# 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

Keywords: text mining, vector space model, support vector machine, systems biology, kinetic modelling.

## Abstract

The understanding and modelling of biological systems relies on the availability of numerical values for physical and chemical properties of biological macro molecules. Kinetic parameters, rate constants, specificities and half-lifes are examples of those properties. This data is mostly published in free text form in scientific journals, which is unsatisfactory for the automatic 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 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 text 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 keyword-based selection of articles.

## 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 modelling and simulating 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 of these systems. Insight into the dynamics of a system aims at the construction and usage of models for further studies and analysis.

For a biochemical reaction system it is practice to use a set of *ordinary differential equations* (ODEs) to describe the changes in the concentration of a biochemical species. The rate of a reaction is a function of the concentrations of the substrates and products of the reaction and of parameters. These parameters may be the concentrations of effectors as well as kinetic constants. The actual expression for a rate depends on the experimental knowledge about the kinetic characterization of a reaction and, partially, on the modelling purpose. Typical expressions for reaction rates are different forms (e.g. reversible or irreversible) of linear kinetics, *Michaelis-Menten*-kinetics, or *Hill*-kinetics. For the quantitative modelling 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 characteristics are missing.

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, focusing on *Saccharomyces cerevisiae*. Manual retrieval and inspection led to unsatisfactory 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 developed methods for the semi-automatic recognition and extraction of kinetic data from full text articles. The problem was divided into two separate steps, i.e., the retrieval of publications relevant to kinetic modelling and the extraction of concrete parameters. In this paper we give first results for the information retrieval step only. Its goal is to identify appropriate documents in a given collection 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 with respect to 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 797 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 processing of the stored documents (full-text articles in PDF-format) included the *conversion* from PDF to plain ascii text as a first pre-processing step (PDFTOTEXT [Noonburg, 1996]). *Tokenization* tries to identify the components of texts, i.e. single words. Word boundaries have to be defined, like white spaces, punctuation, brackets and so on. The final *lemmatizing* of each word is achieved using the Part-of-Speech-Tagger TREETAGGER [Schmid, 1994]. A further reduction could be obtained by predicting the stem (or root) of each word, where homographs often would result in the same stem, loosing semantic information.

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 weighted term frequencies are stored as its coordinates (see below). 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; Ewens and Grant, 2001] to rank and select the most convenient terms:

$$t = \frac{|\overline{X} - \overline{Y}|}{\sqrt{\frac{{S_1}^2}{m} + \frac{{S_2}^2}{n}}}$$

The *t-values* for each term occurring in both the positive and the negative training set state their ability to discriminate both classes. Table 1 contains the 16 words with the highest t-value, and Figure 1 plots the distribution of *t*-values. There is a remarkably rapid drop of *t*-values in the first ≈2000 words, from where on the remaining 64.000 have a relatively constant, yet very small t-value. This range for instance contains common stop words such as 'and' (rank 40.395, t-value 0.5) or 'the' (rank 64.600, t-value 0.1). Those words are considered unappropriate for document discrimination.

| word | *t*-value | word | *t*-value |
|---|---|---|---|
| Km | 8.91 | taining | 5.14 |
| half-life | 8.66 | Michaelis-Menten | 5.13 |
| Vmax | 7.95 | Lineweaver-Burk | 5.09 |
| turnover | 6.75 | Degradation | 4.86 |
| enzyme | 5.85 | 7-fold | 4.86 |
| activity | 5.26 | Vmax/Km | 4.79 |
| radation | 5.24 | degrade | 4.68 |
| di- | 5.18 | enzymatic | 4.65 |

Table 1: The 16 words with the highest *t*-values. The list coincides well with what a expert user would use as keywords, except for the terms 'radation', 'taining', and 'di-', which are probably artifacts of the PDF converter.

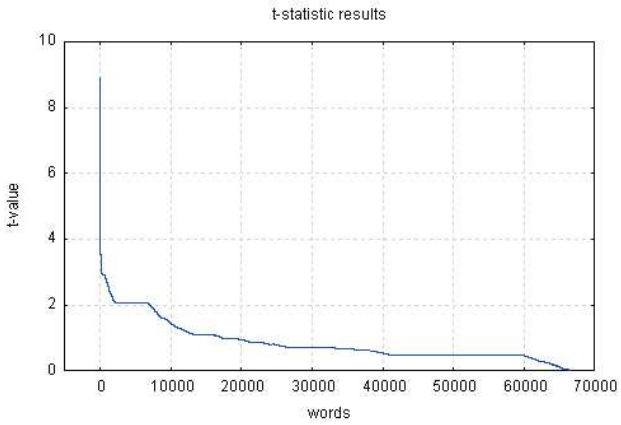


Figure 1: Distribution of the *t*-values from all 66.616 single words.

The word vector used for document representation was sorted by descending *t*-values, facilitating the reduction of dimensions by simply truncating the list after a certain number of words or a cut-off value for their *t*-values, both being parameterizable.

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*:

$$tf \cdot idf = (1 + \log(tf_{t,d})) \cdot \log(N/df_t),$$

where $tf_{t,d}$ is the frequency a term $t$ in document $d$, $df_t$ is the number of documents containing $t$, and $N$ is the total number of documents.

The resulting *tf·idf*-values where used 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

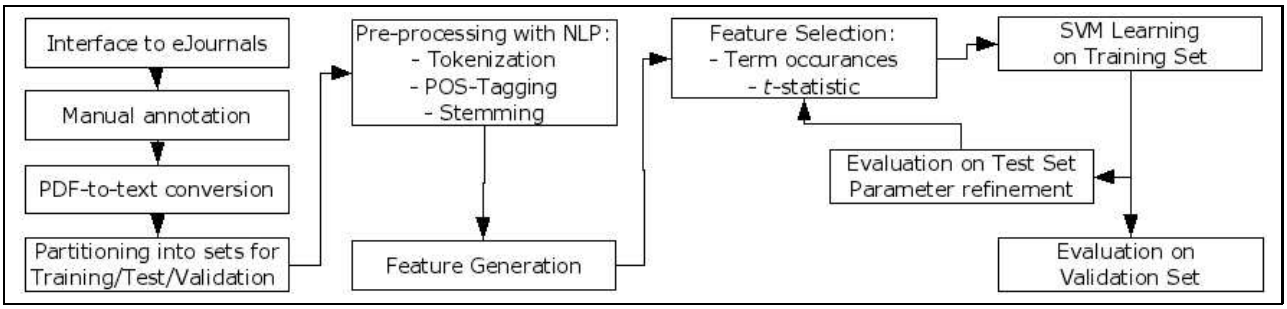


Figure 2: Document retrieval, classification, and validation.

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 polynomial kernels from degree two and three. Finally, different values for 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. The *j*-value can be used to apply different weights for the penalties for positive and negative examples. While searching the optimal values for all these parameters, SVM$^{light}$ was run over 8000 times with different settings.

The overall approach is shown graphically in Figure 2.

## 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 on the test set. The best results for the linear kernel could be observed with a feature vector length of 400 words, and values for $c$ and $j$ being 0.1 and 1, respectively (Figure 3.a-c). Polynomial kernels (degrees 2 and 3) achieved their best classifications in a 150-dimensional vector space, see table 2.

| Kernel, deg | c, j | dim | Precision, Recall |
|---|---|---|---|
| linear | 0.1, 1 | 400 | 100%, 63.16% |
| polyn., 2 | 0.1, 0.8 | 150 | 100%, 52.63% |
| polyn., 3 | 0.0003, 3 | 150 | 100%, 39.47% |

Table 2: Precision and recall for different kernels and various parameter settings. deg: degree of the polynomial kernel. dim: dimension of the word vector, i.e. the number of words.

The figures reflect the results of classifying the test set of 200 articles. When evaluating our method using the validation set (never touched or looked at before), precision and recall degraded considerably. 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 results are obtained on a biased data set, where all documents contain at least of a list of certain keywords.

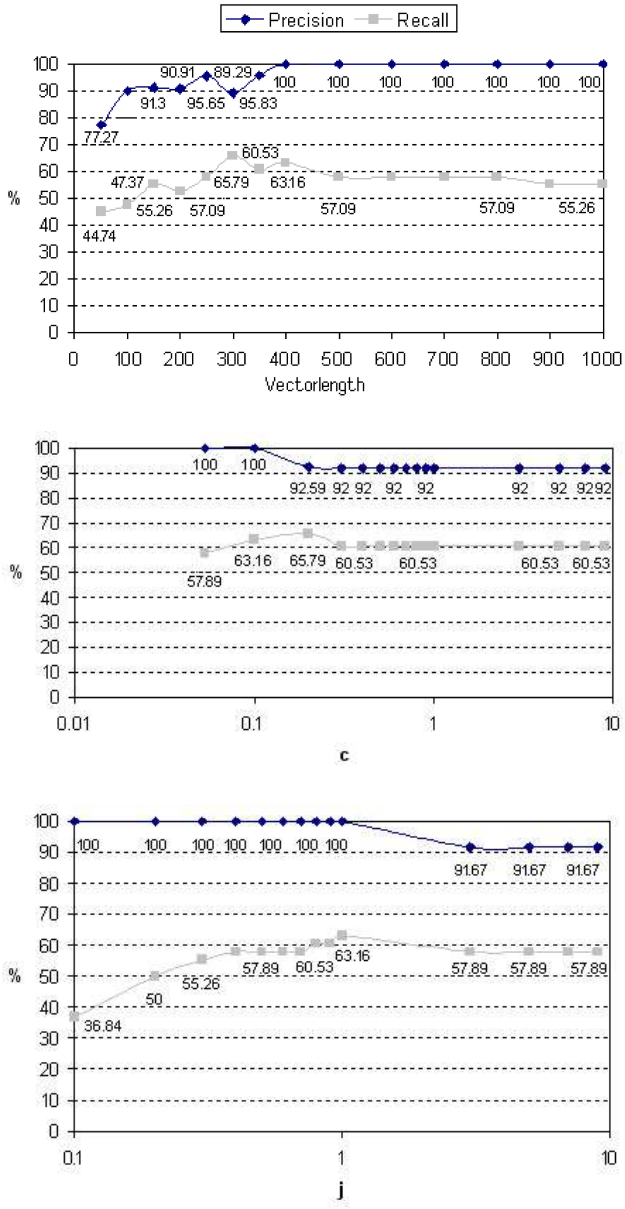


Figure 3: Precision and Recall plotted against varying word vector size (a) and kernel parameters c (b) and j (c).

## 4 Discussion

Recently, text mining has attracted great attention in the biomedical research community. Several studies found these approaches to be 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.

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 too 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 which we are currently investigating. One possible reason is the information content of texts on systems biology itself. Such 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 (e.g. papers on reactions of a particular enzymes, large-scale analysis of metabolism of a species, characterization of molecular reasons for diseases, etc.). 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 the relevant term might also appear, but without the corresponding numerical values. Furthermore, kinetic modelling can be pursued using a number of different modelling techniques which result in different types of reactions and different typical abbreviations of parameters. Our current classifier seems to only capture the most prominent type of model, i.e. the Michaelis-Menten Kinetics.

Another reason could lie in the specific techniques and tools we used. We found problems in several areas. One problem which exists with many pdf-to-text converters is an erroneous recognition and incorrect concatenation of multi-column text, which is a very common format for scientific publications. In some relevant documents used in the training set, problems occurred with hyphenations of the same (or a similar) word. This probably generated and lead to an apparently high t-value for words like 'radation', 'di-' or 'taining', as shown in table 1.

A further factor is the reduction of the word vector dimension by taking into account only the lemma of each word. An appropriate algorithm has to be found to predict the lemma to a word with a high reliability. Finding the right lemma strongly depends on correctly tagging sentences for analysis. Problems occurr when dealing with texts containing many unusual proper nouns, such as protein names. Most lemmatizers where trained on popular corpora (e.g. from news paper articles), and not on domain specific corpora. New word inventions, shortenings, and compoundings, a very common phenomena in life science publications, lead to further difficulties. The TREE-TAGGER used for tokenization and lemmatizing produced reasonable results in most cases, but had difficulties with names of chemical substrates containing brackets and hyphens.

### Related work

SVM-based approaches to information gathering in life sciences has been a major topic of recent research in text mining. [Stapley *et al.*, 2002] deals with the prediction of subcellular protein locations from literature. Each protein is represented by a term vector composed of a set of documents relevant to this protein. A sub-cellular location class is now described using the term vectors for the proteins present at this specific location, and binary classification models for 11 classes are learned using a SVM. At low recall levels for a random classifier, the precision is very high ($\geq$50%), but drops significantly for most of the classes for higher levels.

New and specific kernels have been designed as well, to define similarity measures for vector representations adopted to various problems. A method of classifiying proteins into families by homology detection was presented by [Leslie *et al.*, 2002] using a spectrum kernel. $ROC_{50}$ score plots (area under the receiver operating characteristic curve, up to the first 50 false positives) showed that for most of 33 protein families, the score is less than 0.4.

The prediction of signal peptide cleavage sites using a string kernel is discussed in [Vert, 2002]. For 3% of false positives, the method retrieved an average of 68% true positives, compared to 46% invoking a weight matrix method.

[Donaldson *et al.*, 2003] apply SVMs to the discovery of abstracts describing protein-protein interactions. Using a SVM with a decision boundary of zero, presicion and recall where both found to be at 92%, whereas a naïve Bayes classifier only reached 87%.

Note that most attempts ignore the full text of publications but consider their abstracts only. For some fields, this might be sufficient as enough information is provided in the abstracts. Other problems can possibly only be solved looking into the full texts, as is the case with kinetic data and models.

### Future work

Our main question at the moment is to find an explanation for the drastic reduction in precision and recall from the test to the validation set. It surely is an indication for overfitting. Remarkably, however, smaller feature vectors, i.e. consisting of less words taken into account for the text classification, resulted in a drop of precision and recall as well. We are looking forward to the results of a *leave-one-out* validation test, which will show the variances of our method.

We are approaching this problem from two directions. First, we are implementing a leave-one-out validation to

test the dependency of the model from the specific collections for training and test. Second, we will apply cross-validation for the parameter optimization. The final choices for the c- and j-values and the feature vector length worked best 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 are investigating different classification methods. First, we shall test SVM using other kernel functions, such as a radial basis or a sigmoid kernel. Additionally, we are implementing and optimizing a *naïve Bayes* classifier to be able to compare SVMs with other approaches to document classification. It is an open question whether specific classification algorithms would work differentially well in different domains. To study this question in a systematic fashion, we work towards a software library of algorithms for classification which will eventually enable us to quickly prototype text mining applications.

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 sheer extension of the size of the training data.

## Acknowledgements

This work is supported by the German Ministry for Education and Science (BMBF) under grant contract 0312705B.

## 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. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, pages 60–67, 1999.

[Donaldson *et al.*, 2003] Ian Donaldson, Joel Martin, Berry de Bruijn, Cheryl Wolting *et al.* PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4:11, 2003.

[Ewens and Grant, 2001] Warren J. Ewens and Gregory R. Grant. *Statistical Methods in Bioinformatics*. Springer Verlag, New York, 2001.

[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.

[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.

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

[Leslie *et al.*, 2002] Christina Leslie, Eleazar Eskin, , and William Stafford Noble. The Spectrum Kernel: A String Kernel for SVM Protein Classification. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 7, pages 566–575, 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.

[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.

[Stapley *et al.*, 2002] B.J. Stapley, L.A. Kelley, and M.J.E. Sternberg. Predicting the Sub-Cellular Location of Proteins from Text Using Support Vector Machines. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 7, pages 374–385, 2002.

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

[Vert, 2002] Jean-Philippe Vert. Support Vector Machine Prediction of Signal Peptide Cleavage Site Using a New Class of Kernels for Strings. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 7, pages 649–660, 2002.

[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.