Bayesian Sequence Learning For Predicting Protein Cleavage Points

Michael Mayo

Dept. of Computer Science, University of Waikato, New Zealand
mmayo@cs.waikato.ac.nz

Abstract. A challenging problem in data mining is the application of efficient techniques to automatically annotate the vast databases of biological sequence data. This paper describes one such application in this area, to the prediction of the position of signal peptide cleavage points along protein sequences. It is shown that the method, based on Bayesian statistics, is comparable in terms of accuracy to the existing state-of-the-art neural network techniques while providing explanatory information for its predictions.

1 Introduction

The amount of sequence data generated by experimental biologists and made available via Internet databases is growing at an increasing rate. For example, SWISS-PROT [2], the leading protein sequence database, consists of 170140 entries, with an additional 1.6 million sequences in a supplementary database [2] awaiting addition. One of the significant issues with data of this nature is how to annotate sequences with properties that can occur anywhere along the length of the sequence. Manual experimental annotation in a biologist's laboratory is reliable but time consuming and expensive. Automatic annotation is fast and cheap.

The case study presented in this paper is the problem of determining signal peptides. Given a database of protein sequences with the signal peptides annotated, can a machine learning system discover the rules underlying the form and nature of a signal peptide?

Signal peptides are important because they direct proteins to their correct destination within the cell. Proteins need to have this "address" because they serve a multitude of functions, such as being reaction catalysts and transport molecules [12]. They are also the basic building blocks of the cell itself, and signal peptide failures can lead to diseases such as cystic fibrosis [3]. Knowledge of how signal peptides work is also useful when designing new drugs, which are often created in the form of proteins and therefore must have the correct signal attached to them [3].

Once a protein reaches its destination, its signal peptide is no longer needed. By a careful process of alignment, the signal peptide is cleaved off, severing it from the rest of the protein. An important point is that the signal peptide is always cleaved at exactly the same point along the protein sequence. The question posed here is: is it possible to predict this unique cleavage point for a newly sequenced protein?

The basic process described in this paper involves firstly extracting features from the training sequences. The frequencies of the features are determined and

converted into probabilities, and then Bayes' Theorem is applied to predict the posterior probability of a cleavage point given each feature. When a test sequence is presented, the posterior probability of a cleavage point at each position along the sequence can be calculated and the position with the highest posterior is taken to be the predicted cleavage site.

This relatively simple Bayesian method is comparable to state-of-the-art neural network methods. Furthermore, this method can provide rudimentary explanations (in terms of ranked features) for its predictions. Such explanations are important for biologists trying to understand the nature of signal peptides.

In the next section, the biological and machine learning background to this paper is reviewed. Section 3 describes my proposed Bayesian method, and Section 4 reports on some results using a signal peptide dataset. Section 5 is the conclusion and mentions some issues for future research to address.

2 Background

2.1 Biological Background

All protein molecules are made up of a linear sequence of smaller molecules called amino acid residues. There are twenty amino acid residues in total. Each residue by convention has two abbreviations: a three-letter abbreviation and a one-letter abbreviation. For example, the abbreviations of Alanine are *Ala* and *A*. All twenty residues and their standard abbreviations are listed in Table 1.

Table 1. Amino acid residue abbreviations.

Residue	**Abbreviations**	
Alanine	*Ala*	*A*
Arginine	*Arg*	*R*
Asparagine	*Asn*	*N*
Aspartic acid	*Asp*	*D*
Cysteine	*Cys*	*C*
Glutamic acid	*Glu*	*E*
Glutamine	*Gln*	*Q*
Glycine	*Gly*	*G*
Histidine	*His*	*H*
Isoleucine	*Ile*	*I*
Leucine	*Leu*	*L*
Lysine	*Lys*	*K*
Methionine	*Met*	*M*
Phenylalanine	*Phe*	*F*
Proline	*Pro*	*P*
Serine	*Ser*	*S*
Threonine	*Thr*	*T*
Tryptophan	*Trp*	*W*
Tyrosine	*Tyr*	*Y*
Valine	*Val*	*V*

Computationally speaking, a protein sequence can be viewed as a string of symbols (the residues) drawn from an alphabet of size twenty. Although it is also possible to augment each residue with a set of its properties, in this paper I consider only the basic sequence itself.

Signal peptides have a known structure that can aid in predicting the cleavage point, but within that structure there is considerable variability that makes the task difficult. In this study's datasets, the length of the signal peptide varies from five residues up to 90 residues. The average length is approximately 25 residues. In contrast, the total length of a protein can be thousands of residues. Signal peptides always occur at the beginning (the *N-terminal*) of the protein.

According to von Heijne [13] and Neilson & Krogh [7], a signal peptide consists of three main regions. Firstly, there is the *n-region* near the N-terminal, which comprises positively charged residues and is the greatest contributor to the variability in the length of a signal peptide [3]. This is followed by a so-called *h-region*, which is a longer stretch of eight to fifteen hydrophobic residues. Finally, near the cleavage point, there is typically a *c-region*,

consisting of around five mostly uncharged amino acids. This structure is depicted in Figure 1, using the sequence for human growth hormone as an example.

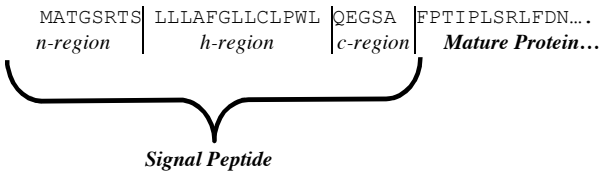

Fig. 1. Structure of a signal peptide for human growth hormone.

The most important part of the signal peptide is the h-region: it serves the dual purpose of both encoding the protein's destination, and it is also used to align the signal peptide for cleavage when it finally arrives [12].

It should be noted that each of these regions are not necessarily a contiguous run of like residues. The hydrophobic h-region, for example, can be interrupted more than once by sequences of non-hydrophobic residues. This contributes to the difficulty of making predictions.

2.2 Signal Peptide Prediction Background

The earliest signal peptide prediction method was known as "the (-3,-1) rule" [12, 8]. This basically followed from the observation that positions –3 and –1 upstream (i.e. to the left) of the cleavage point were often "small and neutral". Using this simple rule seemed sufficient when the number of known signal peptides was small, but it has proved inadequate as the amount of data has increased.

Chou [3] extended the (-3,-1) rule when he introduced the subsite coupling approach. Basically, he formulated an algorithm which takes into account additional positions such as +1, as well as the expected lengths of each of the regions. The algorithm outputs the position on the sequence most likely to be the actual cleavage point. Although Chou reports that the results are encouraging, this method was trained on different data than the other methods were trained on and so it is difficult to make comparisons. An important point is that both of these approaches operate directly on variable-length sequences.

In contrast, more recent machine learning approaches do not operate directly on the variable length sequences but instead preprocess the sequences into fixed length records and transform the problem into one of classification rather than sequence annotation. For example, if the fixed record size is ten positions and the original sequence length is, say, 34, then 24 fixed length records would be produced from this single original sequence. Such preprocessing fits well with existing machine learning tools because they demand fixed-length data, but it does have a number of drawbacks.

The main one is that since each original sequence only has a single cleavage point, there is going to be a high abundance of negative examples (in a single sequence, only one fixed length record ends in the cleavage point and is therefore labelled as positive; the rest are labelled negative). Many machine learning algorithms given this biased data may simply predict every sequence as negative in order to

obtain a high level of testing accuracy. To eliminate this problem and balance the classes more evenly, a considerable number of negative examples have to be discarded – a situation that could result in important information being lost.

The currently best-known and most widely used machine learning solution is the SignalP suite [1, 7, 8]. SignalP version 1 was a solely neural network approach. The neural network had a feedforward architecture and was trained on fixed length, sparsely encoded records derived from a "moving window" [8]. Hidden Markov models were added as a second predictor in SignalP version 2 [7], which increased accuracy slightly but also had the added benefit of being able to discriminate with high accuracy between signal sequences and non-signal sequences. SignalP version 3 [1] is a refinement of both the neural network and hidden Markov model approaches, with a claimed significant increase in prediction accuracy. Table 2 summarises the prediction accuracy results as reported by Bendtsen et al. [1]. Different neural network architectures have failed to provide a significant improvement over Signal P (see, e.g., [4, 9]).

Table 2. Best recorded accuracies of the SignalP suite of predictors

	Eukaryotyes	**Gram-**	**Gram+**
SignalP1	70.2	79.3	67.9
SignalP2	72.4	83.4	67.4
SignalP3	79.0	92.5	85.0

There are number of points worth mentioning about these results. Firstly, separate predictors were trained from data from three different sources: Eukaryotes (being all organisms except viruses, bacteria, and blue-green algae), and two types of Prokaryotes (bacteria): Gram-positive and Gram-negative. Other approaches do not subdivide the data at all and therefore the results are not directly comparable.

One significant weakness of the SignalP evaluations was that they performed only five-fold cross validation. In most cases, 10-fold cross validation is the minimum required for statistical significance [14].

Support Vector Machines (SVMs) have also been applied to this problem. Vert [11] developed a new SVM kernal for strings and applied his method to cleavage point prediction. His dataset was the same as that used to train SignalP1, but he did not subdivide the data. He reports 68% accuracy in predicting the cleavage point.

Some authors have attempted to incorporate residue properties into their systems to improve prediction accuracy. Recently, Smith [10] used a naïve Bayes-based text mining approach and reported accuracy comparable to Vert's SVM approach described above. Maetschke et al. [6] compared a number of different encoding of Blomaps using the WEKA machine learning workbench [14] and came to the conclusion that a particular encoding called BLOSUM62 combined with naïve Bayes produced the best results.

One difficulty when comparing these approaches is the lack of a standard benchmark dataset. It should be noted that Vert [11], Smith[10], and SignalP version 1 [8] all use the same dataset, namely that developed for SignalP version 1. Other authors have generated their own datasets from the SWISS-PROT database, and therefore it is quite possible that differences in accuracy are largely due to differences in data. To date, the SignalP2 dataset is publicly available but the SignalP3 dataset is not available.

3 Fast Bayesian Cleavage Point Prediction

3.1 Dataset Description

Before describing the method, it is necessary to briefly describe the sequence data used.

The dataset in this study is the same dataset used to train SignalP version 2. Each record consists of three lines: firstly, a biological description of the sequence in English which ties the sequence to its original record in SWISS-PROT; second, the residues sequence from the N-terminal all the way to position 29 downstream of the cleavage site; and thirdly, an annotation showing which residues are part of the signal peptide, which are part of the mature protein, and which is the cleavage site (defined as the first residue of the mature protein). Figure 2 below depicts two sample records taken from the dataset.

```
51 11SB_CUCMA     21 11S GLOBULIN BETA SUBUNIT PRECURSOR.
MARSSLFTFLCLAVFINGCLSQIEQQSPWEFQGSEVWQQHRYQSPRACRLE
SSSSSSSSSSSSSSSSSSSSSCMMMMMMMMMMMMMMMMMMMMMMMMMMM

54 41BB_MOUSE     24 T CELL ANTIGEN 4-1BB PRECURSOR.
MGNNCYNVVVIVLLLVGCEKVGAVQNSCDNCQPGTFCRKYNPVCKSCPPSTFSS
SSSSSSSSSSSSSSSSSSSSSSSCMMMMMMMMMMMMMMMMMMMMMMMMMMMM
```

Fig. 2. Two records taken from the dataset.

The method by which this dataset was derived is worth briefly mentioning. SWISS-PROT contains protein sequences both with experimentally verified cleavage points, and without them. The database is constantly being updated as new protein sequences are added, and errors in annotations of existing sequences are corrected. Only sequences with experimentally-verified cleavage points were included in the dataset. Furthermore, some sequences were removed if they met certain criteria, for example having an origin in a virus gene [7].

The next step in the dataset creation was homology reduction. Many protein sequences occurring in nature are homologous, that is, they share long common subsequences which may include the signal peptide. This means that simple string alignment could result in a very high accuracy when predicting cleavage points on test sequences homologous to the training sequences. To eliminate this potential source of bias, for all pairs of homologous sequences in the dataset, Neilson & Krogh [7] discarded one of the sequences. By this method, more than 50% of the sequences in the dataset were discarded.

The final dataset contains 1666 protein sequences, of which 1137 are Eukaryote sequences, 697 are Gram negative Prokaryote sequences, and 280 are Gram positive Prokaryote sequences.

3.2 Training Method and Model

I will now describe the Bayesian method used for building a model based on the training data, and applying it to the prediction of signal peptide cleavage points. This

approach is relatively simple, fast to train, and as shall been seen in the next section, has accuracy comparable to existing systems.

The basic idea is to define a set of features that protein sequences can have, extract from the training set the frequencies of those features, and convert those frequencies into posterior probabilities. This set of features and their posteriors will be referred to as the model. The model is then used to predict the final posterior probability of a cleavage point at each position along a test sequence given all the features on the test sequence.

What are the features? I define two types of feature: a pattern of residues that may occur anywhere along a sequence, and a pattern of residues at a fixed position relative to some other position. Table 3 gives some examples of features extracted from the human growth hormone sequence depicted in Figure 1. I have used an "@" symbol to denote patterns with a position specified.

Table 3. Examples of features extracted from training dataset.

Feature	Description
A	The residue Alanine.
C_L	Cysteine, followed some other residue, followed by Leucine.
L@-10	Leucine at position –10 relative to some position *c*.
C_L@-3	Cysteine at position –3 and Leucine at position –1, both relative to some position *c*.

The following features were extracted from the training set because they resulted in the best accuracies during informal testing: all of the features comprising single residues, without any limits on the distance of the residue from the cleavage point (e.g. see the first and third rows of Table 3); and all the diresidue sequences separated by exactly one position (e.g. see second and fourth rows of Table 3). However, only the position-specific diresidue sequences (i.e. those with an "@" symbol) starting at –3 were extracted. The reasoning for this is that such an approach makes the standard simplifying naïve Bayes assumption (i.e. that the occurrence of a residue at a particular position relative to the cleave point is independent of the residues at other positions given the cleave point). However, this does not hold for (-3,-1), which are considered non-independent. By having a specific feature for the diresidue pattern at (-3,-1), the system can therefore effectively model the (-3,-1) rule mentioned earlier.

Now, for every feature, a probability is calculated. Suppose *f* is a single residue or pattern of residues without a specific position, and *f@p* is the same pattern with a specific relative position. The probability *P(f)* is defined as the prior probability of *f@p*, and is determined by calculating the total fraction of occurrences of *f* in the training set, in both signal and non-signal portions of the sequences. For example, if *f=A*, then *P(f)* is simply the total fraction of residues in the training set that is Alanine.

For each feature, a conditional probability is also calculated. Let *cleave(c)* denote the proposition that position *c* on the sequence is the cleavage point. *P(f@p | cleave(c))* is defined as the fraction of occurrences in the training set of the feature at a particular fixed position relative to the known cleavage point.

For example, from the dataset, the prior probability of the single-residue feature *L*, *P(L)*, is 0.127, but *P(L@-1|cleave(0))* = 0.019 and *P(L@-15|cleave(0))* is 0.285. While the priors capture the general abundance of residues in the training data,

the conditionals capture the distribution of residues across positions relative to the cleavage point. I also compute conditional probabilities for the patterns occurring at positions (-3,-1), as mentioned above.

It is now time to explain how the posterior probabilities used for prediction are computed. Essentially, this is an application of Bayes' Theorem. Equation (1) shows how the priors and conditionals are combined to compute the overall probability of a cleave at some position *c*. *F* is defined as the set of all features on a particular sequence with positions relative to some position *c*. The training model consists of a posterior probability for every feature present in the training data.

$$P(cleave(c) \mid F) \quad \propto \prod_{f @ p \in F} \left[\frac{P(f @ p \mid cleave(c))}{P(f)} \right] \tag{1}$$

We now come to the prediction algorithm. Given a test sequence with an unknown cleavage point, the system predicts a score for every position *c* on the test sequence. The score is the posterior probability as defined in Equation (1) above. When every position is scored, the posteriors are normalised and the position with the highest posterior probability is the predicted cleavage point. Figure 3 depicts the output of the system when tested on the sequence for human growth hormone depicted in Figure 1 after training on the entire dataset minus the human growth hormone sequence. The predicted probability of a cleave at the actual cleavage site is 0.87.

```
MATGSRTSLLLAFGLLCLPWLQEGSAFPTIPLSRLFDNAMLRAHRLHQLAFDTYQE
SSSSSSSSSSSSSSSSSSSSSSSSSSCMMMMMMMMMMMMMMMMMMMMMMMMMMM
…
W     S       0.00277371
L     S       0.000196609
Q     S       0.00609159
E     S       0.00524503
G     S       0.0381914
S     S       0.0577272
A     S       0.0125238
F     C       0.874029              ************
P     M       0.000244162
T     M       0.0015616
…
```

Fig. 3. Normalised predictions for human growth hormone. Only residues with a non-negligible probability of being the cleavage point are shown.

4 Results

I evaluated the method described in the previous section using Leaving One Out Cross Validation (LOOCV) on the SignalP version 2 dataset (the dataset for SignalP3 is different and currently unavailable). LOOCV was applied to the entire dataset, as well as the same three subsets that SignalP was trained on, namely the Eukaryote, Gram positive Prokaryotes, and Gram negative Prokaryotes subsets.

4.1 Accuracy

Compared to computationally more expensive methods such as neural networks, this approach results in comparable testing accuracy. Table 4 compares the accuracies achieved by SignalP version 2 and this method, both of which were trained on the same dataset. A comparison with other versions of SignalP is not as useful because of the different datasets being used.

Table 4. Comparison of SignalP2 and the Bayesian method described in this paper.

	Eukaryotyes	**Gram-**	**Gram+**
SignalP2	72.4	83.4	67.4
Bayesian	69.2	81.5	66.5

As can be observed, the Bayesian method is consistently 1-2% less accurate than SignalP2. However, such a slight difference is likely to be a reflection of the statistical variation arising from Neilson & Krogh's [7] use of the less-rigorous five-fold cross validation for testing. In contrast, the Bayesian method utilised the more reliable LOOCV method. The difference may also reflect the independence assumption made about all positions except (-3,-1): it is possible that including additional diresidue features could further increase accuracy. (Interestingly, treating positions (-3, -1) as non-independent contributes to a large proportion of the accuracy. If this feature is not extracted, and instead only two independent features for positions –3 and –1 are used, then the accuracy is reduced by about 25%.)

I also tested the predictive performance of the Bayesian approach when trained on the entire SignalP2 dataset without subdivision. Again, LOOCV was applied. The accuracy for this experiment was 71.2%, which compares favourably with Vert's SVM approach [11] that achieved 68% accuracy, albeit on the (mostly similar) SignalP1 dataset.

Aside from raw accuracy, one can also consider how close erroneous predictions are from the actual predictions. In Figure 4, the distribution of predicted cleavage sites against proximity to the real cleavage site are depicted following LOOCV on the entire SignalP2 dataset. The diagram clearly shows that the majority of predictions (91.4%) lie within –5 and +5 of the actual cleavage site even though the raw accuracy is 71.2%. It is quite possible that many of these predictions are correct, but have been misclassified by the experimental biologist, as suggested by Hiller et al. [4].

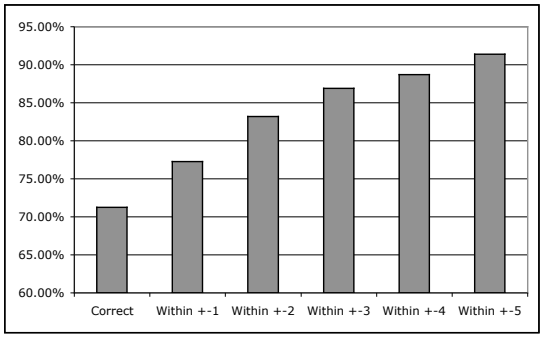

Fig. 4. Accuracy of prediction vs. percentage frequency after LOOCV on entire dataset.

Finally, it is necessary to comment on the relationship between the value of the posterior probability and the confidence of the prediction. In other words, is the posterior probability calculated a good indicator of the reliability of the prediction? I performed an analysis on the results of the LOOCV

experiment applied to the entire dataset, and found a positive correlation between posterior probability and true positive rate. The result of this analysis is depicted graphically in Figure 5.

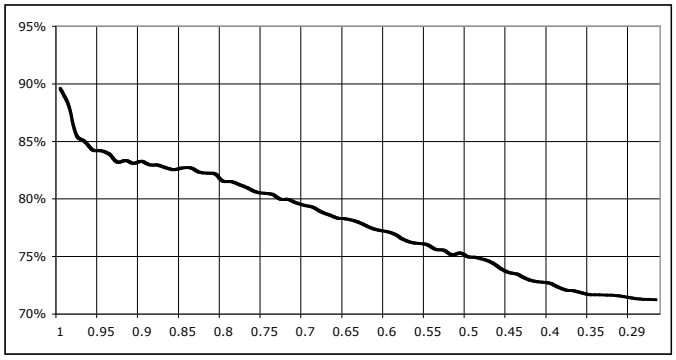

Fig. 5. Posterior probability of predicted cleavage site vs. true positive percentage.

Clearly, predictions with a high posterior probability are to be considered more confident than predictions with a low posterior probability. For example, where the best predicted cleavage point has a probability of only 0.5 or above, the true positive rate was only 75%. However, for predictions with a posterior of 0.95 and above, the true positive rate is between 85% and 90% - quite a significant increase.

4.2 Explanations

The Bayesian method has one significant advantage over neural network approaches: namely, the ability to extract the reason for the system making a particular prediction. Since the overall posterior probability of a cleave is simply the product of the individual posteriors of a cleave given a single feature, it is possible to rank the features in a test sequence by how much they contribute to the final prediction. In Figure 6, the features contributing to the human growth hormone prediction shown in Figure 3 are listed in decreasing order of individual posterior.

```
A@-1 (5.3), GA@-3,-1 (3.7), L@-12 (2.8), P@1 (2.8), L@-11 (2.6),
L@-9 (2.6), L@-16 (2.1), L@-17 (1.6), L@-6 (1.6), P@4 (1.5),
F@-14 (1.4), Q@-5 (1.4), T@2 (1.4), L@-18 (1.4), W@-7 (1.3),
A@-15 (1.3), S@-2 (1.2), S@-19 (1.1), R@7 (0.9), S@6 (0.9),
C@-10 (0.9), T@-20 (0.9), G@-3 (0.8), F@9 (0.8),I@3 (0.7),
G@-13 (0.6), F@0 (0.6), E@-4 (0.5), L@8 (0.5), L@5 (0.4), P@-8 (0.2)
```

Fig. 6. Features and their posteriors ranked from most significant to least significant, for the human growth hormone prediction.

It can be seen that the biggest contributor to the prediction is the presence if *Ala* at position –1. The pattern of *Gly* and *Ala* at positions (–3, –1) is the second largest predictor, and this is followed by the occurrence of *Leu* at multiple positions from –6 to –19, which is where the hydrophobic region is expected to be. *Pro* at positions 1 and 4 also has a high posterior.

5 Conclusion

To conclude, an efficient and effective method of predicting signal peptide cleavage points along protein sequences has been presented. I have shown that computationally more expensive approaches are not necessarily better in terms of accuracy than simpler Bayesian approaches, and the Bayesian approach described here can offer some degree of explanations for its predictions. Some of the issues involved in applying data mining techniques to biological datasets (such as dealing with variable length sequences) have also been explored.

References

1. Bendtsen J., Neilson H., von Heijne G., Brunak S.: Improved Prediction of Signal Peptides – SignalP 3.0. Journal of Molecular Biology **340** (2004), 783-795.
2. Boekmann B. et al.: The SWISS-PROT Protein Knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research **31**(1) (2003) 365-370.
3. Chou K.: Prediction of Protein Signal Sequences. Current Protein and Peptide Science **3** (2003), 615-622.
4. Hiller K., Grote A., Scheer M., Munch R., Jahn D.: PrediSi: Prediction of Signal Peptides and their Cleavage Positions. Nucleic Acids Res. **1** (2004) W375-W379.
5. Hua S., Sun Z.: Support Vector Machine Approach for Protein Subcellular Localization Prediction. Bioinformatics **17**(8) (2001), 721-728.
6. Maetschke S., Towsey M., Boden M: BLOMAP: An Encoding of Amino Acids which Improves Signal Peptide Cleavage Site Prediction. In Chen Y., Wong L: Proc. 3rd Asia-Pacific Bioinformatics Conference, Imperial College Press (2005).
7. Neilson H., Krogh A.: Prediction of Signal Peptides and Signal Anchors by a Hidden Markov Model. In: Glasgow J et al.: Proc Sixth Int. Conf. on Intelligent Systems for Molecular Biology. AAAI Press (1998), 122-130.
8. Nielson H., Englebrecht J., Brunak S., von Heijne G.: Identification of Prokaryotic and Eukaryotic Signal Peptides and Prediction of their Cleavage Sites. Protein Engineering **10**(1) (1997), 1-6.
9. Reczko M., Fiziev P., Staub E., Hatzigeorgiou A: Finding Signal Peptides in Human Protein Sequences using Recurrent Neural Networks. In Guigo R., Gusfield D.: Algorithms in Bioinformatics, Proceedings of the 2nd Int. Workshop WABI 2002, Rome, Italy, Lecture Notes in Computer Science, Springer, **2452** (2002), 60-67.
10. Smith T: A Text-Classification Approach to the Prediction and Characterization of Signal Peptides. In: Informatica 2004: World Congress on Bioinformatics, Havana, Cuba (2004).
11. Vert J.: Support Vector Machine Prediction of Signal Peptide Cleavage Sites Using a New Class of Kernals for Strings. In: Proc. Pacific Sym. on Biocomputing (2002), 649-660.
12. von Heijne G.: Life and Death of a Signal Peptide. Nature **396** (1998) 111-113.
13. von Heijne, G: A New Method for Predicting Signal Sequence Cleavage Sites. Nucleic Acids Research **14**(11) (1986), 4683-4690.
14. Witten I., Frank E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kauffman (1999).