Chapter 19

Text Mining for Systems Modeling

Axel Kowald and Sebastian Schmeier

Abstract

The yearly output of scientific papers is constantly rising and makes it often impossible for the individual researcher to keep up. Text mining of scientific publications is, therefore, an interesting method to automate knowledge and data retrieval from the literature. In this chapter, we discuss specific tasks required for text mining, including their problems and limitations. The second half of the chapter demonstrates the various aspects of text mining using a practical example. Publications are transformed into a vector space representation and then support vector machines are used to classify papers depending on their content of kinetic parameters, which are required for model building in systems biology.

1. Introduction

Since the advent of written language scientific advances are communicated in the form of text-based scientific publications. One of the major aims of text mining (TM) in the life sciences is to transfer the text based information into databases for storage, easy accessibility, and further processing. Up to now, this information transfer is heavily dependent on human experts who curate biological information in the text and further map it onto database entities utilizing ontologies or controlled vocabularies. Despite the endless number of biological databases, most information is still contained within the wealth of scientific publications. The sheer volume of the documents makes automated systems for searching and indexing the contained information indispensable to aid the human curation effort.

Two terms often encountered in TM are "Information Retrieval" and "Information Extraction." Information retrieval relates to the task of finding documents with relevance to a prespecified search query. The query can be of arbitrary complexity (e.g., all documents related to systems biology, all documents

Michael Hamacher et al. (eds.), *Data Mining in Proteomics: From Standards to Applications*, Methods in Molecular Biology, vol. 696,
DOI 10.1007/978-1-60761-987-1_19,

that contain the terms "polymerase" and "DNA," etc.). As one can already deduce, information retrieval has a huge impact on all forms of information technology, e.g., search engines for the World Wide Web where a document would be considered a web page. Information extraction, on the contrary, is a type of information retrieval with the task of automatically extracting structured information from within unstructured documents. The structured information to be extracted has a well-defined domain (e.g., protein names, gene names, numbers, etc.).

In a perfect world scenario, an automated computerized system would combine the concepts of information retrieval and information extraction. A collection of scientific documents is searched regarding pre-defined criteria (e.g., all documents relevant to systems biology). Each scientific text of the sub-collection is parsed by the system and analyzed toward identification of biological entities (e.g., proteins, genes, chemicals, drugs, species, etc.), physicochemical entities (e.g., constants, rates, etc.), numerical entities (e.g., numbers), and relationships between them (e.g., reactions, interactions, processes, etc.). The found relationships are mapped onto existing database entities and accompanying information of the relationships is stored (e.g. binding constants, half-life data, reaction velocities, etc.). No human interaction would be necessary to extract these relationships and the system is able to populate such a database for any volume of documents (e.g., the whole of Medline.). Unfortunately, up to now several problems influence the quality of a system that would fulfill all these requirements without any human interaction. For a better understanding of the encountered obstacles, the following sections contain a closer look at specific tasks that are required for the implementation of such a system.

2. Specific Tasks Within Text Mining, Their Problems, and Limitations

2.1. Format Conversion

The first problem that one encounters when dealing with text documents is the digital format of the text itself. Automated TM almost always requires the underlying text to be in ASCII or related format (e.g., Unicode for more complex encoding). The biggest resource of biological knowledge is scientific publications. Full-text articles of these publications are often only distributed in PDF format. A straight-forward conversion from PDF to ASCII text is currently not possible without the loss of at least some information, which could prove critical in the assessment of the information contained within the document itself. Research in this field is currently conducted outside the field of life sciences which is natural, given that the roots of TM lie within the field of information technology. Apart from PDF documents, several

projects such as, for example, PubMed Central (http://www.pubmedcentral.nih.gov) focus on gathering full-text articles that do not violate copyright restrictions in XML format. XML format is ASCII based and offers on top a simple annotated structure within a document that can be easily parsed and further processed by a computer program (e.g., specific tags for titles, sections, etc.). With more and more publishers (e.g., BioMed Central (http://www.biomedcentral.com), Public Library of Science (http://www.plos.org), etc.) adopting an open access policy of their content, collections of full-text scientific articles, such as PubMed Central, are steadily growing, but the mass of information is still only available in formats that are not easily translated into a machine readable encoding.

2.2. Identification of Word and Sentence Boundaries

Despite the rudimentary structure within the XML formats that e.g., PubMed Central offers it is still a challenge for automated computerised systems to identify single word and sentence boundaries (often denoted to as *Tokenization*). The former is of importance while identifying entities of interest within the text, while the latter influences more the semantic and syntactic ambiguity while establishing relationships between identified entities. For example, an interaction of two proteins is most likely but not exclusively conveyed within the same sentence:

> " ... and it could be shown that *protein A* and *protein B* interact."
>
> vs.
>
> " ... as could be shown for *protein A*. An interacting partner, *protein B*, is similarly ..."

2.3. Part-of-Speech Tagging

Automated *part-of-speech (POS)* tagging of words in sentences is another field of research conducted in information technology. Here, the aim is to annotate words in a sentence or phrase with its corresponding part of speech (e.g., verb, noun, adjective, etc.). This annotation is of help while identifying entities and establishing relationships between them (e.g., identify nouns in the sentence, identify verbs that connect nouns, etc.). Tasks related and often processed together with *POS* tagging are *stemming* and *lemmatization*. While either are closely related, *stemming* reduces words to their word stem or root (e.g., "interacted" and "interaction" are reduced to "interact"), whereas *lemmatization* maps words to their lemma or base form (e.g., "good" is a lemma of "better").

2.4. Named Entity Recognition and Word Sense Disambiguation

The task of identifying entities in text is often denoted to as "Named entity recognition" (NER). Word sense disambiguation plays an important role in NER. Especially in the life sciences, words with several meanings often appear disproportional. Examples are words used for genes or proteins that resemble

English words in natural speech, such as, for example the Drosophila genes "decay," "off," "blue," etc., which might relate to a property of the gene but which would not be easily identified by an automated system as a gene. Another example would be the denomination of a gene that resembles another biological entity, e.g., a protein. These problems could be avoided with the establishing of a formal naming convention for all biological entities. Despite efforts in this direction (1, 2), it is still far from being complete, commonly accepted, or utilized (3). Automated systems for word sense disambiguation try to overcome such problems by taking, for example, *POS* tags into consideration to identify nouns in text, which does not necessarily help in mapping the found text entities onto existing database entities.

2.5. Identification of Relationships Between Entities

The aforementioned tasks and their individual limitations influence the identification of relationships between entities. Automated systems still struggle with semantic ambiguity that is often encountered in the English language. A sentence read by a human reader can have several different meanings, depending on where the reader puts the stress within the sentence. Such sentences are generally difficult for computer software to analyze. It becomes even more difficult when relationships among entities span through several sentences. Even trained human curators with a sufficient biological background are not able to fulfill the task with a 100% accuracy.

3. Biomedical Ontologies `and Text Mining

Ontologies are foremost conceptual models. They try to establish a unifying representation and systematical order for entities, concepts, and relationships between them in a hierarchical manner for unambiguous and consistent sharing of knowledge over different domains. The Open Biomedical Ontologies (OBO, www.obofoundry.org) initiative is a collaborative effort to create guidelines for the development of biomedical ontologies. In addition, it gives an overview of biomedical ontologies currently under development. An example for a biomedical ontology is the well-known and studied Gene Ontology (2) (GO, www.geneontology.org). An example of an early TM system that focuses on the GO is GoPubMed (4) (www.gopubmed.org), which categorizes results of a PubMed (www.ncbi.nlm.gov/pubmed) search based on GO terms and concepts, thus letting a possible user navigate abstracts through these categories rather than through a list of, e.g., authors or publication titles.

One of the main criticisms of ontologies and their application in the biomedical domain is that an ontology will always be an unfinished product that can be improved and that they often do not follow stringent standards (5, 6). In addition, the creation and the research of ontologies were not driven by the need of controlled vocabularies with hindsight to biomedical TM. The main obstacles for the application of ontologies within the scope of biomedical TM are the nonstandardized ontology language, the earlier mentioned inconsistency in naming convention for biological entities and concepts, and the incompleteness of ontologies (7). Nevertheless, research into ontologies and their application within biomedical TM is currently of a huge interest and more and more TM systems are developed that rely on ontologies.

4. Examples of General Text Mining Systems in the Biomedical Domain

TM systems in the biomedical field can be categorized broadly into two categories: (1) general TM systems and (2) specialized TM systems. The former systems do not focus on a specialized field of biology and are capable of retrieving either documents or co-occurrences to a variety of biological questions. The main aims of a researcher in utilizing such a tool are twofold. First, to filter out documents of interest to a particular search question (e.g., "Retrieve all documents that contain a particular gene or protein" "Retrieve all documents where *protein A* occurs with another protein in the same sentence" etc.) and, second to find literature evidence for testing a hypothesis (e.g., "Does evidence in the scientific literature exist that *protein A* and *protein B* interact?" "Does evidence in the literature exist that the *drug X* is related to the *disease* Ƴ?" etc.). Examples of such tools are manifold. For example, *iHOP* (8) uses gene/protein names as hyperlinks between sentences and abstracts of the PubMed database. The TM system is gene/protein centered, which means that the starting point for utilizing the system is a gene/protein name. Based on the name, *iHOP* finds sentences in the literature that contain the gene/protein name with other genes/proteins, thus creating an easily searchable network around the input gene/protein linked to the underlying literature. *EBIMed* (9) also, based on PubMed abstracts, has the goal to present information about UniProtKB/Swiss-Prot proteins, GO annotations, drugs, and species found in the abstracts in the form of an easy accessible table. The advantage of this TM system is that the input query into the system can be of arbitrary complexity. Standard search queries that would be utilized to query PubMed directly can be used. The resulting set of abstracts is then analyzed toward the former mentioned biological concepts and the results are presented in the form of a table, where each entry is linked to the

underlying sentence and abstract where the information was found, as well as to biological databases for more information on the biological entity/concept. This table highlights all co-occurrences of biological entities and concepts found in the corpus of abstracts that was retrieved by the search query. The table can be ordered in manifold ways to satisfy the user needs. Another similar approach is the TM system *AliBaba* (10) that also works on PubMed abstracts. Based on a protein or disease, it creates a network in the form of a graph, which visualizes interacting concepts such as cells, compounds, diseases, drugs, enzymes, proteins, genes, species, and tissues mined from the PubMed abstracts. The extracted information is again linked to the underlying text source, which is made readily accessible to provide the means for the user to confirm the accuracy of the extracted associations by hand.

All these systems provide in essence a method to query a literature corpus and retrieve abstracts/sentences that match a prespecified search query. The results are presented in different formats, while the focus is on different biological entities. It is a quick way to find fast information about a biological entity of interest. The extracted information is linked to the text source, and in most cases, to other biological databases, which enables a user to verify by hand how much confidence he gives to certain extracted information.

The biggest downside of these TM systems is that they only work on PubMed abstracts. The wealth of information buried in the full-text articles is thus not considered at all. Many of the problems and limitation in TM systems mentioned above play a role for disregarding the full-text articles in the first place. The main reason for considering only the abstracts is their easy accessibility, which in case of PubMed can be obtained free of charge in XML format for public institutions.

5. Measuring Success

After a system for information retrieval or extraction has been developed, its performance has to be measured. This can, for instance, be done by calculating sensitivity and specificity, while in the field of information retrieval, more often recall and precision are used. To make things even more confusing, sometimes also the positive, and, respectively negative predictive values are used to characterize a classifier.

The connection between all these terms is displayed in Fig. 1 for a binary classification problem. An example can in reality be true or false and the classifier can give a positive or a negative result, leading to four possible outcomes. In two cases

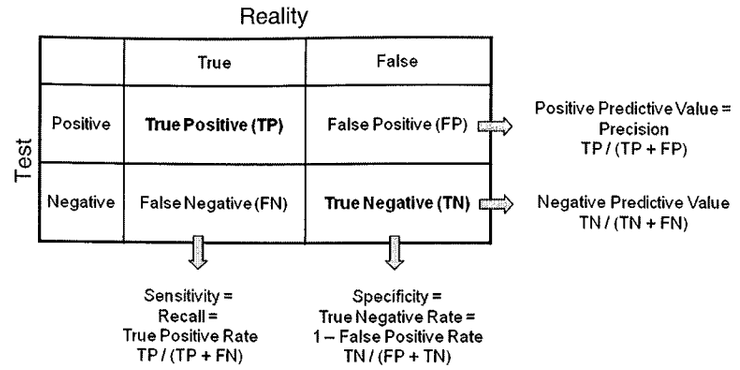

Fig. 1. Possible outcomes of a two-class classification problem. In the different scientific communities, different measures for classification success are used. Examples are the sensitivity and specificity system or the recall and precision system. For further details, see text.

(true positive and true negative), the prediction was correct, whereas in the other two cases (false positive and false negative), it was wrong. Sensitivity is now defined as the number of true positive predictions divided by all positive examples and specificity is the number of the true negative predictions divided by all negative examples. Thus, sensitivity measures how well a classifier recognizes true examples and specificity measures how well false examples are recognized. The term recall is actually identical to sensitivity, while precision is identical to the positive predictive value. Thus, precision is the fraction of positive predictions that are correct. A noteworthy feature of the sensitivity/specificity system is its independence of the ratio of true to false examples. Precision, in contrast, does vary with sample composition.

A specific pair of sensitivity and specificity values often depends on the discrimination threshold used by the classifier, and thus a single classifier can produce a whole range of sensitivity/specificity pairs. Consider, for example, the PSA level that is used in prostate cancer diagnostics. If the threshold level used for positive classification is low, the test will generate many positive predictions but with a high error rate, i.e., sensitivity is high, but specificity is low. If, however, a high threshold is used, there will be only few positive predictions but most of them will be correct. This means sensitivity is low, but specificity is high. To compare classifiers, it is, therefore, not sufficient to compare single sensitivity/specificity values, but instead the whole range of generated values has to be considered. A convenient way to do this is the use of a receiver operator curve (ROC), which displays true positive rate (sensitivity) as function of the false positive rate (1-specificity). The area under the receiver operator curve (AUC) ranges from 0 to 1 and is a popular measure for the quality of a classifier. For a more in-depth discussion on the use of ROCs see (11).

6. An Example of Text Mining for Systems Biology

As a specific example, the problem of finding scientific publications that contain kinetic parameters is now described. Biochemical reaction systems are usually modeled by a set of ordinary differential equations (ODEs) that describe the changes in the concentration of a biochemical species. The rate of a reaction is a function of the concentrations of the substrates, products, and of kinetic parameters that are part of the kinetic law. The irreversible Michaelis–Menten kinetics is a simple kinetic for the case that one substrate, with concentration c_S, is irreversibly converted into a product:

$$v = \frac{V_{\max} \cdot c_S}{K_M + c_S}$$

$V_{\max}$ denotes the maximal rate for high substrate concentrations and K_M is the half-saturation concentration (Michaelis–Menten constant). Other, more complicated, kinetic laws exist that depend on further parameters such as half-life and activation, respectively, inhibition constants. For the quantitative modeling of biochemical reaction networks, it is important to know the values of the various parameters and to know to which kinetic type they belong. Whereas most reaction networks are well described qualitatively, detailed quantitative values are missing or scattered in various scientific publications.

The aim was, therefore, to build a classifier that could separate few publications that contain values for kinetic parameters from those that do not (see also (12)). For this purpose, 4,582 randomly chosen full-text documents were downloaded from 12 different journals. From the full set, a keyword search generated 791 candidate articles. The keywords consisted of names and identifiers of constants (such as "Michaelis–Menten" or "Km") and words describing functions (such as "degradation," and "activation") or components ("enzyme"). Reading those 791 documents revealed that only 155 actually contained kinetic parameters, corresponding to a precision of 20% of this method. However, this first selection step was necessary, because it would have been a prohibitive amount of work to read all 4,582 articles.

6.1. Document Representation

The representation most often used for the application of certain machine learning techniques is the vector space model (VSM) (13). This model describes each document as a set of properties called features. This leads to a comparable representation of texts, regardless of their prior format, size, or structure (book, journal, article, and paragraph). It becomes irrelevant whether the information is presented in the Results or the Methods section of a research article, or what the exact content is (e.g. differences in nomenclature usage or spelling variants). Another advantage is the suitability of such vector formats for machine learning techniques, which can easily gather hints on the importance and influence of a particular fact (a feature) or certain nonlinear combinations of those.

Representing documents using the VSM, a fixed vector of features observed in the entire document collection (a feature vector) is calculated. Next, for each single document, an instance of this feature vector is filled with values describing the relevance of each feature for this particular document. Some features or properties might be present (to some degree) in one document, but absent in others. A single document can contain a certain term, with a certain number of occurrences, or not. The corresponding coordinate in the document vector, an instance of the feature vector, is assigned a value reflecting this occurrence, that is, the term frequency (*tf*). After tokenization and stemming of the texts, a fixed feature vector can be extracted consisting of every word stem encountered. Instances of the feature vector are then filled with the corresponding occurrences of each term for this particular document, resulting in one document vector per publication. The underlying approach is called a bag-of-words, as all words are represented by their frequency only, regardless of co-occurrences, collocations, and context. Additionally, one might think of different weighting schemes to represent the significance of a term for describing a certain document. Most weighting schemes (14, 15) comprise a combination of a term's local weight (i.e. within the document) and its global weight (i.e. in the document collection). However, in this study, only *tf* was used to construct the feature vector. Processing of the complete corpus (791 documents) resulted in approximately 44,000 different features.

6.2. Feature Ranking and Dimensionality Reduction

The described way to represent documents leads to a very high dimensional feature vector. These extreme dimensionalities can negatively affect the classification performance. On the contrary, one can argue, that the more information is used to describe the documents, the better will be the classification model generated by the machine learning algorithms. It is, therefore, an important step to find an appropriate balance between these opposing effects. To pick the most relevant features of a document (or the whole document collection), different ideas were applied. In every language, there are a lot of so-called stop-words, common terms which do not provide any information toward discriminating documents, as they tend to appear with the same frequency in every kind of text (e.g. and, are, it, and with). These words can be removed, as well as very rare words, appearing in only a few (or a single) documents. A pruning of such words helps to reduce the dimensionality of the vector space.

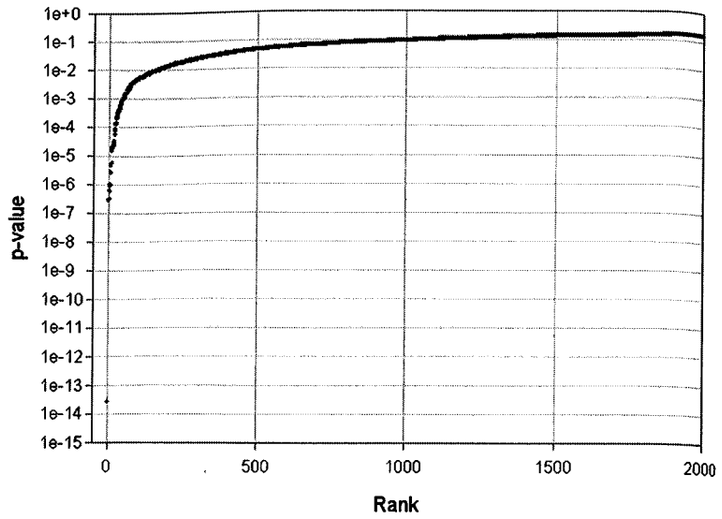

Fig. 2. Sorted results of the nonparametric Mann–Whitney test used to rank all words in the feature list obtained from the analysis of the document corpus (796 scientific papers). From the 44,000 features, only 532 have a *p*-value smaller than 0.05.

Furthermore, the remaining features can be ranked according to their importance by some appropriate statistical test and then only the most important terms are used for the classification algorithm. To calculate such a ranking, the non-parametric Mann–Whitney test was used, which does not rely on special assumptions about the data distribution (such as normality). The test calculates for each of the approximately 44,000 different features a *p*-value, indicating how important this feature is for separating the two classes. Figure 2 shows the *p*-values for the 2,000 most significant features. There are only relatively few features with small values, while the large majority of terms seems to be evenly distributed between the two classes of documents (resulting in large *p*-values). Although we perform multiple tests (namely, 44,000), corrections for multiple testing are not required since we are only interested in the relative ranking and not the absolute significance of each term.

6.3. Classification Performance and Feature Number

Several classification runs were performed to study the dependency between feature number and classification performance. For this purpose, only a certain number of top ranked (Mann–Whitney) features were included in the support vector used by a support vector machine (SVM) (16). Classification was performed with RBF (radial basis function) kernel and tenfold cross-validation to avoid over-fitting. Figure 3 shows the connection between the area under the receiver operator curve (AUC) and the number of used features. As can be seen, the AUC rapidly increases

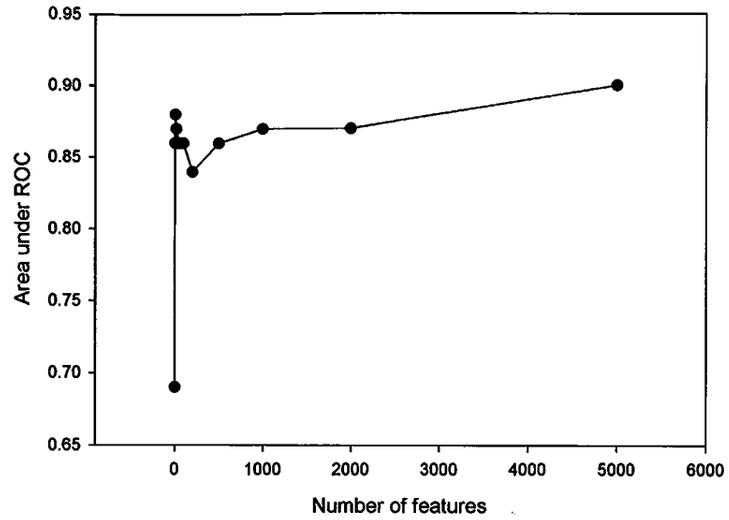

Fig. 3. Diagram showing the dependence between the area under the ROC curve (*AUC*) and the number of best ranked features used for classification. The features were ranked using a Mann–Whitney test (see Fig. 2).

with increasing feature number and then approaches a maximum at 5,000 features (which was the number of features given as input to the SVM). Thus, in this case, already a small number of top-ranked features are sufficient to give a good classification performance. Furthermore, the classification ability of the SVM does not degrade with feature number (it even seems to increase slightly). This confirms the well-known observation that the performance of SVMs is quite robust against a surplus of features.

6.4. Classification Performance with 5,000 Features

Finally, the classification performance is examined when using a feature vector with 5,000 features, which gave the best AUC value of the studied cases. Figure 4 shows the ROC curve for this situation with an AUC of 0.90.

Support vector machines can provide a probability estimate on how likely it is that an example belongs to one class or the other. By using different probabilities as threshold for the classification (normally 0.5 is used), different combinations of sensitivity (true positive rate) and specificity (1-false positive rate) can be obtained. All points on the surface of the ROC can be reached by an appropriate choice of the classification threshold.

Another way to visualize this connection is displayed in Fig. 5. The diagram shows directly how sensitivity and specificity vary with the used threshold. In general, there is a trade-off between sensitivity and specificity. However, depending on the problem, it might not be necessary to have high values for both measurements. In our case, sensitivity is not as important. Since a potentially

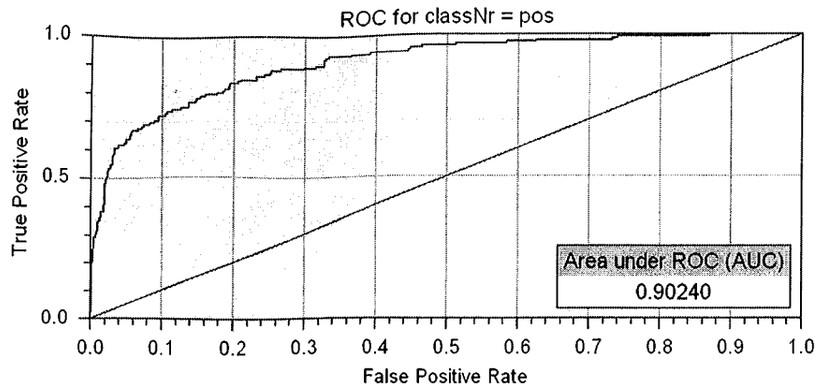

Fig. 4. Receiver operator characteristic (*ROC*) curve for a support vector machine classification using a feature vector with the 5,000 top-ranked features.

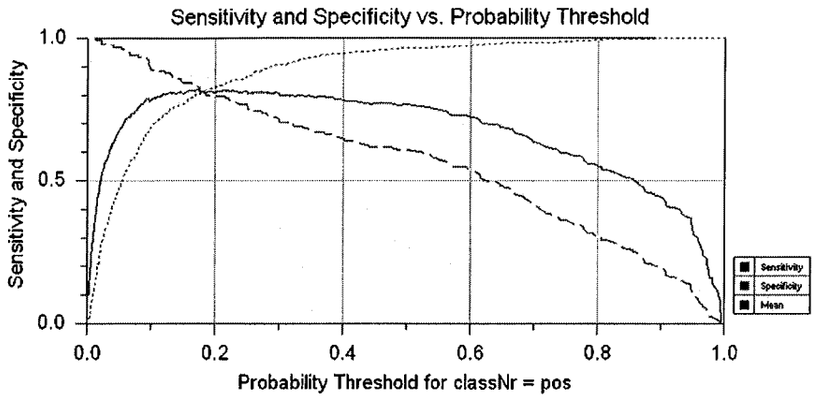

Fig. 5. Achieved sensitivity and specificity (and geometric mean of both) as a function of the used probability threshold. Support vector machines can calculate a probability value indicating how "sure" the classifier is that the example belongs to the predicted class. If different probability thresholds are used for classification, different combinations of sensitivity and specificity are obtained.

very large number of publications with kinetic parameters does exist in the literature, it is not so important if one is not found (false negative). But false positives are very costly, because those papers have to be inspected manually before the error is detected (labor costs). Therefore, a high specificity is desirable. That means a large threshold value will be chosen to obtain a high specificity.

7. Conclusions

Interest and research in biomedical TM has increased greatly over the last decade. Currently, information retrieval and extraction provide the means to support a variety of biomedical studies.

An example study of a TM approach set in the field of systems biology has been described. The aim was to train a machine learning classifier to distinguish relevant from irrelevant scientific publications. The relevance is defined by their content of kinetic parameters that are necessary for the in silico modeling of biological pathways. Several TM sub-tasks such as format conversion, POS tagging, stemming, feature representation with the help of the vector space model approach, and machine learning. have been discussed during the analysis. It could be shown that with the help of TM techniques it was possible to fulfill the task with an acceptable performance. However, several difficulties were encountered during the course of the study. The automatic conversion from PDF documents to plain ASCII text was imperfect. The used software was not able to resolve all words and symbols encountered in the PDF documents correctly. Future advances in conversion technology and optical character recognition (*OCR*) software will definitively improve PDF-based TM. Shifting the focus away from PDF documents toward full-text publications in HTML or XML format would solve this problem. An example of such a format is ePub, which has in 2007 been endorsed by the International Digital Publishing Forum (www.idpf.org) as a new standard for electronic publishing. Furthermore, other feature representation schema or machine learning algorithms might lead to improvements as well. However, even though the created system for the automatic classification of documents from a specialized biological domain is not perfect, it could be demonstrated that such a system can already now be of great value for scientists seeking kinetic information from text sources.

References

1. White J, Wain H, Bruford E, Povey S (1999) Promoting a standard nomenclature for genes and proteins. Nature 402(6760):347
2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29
3. Chen L, Liu H, Friedman C (2005) Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics 21(2):248–256
4. Doms A, Schroeder M (2005) GoPubMed: exploring PubMed with the gene ontology. Nucleic Acids Res 33:W783–W786 (Web Server issue)
5. Soldatova LN, King RD (2005) Are the current ontologies in biology good ontologies? Nat Biotechnol 23(9):1095–1098
6. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W et al (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25(11):1251–1255
7. Spasic I, Ananiadou S, McNaught J, Kumar A (2005) Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinform 6(3):239–251
8. Hoffmann R, Valencia A (2004) A gene network for navigating the literature. Nat Genet 36(7):664
9. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P (2007) EBIMed-text crunching to gather facts for proteins from Medline. Bioinformatics 23(2):e237–e244

10. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U (2006) AliBaba: PubMed as a graph. Bioinformatics 22(19):2444–2445
11. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8):861–874
12. Hakenberg J, Schmeier S, Kowald A, Klipp E, Leser U (2004) Finding kinetic parameters using text mining. OMICS 8(2):131–152
13. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620
14. Strasberg HR, Manning CD, Rindfleisch TC, Melmon KL (2000) What's related? Generalizing approaches to related articles in medicine. Proc AMIA Symp 838–842
15. Glenisson P, Antal P, Mathys J, Moreau Y, De Moor B (2003) Evaluation of the vector space representation in text-based gene clustering. Pac Symp Biocomput 391–402
16. Vapnik VN (1995) The nature of statistical learning theory. Springer, Berlin

Chapter 20

Identification of Alternatively Spliced Transcripts Using a Proteomic Informatics Approach

Rajasree Menon and Gilbert S. Omenn

Abstract

We present the protocol for the identification of alternatively spliced peptide sequences from tandem mass spectrometry datasets searched using X!Tandem against our modified ECgene resource with all potential translation products and then matched with the Michigan Peptide to Protein Integration (MPPI) scheme. This approach is suitable for human and mouse datasets. Application of the method is illustrated with a study of the Kras activation-Ink4/Arf deletion mouse model of human pancreatic ductal adenocarcinoma.

1. Introduction

By means of alternative splicing and posttranslational modifications, one gene can generate a variety of proteins. Alternative splice events that affect the protein coding region of the mRNA will give rise to proteins which differ in their sequence and activities. Alternative splicing within the noncoding regions of the RNA can result in changes in regulatory elements, such as translation enhancers or RNA stability domains, which may dramatically influence protein expression (1).

Alternative splicing has been associated with such diseases as growth hormone deficiency, Fraser syndrome, cystic fibrosis, spinal muscular atrophy, and myotonic dystrophy (2, 3). In cancers, there are examples of every kind of alternative splicing, including alternative individual splice sites, alternative exons, and alternative introns (4). A number of public alternative splice databases have recently become available, including ASD, HOLLYWOOD, and ASAP II. Each of these repositories contains transcript models that have been constructed from either expression data

Michael Hamacher et al. (eds.), *Data Mining in Proteomics: From Standards to Applications*, Methods in Molecular Biology, vol. 696,
DOI 10.1007/978-1-60761-987-1_20,