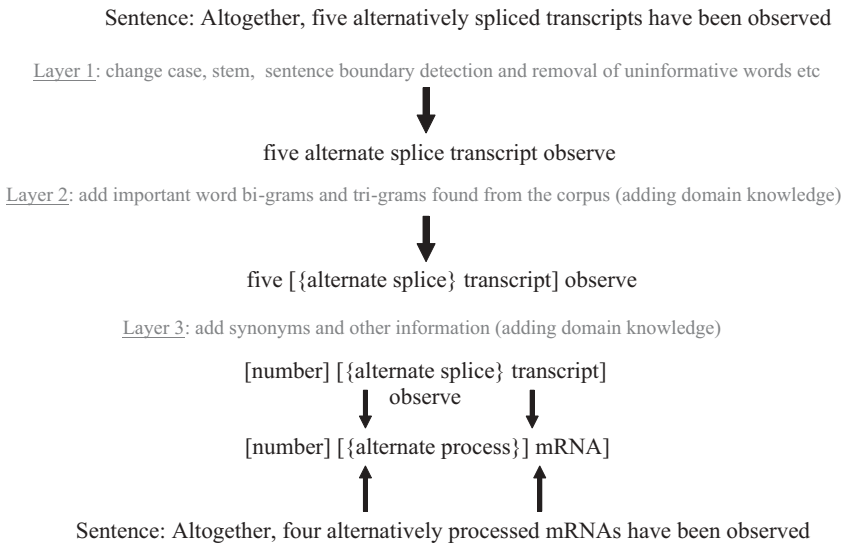

Fig. 2. The procedure of feature enrichment that allows two seemingly different sentences to merge into a single pattern is shown here. The usage of reducing numbers in to single features, use of bi-grams (e.g. AS; see Supplementary material) and synonymous (e.g. mRNA and transcript) are illustrated.

1725 of these records also contained the MeSH term 'mutation', usually referring to cases of aberrant transcripts and 489 records were without any abstract text. Hence in order to maintain consistency, we removed 2214 records from the list and used the remaining records while benchmarking for the recall.

From the sentences retrieved by the SVM classifier, we extracted instances of eight semantic categories and evaluated the precision and recall by manually inspecting 300 randomly selected sentences for each semantic category.

Extraction of eight semantic categories

We tagged gene names using NLprot tagger (Mika and Rost, 2004). Tissues and species were tagged using a dictionary made from compilations provided by Swissprot and Refseq. Tissue and species specificity were identified by tagging the word 'specific*' that may follow the tagged tissue/species name or part of the word describing the tissue/species (e.g. brain-specific). Event mechanisms, differences in structure and function of alternative transcripts and experimental methods were tagged using rules based on predicate argument structures (see Supplementary material). 'Number of isoforms' was extracted by the fact that words indicating number of isoforms (with part of speech tag CD [cardinals]) always preceded the tagged event mechanisms. Apart from phrases extracted using the predicate argument structure analysis, event mechanisms were also extracted using the bi-gram and tri-gram they were part of.

Definitions of precision, recall and F_β-measure

The precision and recall of the classifier are defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{and} \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

where TP, TN, FP and FN are true positives, true negatives, false positives and false negatives. The F_β-measure is defined as below.

$$F_\beta = \frac{(1 + \beta^2) \times \text{Prec} \times \text{Rec}}{\beta^2 \times \text{Prec} \times \text{Rec}}.$$

We give equal weight to precision and recall and take $\beta = 1$. Hence, the F_β-measure is the harmonic mean of precision and recall.

RESULTS

Selecting the best sentence classification method

Our aim was to find out the best sentence classification method for the training set involving more than 23 000 features. Intuitively, while training with the most basic set the learning method good at feature selection would outperform the rest. Hence, we checked the performance of various methods with different fractions of the training set while utilizing the simplest feature set (bag of words).

The classification performance of all methods, defined as the F_β-measure, improved with increase in the amount of training data, as expected (Fig. 3). However, SVM and maximum entropy classifiers consistently outperformed the others. Also, the SVM classifier with the radial basis function (RBF) kernel outperformed that with the linear kernel even though text data are considered to be linearly seperable. The KNN algorithm with number of neighbors ranging from 5 to 50 either suffered from memory problems or does not seem to learn the classification rule (data not shown). Adding part of speech tags as additional learning features did not improve classification accuracy. It was clear that the SVM with three different kernels performed better than the other methods (Fig. 3) and were taken for parameter optimization.

Parameter optimization for the SVM learning

We explored classification performance of the SVM with three different kernel functions; the linear function, the RBF and the sigmoid function and associated learning parameters (see Supplementary material). For all kernel functions the value of parameter C in the SVM optimization problem controls the trade-off between the training error and the margin (Joachims, 2001). The optimal value of C depends on the training data and it was determined empirically. In addition, the RBF and sigmoid functions have one and two model parameters respectively that can affect the learning process.

We characterized the value for parameter C with different values of gamma for the RBF kernel and the value of r for the sigmoid

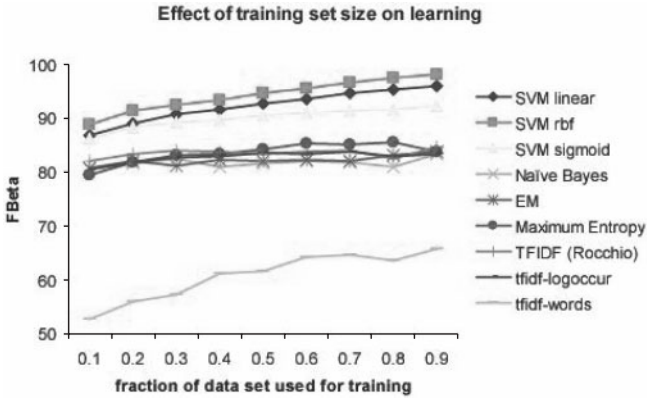

Fig. 3. Classification performance of various text categorization methods studied as F_β-measure (y-axis) with different fractions of training set. EM is a short-form for Expectation Maximization methods and tf $*$ idf stands for term frequency-inverse document frequency methods. Each data point is a mean of 4-fold cross validation. See Supplementary Table 1 for the data points used to plot this figure.

kernel with bag of words as the input feature set (Fig. 4, left panel). The value of 1.5 for gamma and the value of 10 for C were the optimal classification parameters for the SVM with the RBF kernel. Similarly, the value of 0.01 for r and the value of 1000 for C were the optimal parameters for the SVM with the sigmoid kernel.

Effect of enriched feature sets on SVM learning

We checked the effect of all four input feature sets on the classification performance of the SVM with different values of C. The SVM with the RBF kernel outperformed SVM with linear and sigmoid kernels in the classification accuracy while learning with all four feature sets (Fig. 4, right panel). It achieved a mean F_β-measure of 91% (precision = 92.43; recall = 89.94) when performing four randomized trials with 60% of total corpus as training set and the rest as the test set (Table 1, in bold). The input feature set 'bag of words and phrases' performed best for the SVM with linear and RBF kernels but not with sigmoid kernel (Fig. 4 and Table 1). Thus, the contribution of bi- and tri-gram phrases depends on the classification method and learning parameters. This is in accordance with previous text categorization experiments where both performance degradation and performance increase were reported as a result of use of phrases as learning features (Tan *et al*., 2002). Utilization of feature selection methods may have the potential of reducing the total number of input features and the classification performance. However, one of the aims of our work was to find out a sentence classification method that is inherently good at performing feature selection. Therefore we did not do any feature selection before machine learning

Learning performance of the SVM

Support vectors are the training examples closest to the hyperplane (Supplementary Figure 2) and their use allow SVM to perform classification independent of total number of input features (Joachims, 2001). The number of support vectors is an indication of the complexity of an SVM model. The number of support vectors used by the SVM with linear, sigmoid and RBF kernels increased in that order (Fig. 4; see Methods). Also, for SVM with linear and sigmoid kernels the required number of support vectors decreased with the richness of input feature sets, in contrast to the SVM with the RBF kernel (Fig. 5a).

We measured the training errors using the $X_{i\alpha}$ estimators supplied with the SVM$^{\text{light}}$ software. In our experience they gave reliable estimates of the classifier performance on the test set. The training error is lowest when bag of words and phrases was used as a feature set in case of each kernel (Fig. 5b). It is clear from the figure that RBF kernel function brought the lowest training error. The polynomial functions of order more than one performed equivalent or poorer to the linear function in the experiments described above (data not shown). Hence, we used the SVM with the RBF kernel with a gamma value of 1.5, C value of 10 and bag of words and phrases as the feature set for classification of entire MEDLINE.

Benchmarking of the classifier performance on MEDLINE

The trained SVM classifier was used to rank all sentences in MEDLINE. It assigned positive scores to 31 123 sentences from more than 12 million MEDLINE abstracts, identifying them as positive sentences. A manual inspection of the extracted sentences resulted in retaining 20 549 sentences describing TD. This gives precision of 66% to the classifier while classifying all sentences in MEDLINE. An appropriate threshold could be used to reduce the false positives from the ranked sentence. However, we chose not to use the threshold values in this study.

The sensitivity of the classifier was assessed against manual annotations of AS provided by the MEDLINE curators. All entries (5919) with the MeSH term alternative splicing were taken form the MEDLINE 2004 database (see Methods). The classifier detected 4400 out of 5919 abstracts, resulting in a recall of 74% and F_β-measure of 70%.

We manually checked the abstracts missed by the SVM classifier. In many cases the sentences (abstracts) missed by the classifier were describing AS in the normal versus the diseased tissues and these abstracts did not explicitly mention changes in gene sequence as the basis of AS. Since the SVM classifier was trained to identify only physiological (natural) alternative transcripts, these abstracts were counted as true negatives. The final recall of the classifier was 84% and the F_β-measure while classifying all sentences in MEDLINE was 74%.

Semantic role labeling

During the task of semantic role labeling, for each verb in a sentence, the goal is to group sequence of words that fill a semantic role and to determine their roles (Supplementary material). Semantic role labeling is an important task towards natural language understanding, and has immediate applications in the task of information extraction. For example, in the following sentence containing the verb express and describing alternative transcripts, constituents have roles such as gene name (a), number of isoforms (b), tissue-specificity (c) and mechanism responsible for generating alternative transcripts (d).

The [calcitonin/CGRP gene $_{a}$] [expresses $_{verb}$] [two $_{b}$] different mRNAs by [tissue-specific $_{c}$] [alternative splicing $_{d}$].

We analyzed the sentences extracted by the classifier for identifying frequently occurring, biologically meaningful categories.

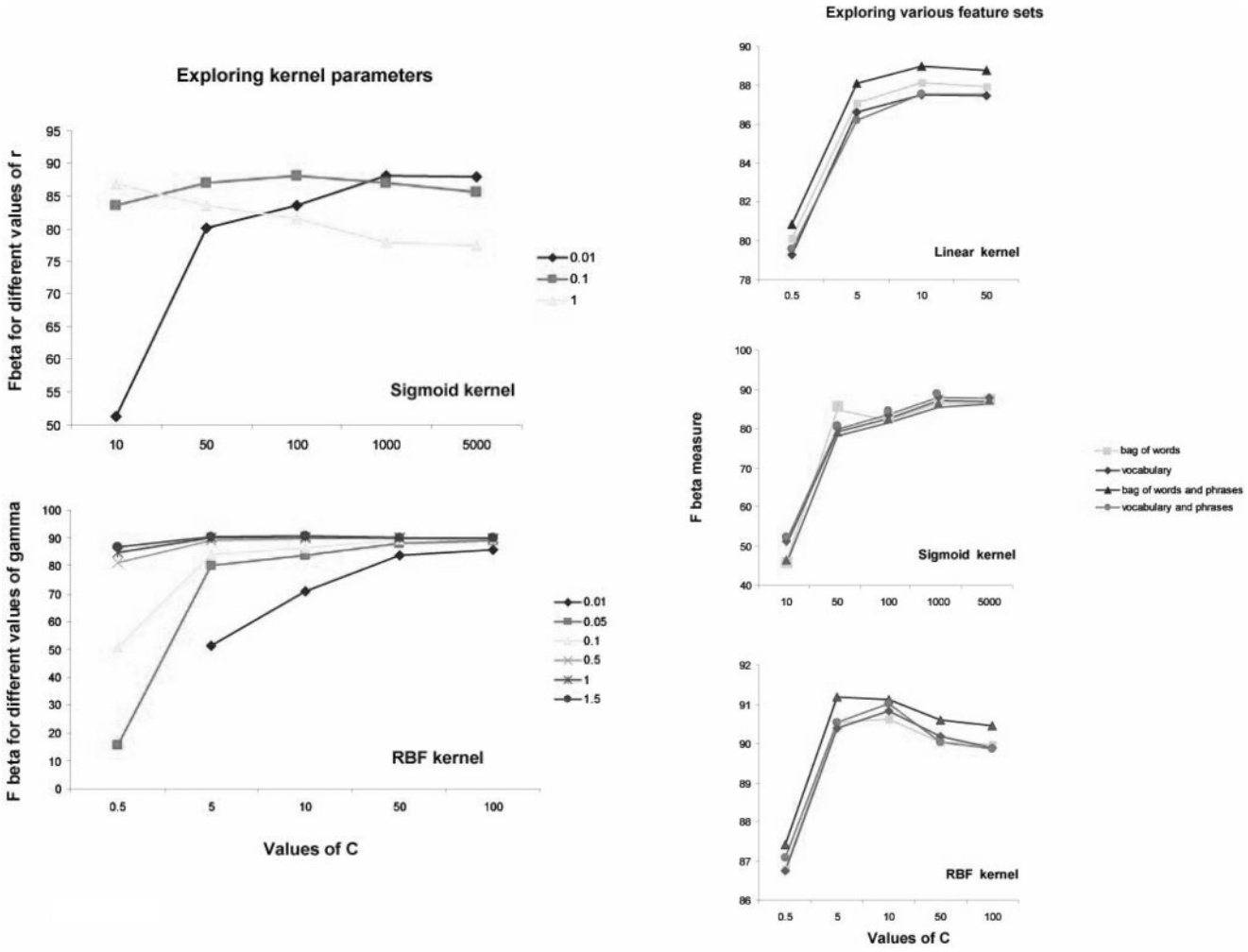

Fig. 4. Values of gamma for the RBF kernel and *r* for sigmoid function at different values of parameter *C* were studied as a function of F_β-measure; 'bag of words' was used as the input feature set (left panel). Classification performance of different SVM kernels with different values of parameter *C* was assessed for four different sets of input features (right panel). Each data point is a mean of 4-fold cross-validation. See Supplementary Tables 2 and 3 for the data points used to plot this figure.

Table 1. Performance of the SVM classifiers with best performing learning parameters

SVM kernel	Bag of words		Vocabulary		Bag of words and phrases		Vocabulary and phrases	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Linear	90.85	85.55	90.32	84.87	91.61	86.56	90.71	84.56
RBF	91.98	89.28	92.24	89.44	**92.43**	**89.84**	92.68	89.42
Sigmoid	90.19	84.87	90.85	85.55	89.75	83.50	91.62	86.53

These categories include gene names, tissues, species, differences in structure/function of alternative transcripts, expression-specificity, number of isoforms, event mechanisms and experimental methods (Supplementary Table 5). Frequent presence of these categories is indicative of their biological importance. Indeed, manual annotation of these categories could be found in the Alternative Exon database (Thanaraj *et al*., 2004) for studying the functional complexity and evolution of AS in mammalian genomes. Automated extraction of gene names, species, tissues and functional differences will also help in associating literature knowledge to sequence entries in databases like Swiss-Prot.

We got satisfactory values (Supplementary Table 5) for recall and precision for tagging of semantic categories identified from the sentences extracted by SVM classifier. The performance at the tagging boundaries was not evaluated in this study. It should be noted that not all extracted sentences provide all types of

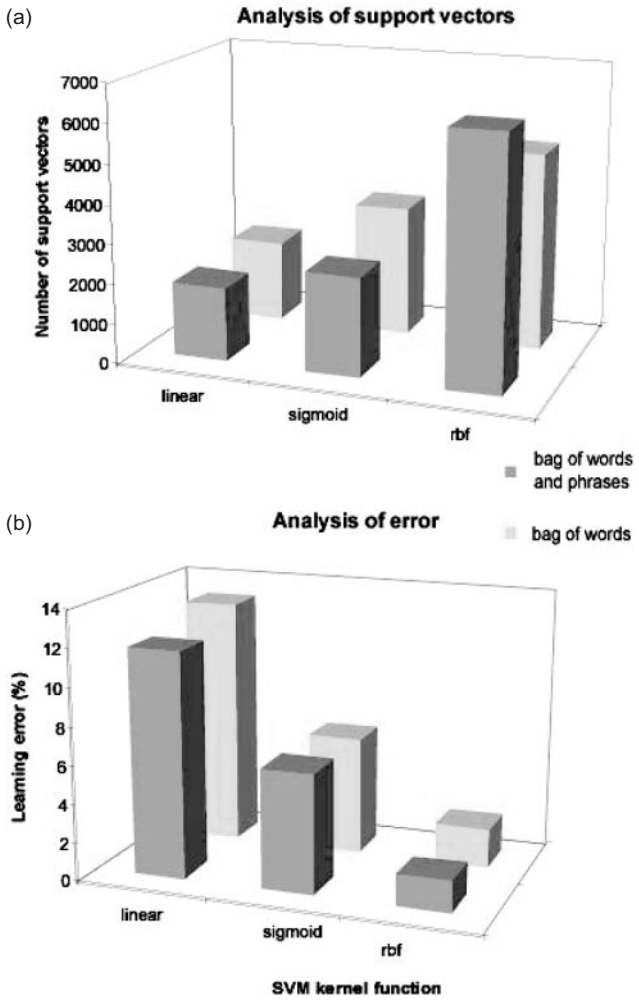

Fig. 5. Number of support vectors (**a**) and associated learning errors (**b**) brought about by the SVM with the best learning parameters and three different kernels. Their learning performances with 'bag of words' and 'bag of words and phrases' as feature sets are shown in the figure. The *C* values were 10, 10 and 1000 for linear, RBF and sigmoid kernels, respectively. The gamma-value of 1.5 for the RBF kernel and *r*-value of 0.1 for sigmoid kernel were used. All data points are mean values of 4-fold cross validation. See Supplementary Table 4 for the data points used to plot this figure.

information. For example, gene names are present in ~70% of the extracted sentences. On the other hand information about tissue-specificity was found only in 5% of the sentences, reflecting fewer known examples in the literature. Missing information could be derived from neighboring sentences using advanced methods.

LSAT—the database of alternative transcripts described in the literature

LSAT (Literature Support for Alternative Transcripts) was generated as a result of the two step procedure involving sentence classification and information extraction as described above. LSAT entries (Fig. 6) are abstract-based and provide article title, abstract text and the information identified from the sentences extracted by the SVM classifier. It also provides identifiers to Swiss-Prot, Refseq, Genbank and Ensembl, when possible. LSAT is available for interested researchers at http://www.bork.embl.de/LSAT/. LSAT contents could be searched using SQL queries as well as identifiers from Swiss-Prot, Refseq and Ensembl. We plan to update LSAT yearly.

The potential of AS, differential promoter usage and alternative polyadenylation in creating alternative transcripts in different tissues is currently being explored using specialized gene expression microarray platforms and computation tools (Boue *et al*., 2003; Lee and Roy, 2004). However, data from high-throughput analysis are usually noisy (Nadon and Shoemaker, 2002) and experimental verifications for their results are often required. Computational tools are also being developed for predicting existence of alternative transcripts. We believe that the information about alternative transcripts in LSAT will provide an ideal test set for such experiments.

Moreover, information like gene names, species, tissue-specificity, and instances of mechanisms like AS, alternative polyadenylation and differential promoter usage, extracted from the sentences will speedup the function annotation in databases like Swiss-Prot (Boeckmann *et al*., 2003), Alternative Exon database (Thanaraj *et al*., 2004) and computationally generated transcripts. The knowledge residing in LSAT has already been applied for assignment of MeSH terms to abstracts, function annotations to genes and studying usage of various TD generating mechanisms proving effectiveness of our two-step approach (Shah *et al*., 2005). We have also deposited our results to curators of the Alternative Exon database.

DISCUSSION

Most reported efforts for relationship/event extraction in biomedical texts are geared towards extraction of molecular interactions (Blaschke and Valencia, 2001; Daraselia *et al*., 2004; Novichkova *et al*., 2003; Shatkay and Feldman, 2003). However, molecular interactions are only one of the two important factors behind the phenotypic diversity in eukaryotes. The other factor is the generation and expression of multiple mRNA transcripts from a single gene. Alternative transcripts generated using mechanisms like AS has a potential to modify molecular interactions.

In the first part of this work we showed the feasibility of identification of sentences describing TD as a classification task using machine learning methods. This task is analogous to text categorization for obtaining documents of interest. We manually prepared a large training set so that classification performance of various methods could be compared while utilizing different fractions of training set and different feature sets. Creating a suitable training set is usually the rate-limiting step for machine learning methods and we aim to maintain it and make it accessible to the machine learning community.

SVM proved to be the best classifiers with 'bag of words' as a feature set containing 23 742 training features and 17 760 training examples. This set had no enriched features (e.g. phrases or synonyms) and it was expected that the method inherently better at feature selection would outperform others. Maximum entropy classifier that estimates the conditional distribution of class label given a training sentence was the second best classifier followed by the naïve Bayes, EM and tf $*$ idf methods. The KNN classifier was not able to learn classification rule and suffered with memory

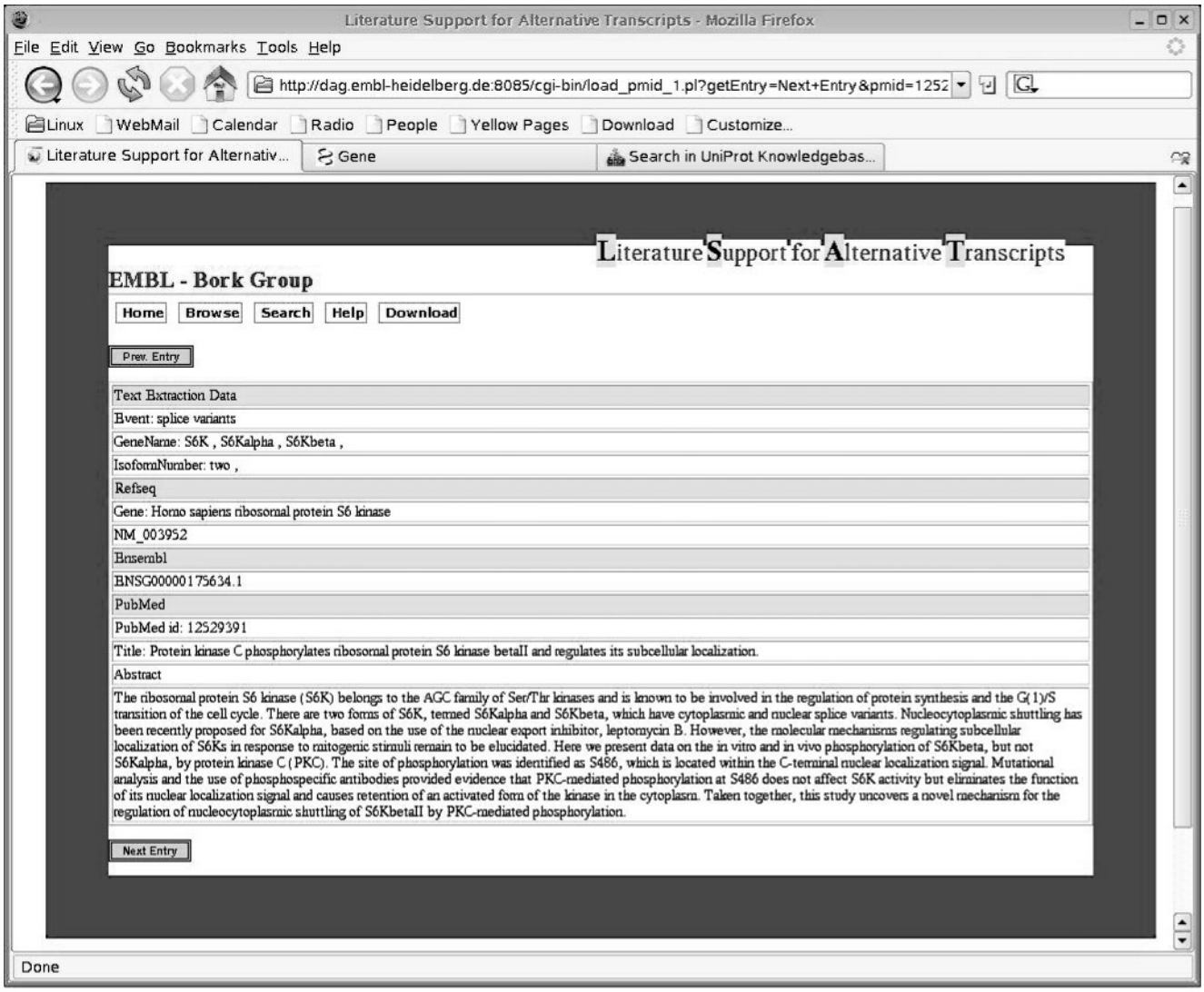

Fig. 6. An example LSAT entry is shown here. This entry identifies existence of alternative splice variants (S6Kalpha and S6Kbeta) for ribosomal protein S6 kinase gene in human as reported in the literature. It also provides identifiers to sequence databases like Refseq and Ensembl. It should be noted here that AS of S6K is not annotated in any existing sequence databases.

problems in our experiments. It is not surprising since KNN classification requires storage of huge amount of training examples with very high-dimensional feature space and its comparison with each test example. Such a trend of classification accuracies has been observed before for the text categorization task (Dumais *et al*., 1998; Yiming Yang, 1999).

SVM with the RBF kernel was the best classifier for sentence classification. The F_β-measure while classifying all sentences in MEDLINE was 74%. This performance is better than the previous sentence classification approaches described in the biomedical literature (Ray and Craven, 2001). An SVM classifier with the RBF kernel was also used by curators of BIND for their Pre-BIND and Textomy system (Donaldson *et al*., 2003). Performance of the RBF kernel results from the fact that it transforms the input features into an infinite-dimensional feature space and allows enclosed decision boundaries.

Hence, SVM classifier was able to learn multiple patterns present in the training set while handling a relatively large amount of features and provided good values for precision and recall over a huge repository of biomedical text. Moreover, there was no need to select a subset of abstracts or write rules to achieve this performance. The acquisition of domain knowledge was satisfactory when trained with appropriate examples in the training set and feature enrichment. For this reason we expect the classifier to scale-up for the mining of mRNA TD from full text articles.

The subsequent role labeling step also achieved good F_β-measures for all eight categories and it was instrumental in generation of LSAT. Machine learning of semantic role labeling is an important task and many community wide efforts are organized for it for the general English (e.g. CoNLL-2005 shared task defined at http://www.lsi.upc.edu/~srlconll). The limiting step for a similar learning task for the biomedical NLP is the availability

of a comprehensive database of predicate argument structures and an annotated corpus. We have already prepared PASBio database for predicates common in biological texts and will be including predicate frames for the additional verbs present in the sentence identified in this work (Wattarujeekrit *et al*., 2004). At present we are re-annotating the tagged sentences to prepare a learning corpus. We aim to train another SVM classifier for machine learning of semantic role labeling using the structural features derived form the parse tree and the semantic knowledge in the PAS frames. The corpus tagged by rules is available for researchers interested in semantic role labeling at http://www.bork.embl.de/LSAT/. We propose that the sentence classification and semantic learning tasks should become part of community wide competitions like BioCreAtIve (Hirschman *et al*., 2005) or KDD Challenge cup (Yeh *et al*., 2003) for the biomedical text-mining.

ACKNOWLEDGEMENTS

The authors would like to thank Ms Stephanie Boué for helping with the generation of the training set.

Conflict of Interest: none declared.

REFERENCES

Black,D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367–370.

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

Boeckmann,B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Boue,S. *et al.* (2003) Alternative splicing and evolution. *Bioessays*, **25**, 1031–1034.

Cohen,J.A. (1960) Coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, **20**, 37–46.

Craven,M. and Kumlien,J. (1999) Constructing biological knowledgebases by extracting information from text sources. In *Proceedings of the AAAI Conference on Intelligent Systems for Molecular Biology*, Heidelberg, Germany, pp. 77–86.

Daraselia,N. *et al.* (2004) Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, **20**, 604–611.

Donaldson,I. *et al.* (2003) PreBIND and Textomy—mining the biomedical literature for protein–protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, 11.

Dumais,S.T., Platt,J., Heckerman,D. and Sahami,M. (1998) Inductive learning algorithms and representations for text categorization. *Proceedings of ACM-CIKM98*, Bethesda, MD, pp. 148–155.

Edwalds-Gilbert,G. *et al.* (1997) Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res.*, **25**, 2547–2561.

Hirschman,L. *et al.* (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, **6** (Suppl. 1), S1.

Joachims,T. (2001) Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer Academic Publishers, Boston.

Krallinger,M. and Valencia,A. (2005) Text-mining and information-retrieval services for molecular biology. *Genome Biol.*, **6**, 224.

Lee,C. and Roy,M. (2004) Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol.*, **5**, 231.

Mika,S. and Rost,B. (2004) Protein names precisely peeled off free text. *Bioinformatics*, **20** (Suppl. 1), I241–I247.

Mitchell,T.M. (1997) *Machine Learning*. McGraw Hill.

Nadon,R. and Shoemaker,J. (2002) Statistical issues with microarrays: processing and analysis. *Trends Genet.*, **18**, 265–271.

Nigam,K., Lafferty,J. and McCallum,A. (1999) Using maximum entropy for text classification. *In IJCAI-99 Workshop on Machine Learning for Information Filtering*, Stockholm, Sweden, 61–67.

Novichkova,S. *et al.* (2003) MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*, **19**, 1699–1706.

Pradhan,S., Ward,W., Hacioglu,K., Martin,J. and Jurafsky,D. (2004) Shallow semantic parsing using support vector machines. *In Proceedings of NAACL-HLT*. Boston, MA.

Ray,S. and Craven,M. (2001) Representing sentence structure in hidden Markov models for information extraction. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Seattle, WA, 1273–1279.

Ribeiro-Neto,R.B.-Y.a.B. (1999) *Modern Information Retrieval*. Addison Wesley, New York.

Schmid,H. (1994) Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, pp. 44–49.

Shah,P.K. *et al.* (2005) Extraction of transcript diversity from scientific literature. *PLoS Computat. Biol.*, **1**, e10.

Shatkay,H. and Feldman,R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.*, **10**, 821–855.

Tan,C.M. *et al.* (2002) The use of bigrams to enhance text categorization. *J. Inform. Process. Manage.*, **30**, 529–546.

Thanaraj,T.A. *et al.* (2004) ASD: the alternative splicing database. *Nucleic Acids Res.*, **32**, D64–D69.

Wattarujeekrit,T. *et al.* (2004) PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, **5**, 155.

Yakushiji,A. *et al.* (2001) Event extraction from biomedical papers using a full parser. *Pac. Symp. Biocomput.*, 408–419.

Yeh,A.S. *et al.* (2003) Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*, **19** (Suppl. 1), i331–339.

Yiming Yang,X.L. (1999) A re-examination of text categorization methods. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkley, CA, pp. 42–49.

Zavolan,M. *et al.* (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.*, **13**, 1290–1300.