Text Mining for Systems Biology Using Statistical Learning Methods

Sebastian Schmeier[†], Jörg Hakenberg[‡*], Axel Kowald[†], Edda Klipp[†], and Ulf Leser[‡]
[†] Max-Planck-Institute for Molecular Genetics, Berlin, Germany
[‡] Humboldt-University Berlin, Germany
[*] Corresponding author, Dept. Knowledge Management in Bioinformatics, Humboldt-University,
Rudower Chaussee 25, D - 12489 Berlin, Phone: +49.30.2093.3903, eMail: hakenberg@informatik.hu-berlin.de

Abstract

The understanding and modelling of biological systems relies on the availability of experimental results measuring chemical and physical properties and dynamic changes of components in these systems. As of now, this data and results is merely published in free text form in scientific journals, which is unsatisfactory for the search and retrieval of specific information. No individual nor a group is able to keep up with the huge amount of input coming from new and old publications. The gathering of documents relevant to any field of kinetic modelling and the extraction of needed data has to be supported by automated processes. This work describes first steps towards the automatic recognition and extraction of kinetic parameters from full test articles. We describe the processing of full-text publications by text mining methods to classify the texts regarding their relevance to kinetic modelling. Using support vector machines as classification basis, we were able to improve the precision of the process by a factor of 2.5 compared to a manual selection and processing of articles.

EXTENDED ABSTRACT

1 Motivation

The emerging field of systems biology aims at understanding and modelling biological systems at the molecular level [Kitano, 2002]. In contrast to previous approaches, systems biology aims at quantitatively model and simulate complex biological processes comprised of thousands of chemical components and reactions. Therefore, one has to gain a very deep understanding of structure, dynamics, control, and design methods of these systems. Insight into the dynamics of a system aims at the construction and usage of models for further studies and analysis. The kinetic modelling of biological systems depends on sets of different kinetic data and values measured in laborious and expensive experiments. Such data is continuously published in thousands of scientific articles that can hardly be overlooked by individuals. However, to build a concrete model, e.g. for a certain metabolic pathway of a given species, one needs only a selected subset of kinetic parameters which is very likely to be found in only a few papers - but finding those is a challenge. To build up a model one has to gather, sort out, read, and understand a large number of publications.

Our project aims at the generation of a database containing the exact information for a multitude of species. Manual retrieval and inspection of a led to unsatisfying results regarding time and precision of the method, as less than 20% of the articles found by a simple text search performed with a set of characteristic keywords contained relevant information. To achieve a comprehensive data set in an efficient manner, we research methods for the semi-automatic recognition and extraction of kinetic data from full text articles. The problem was divided into the retrieval and the extraction of publications relevant to kinetic modelling. In this paper we give first results for the information retrieval step. Its goal is to identify appropriate documents obtainable from online journals by using text mining methods. To this end, we implemented and tested different methods for natural language processing, text processing, and text classification. A number of combinations were evaluated wrt. their individual strength and the overall performance of the classification process. Our first experiences are encouraging and already resulted in a drastic reduction of the manual work necessary to filter out irrelevant documents.

2 Methods

Prior to this project, publications were gathered by keyword based queries using the PubMed [Wheeler *et al.*, 2003] interface to MEDLINE®with a subsequent manual inspection of each individual article. As a first step towards automatic text classification, we implemented a tool for randomized access of articles contained in a given set of online journals focusing, at least in parts, on the topics of kinetic modelling. From the full set, candidate articles were selected by using a keyword search. The keywords consisted of names and identifiers of constants (such as 'Michaelis-Menten' or 'Km') and words describing functions ('degradation', 'activation') or components ('enzyme').

Using this method, we gathered a collection of 4582 papers of which 797 contained at least one of the given keywords. Reading each of these papers over a period of several months, we found that only 155 of them actually contained appropriate kinetic data, leaving 642 which had to be read but revealed no useful information whatsoever.

We aimed at improving this process through text mining methods. We took the set of 797 manually annotated publications (either *positive* or *negative*, depending on their relevance) to train and test different text mining methods. First, we fixed three data sets for training, testing, and final validation, consisting of 400, 200, and 197 publications, respectively.

The pre-processing of the stored documents (full-text articles in PDF-format) included the *conversion* from PDF to plain text, *tokenization*, *tagging*, and *lemmatizing* of

each document (pdftotext [Noonburg, 1996], TreeTagger [Schmid, 1994]).

In order to get to a comparable representation of different documents, we chose the *vector space model* approach [Salton, 1983]. A fixed feature vector is chosen by taking into account all terms contained in the document base. The dimension of the feature vector is determined by the number of terms. An instance of this vector is used for the representation of each document, where the weights for each term appearing in the document are stored as its coordinates. Two different texts would result in two unique vector representations, where a similarity measure can be defined very easily on.

The performance of any text classifier strongly depends on the selection of an appropriate and fixed feature vector used for the representation of all documents. As we found the 600 articles from the training and test set to contain more than 100.000 different terms, we decided on two additional steps to identify the terms which would be most suitable at discriminating relevant from irrelevant articles. First of all, we excluded all words with only one appearance in the document collection. This led to a vector size of 66.616 words. This particulary removed many artificial terms generated erroneously by the PDF-to-text-converter. To cope with the problem of overfitting, where two classes are separable too easily because of the high dimensionality of the underlying feature vector, we applied Student's *t-statistic* [Gosset, 1908] to rank and select the most convenient terms. The *t-values* for each term occurring in both the positive and the negative training set state their ability to discriminate both classes.

For each document in the training set of 400 articles, we calculated the weights for each term in the fixed feature vector. For the local weight, the *term frequency* was used. The relative measure for the global weighting of terms was computed as the *inverse document frequency*. The resulting *tf·idf*-values where then applied to train the *Support Vector Machine* [Vapnik, 1995] SVM^{light} [Joachims, 1998].

In order to find the best set of parameters for the Support Vector Machine, an optimization step was necessary. The aim was to improve precision and recall. Several parameters can be optimized. The vector length is one of these. We varied this parameter from length 50 to length 400 in steps of 50 words and from 400 to 1000 in steps of 100 words. Furthermore, we experimented with both linear kernel and the polynomial kernels from degree two and three. Finally, two kernel parameters, the *c*-value and the *j*-value were investigated. The *c*-value penalizes misclassifications. It is the pre-factor of the sum of the distances from the misclassifications to the hyperplane. A weighting of misclassifications into the classes can be parameterized by altering the *j*-value. While searching the optimal values for all these parameters, SVM^{light} was run over 8000 times with different settings (data not given in detail).

The method is shown graphically in Figure 1.

3 Results

We trained the support vector machine using the training set and optimized the parameters using the test set. For every combination of kernel, kernel parameters and feature vector length, a new model was learned and precision and recall calculated with the predictions on the test set. The best results for the linear kernel could be observed with a feature vector length of 400 words. Polynomial kernels (degrees 2 and 3) achieved their best classifications in a 150-dimensional vector space.

kernel, degree	c, j	dim	precision & recall
linear	.1, 1	400	100% & 63.16%
polynomial, 2	.1, .8	150	100% & 52.63%
polynomial, 3	.0003, 3	150	100% & 39.47%

All these figures reflect the results of classifying the test set of 200 articles, which had an influence only on the composition of the feature vector. When evaluating our method using the validation set never touched or looked at before, precision and recall degraded immensely. Precision degraded to 50.0%, while the recall reached 30.77%. The accuracy on this data set was 80.20%. Despite these low figures, the text classification method promises to be a considerable advance since the number of superfluously read papers decreases by a factor of 2.5 compared to the simple keyword search mechanism. However, this improvement has to be further verified since the current figure depend on a preselected data set containing certain keywords.

4 Discussion

Recently, text mining has attracted great attention in the biomedical research community. Several studies found these approaches to be very helpful in assisting information extraction and knowledge discovery (see e.g. [Rindflesch *et al.*, 1999] and [Blaschke *et al.*, 1999]). However, most of these studies try to classify texts for describing genes or the function of gene products, which is quite a different problem. We are not aware of any other attempts to apply text mining for the detection and extraction of parameters for kinetic modelling.

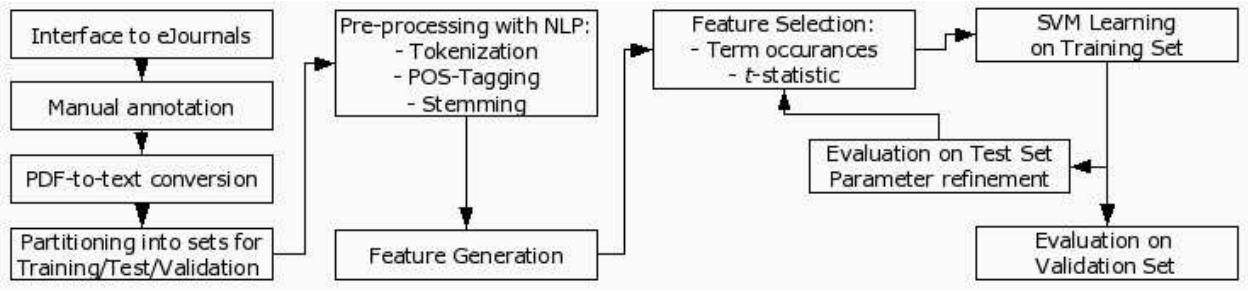

Figure 1: Document retrieval, classification, and validation

Our goal in this project is the building of a database holding kinetic data for various species. One comparable database available is BRENDA [Schomburg *et al.*, 2002], which is manually curated and provides to few data on only a small set of species. Especially, data on yeast enzymes is included only to a very low extent. On a complete data set, models can be designed and then represented and exchanged using the *Systems Biology Markup Language* (SMBL) [Hucka *et al.*, 2003].

Despite the low figures for recall and precision, our results are encouraging given our particular application. In our current project phase, a low recall is not as dramatic as it may appear, as there is currently little hope to catch all relevant texts. Much more important is the precision of the process, as it determines the amount of time a human reader has to waste on irrelevant documents. Our classifier enhances the precision by a factor of 2.5 compared to the simple keyword selection method, saving months of work for the biologists.

However, compared to other studies in text mining for biomedical applications, our figures for recall and precision both for the test and the validation set are surprisingly low. This may be explained by different reasons. One would be the information content of the texts on systems biology itself. For instance, the publications are more diverse than in other fields, as kinetic modelling parameters can appear in very different types of work which are difficult to characterize uniformly. Furthermore, kinetic data is not only presented in continuous text, but very often in tables and figures which are hard to translate and hard to capture for parsers. Additionally, we are not interested in papers dealing with theoretical aspects of kinetic modelling, where all names of constants and differential equations also appear, but without any real experimental data. Aside from the quality and quantity of the data sets, the text mining methods have to be examined as well, as our current tools where found to have problems in several regions. E.g., many pdf2txt tools have difficulties in correctly concatenating multi-column text, and our stemmer mistreated many scientific verbs and nouns.

Future work

The step directly ahead is the testing of other kernel functions, such as a radial basis or a sigmoid kernel. Additionally, we are currently implementing and optimizing a *naïve Bayes* classifier for comparing SVMs with other approaches.

Further on, the different parameters altering our processes have to be evaluated more systematically. The final choices for the c- and j-values and the feature vector length worked best for this topic and for the given data set sizes. We have no understanding yet of how robust these parameters are with respect to changing data sets. The size of the data set itself will be a subject of variation in the further project. Furthermore, we plan a systematic analysis of the effect of changing the training data.

Another point is the possible weighting of terms specific to the field (in this context, common keywords from kinetic modelling data), either manually or automatically, to improve the classifiers predictions.

As the annotated documents accumulate, we furthermore hope to improve the performance and accuracy of the model learned by the different classifiers by shere extension of the size of the training data.

Acknowledgements

This work is supported by the Bundesministerium für Bildung und Forschung, BMBF.

References

[Blaschke *et al.*, 1999] C. Blaschke, M.A. Andrade, C. Ouzounis, and A. Valencia. Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol*, pages 60–67, 1999.

[Gosset, 1908] William Sealy Gosset. The Probable Error of a Mean. *Biometrika*, 6:1–25, 1908.

[Hucka *et al.*, 2003] M. Hucka, A. Finney, H.M. Sauro, H. Bolouri *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, Mar 1 2003. http://www.sbml.org.

[Joachims, 1998] Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning*. Springer, 1998. http://www.svmlight.joachims.org.

[Kitano, 2002] Kiroaki Kitano. Systems Biology: A Brief Overview. *Science*, 295:1662–1664, March 1 2002.

[Noonburg, 1996] Derek B. Noonburg. pdftotext ©1996-98. 1996. http://www.aimnet.com/∼derekn/xpdf/.

[Rindflesch *et al.*, 1999] Thomas C. Rindflesch, Lawrence Hunter, and Alan R. Aronson. Mining molecular binding terminology from biomedical text. *Proc AMIA Symp*, pages 127–131, 1999.

[Salton, 1983] Gerard Salton. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[Schmid, 1994] Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, 1994. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

[Schomburg *et al.*, 2002] Ida Schomburg, Antje Chang, and Dietmar Schomburg. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.*, 30(1):47–49, Jan 1 2002. http://www.brenda.uni-koeln.de.

[Vapnik, 1995] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[Wheeler *et al.*, 2003] D.L. Wheeler, D.M. Church, S. Federhen, A.E. Lash *et al.* Database resources of the national center for biotechnology. *Nucleic Acids Res.*, 31(1):28–33, Jan 1 2003. http://www.ncbi.nlm.nih.gov/-entrez/query.fcgi?db=PubMed.