View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Queensland University of Technology ePrints Archive

COVER SHEET

This is the author version of article published as:

Maetschke, Stefan and Towsey, Michael W. and Boden, Mikael (2005) Blomap: an encoding of amino acids which improves signal peptide cleavage site prediction . In Chen, Yi-Ping Phoebe and Wong, Limsoon, Eds. Proceedings Third Asia Pacific Bioinformatics Conference 1, pages pp. 141-150, Singapore.

Copyright 2005 (please consult author)

Accessed from http://eprints.qut.edu.au

BLOMAP: AN ENCODING OF AMINO ACIDS WHICH IMPROVES SIGNAL PEPTIDE CLEAVAGE SITE PREDICTION

STEFAN MAETSCHKE[1], MICHAEL TOWSEY[2] AND MIKAEL BODÉN[1]

[1] *ITEE, University of Queensland,*
Brisbane, Queensland 4072, Australia

[2] *CITI, Queensland University of Technology*
Brisbane, Queensland 4001, Australia

Research on cleavage site prediction for signal peptides has focused mainly on the application of different classification algorithms to achieve improved prediction accuracies. This paper addresses the fundamental issue of amino acid encoding to present amino acid sequences in the most beneficial way for machine learning algorithms. A comparison of several standard encoding methods shows, that for cleavage site prediction the frequently used orthonormal encoding is inferior compared to other methods. The best results are achieved with a new encoding method named BLOMAP – based on the BLOSUM62 substitution matrix – using a Naïve Bayes classifier.

1. Introduction

Transport of proteins is controlled by signal peptides, sequences of 15 to 25 amino acid residues attached to the N-terminal end of a protein [6]. Signal peptides basically serve as *zipcodes,* ensuring that a protein is delivered to its correct secretory pathway. The signal peptide is removed by signal peptidase when the mature protein is translocated through the membrane. Since defects in the protein sorting process cause many diseases, there is considerable scientific and commercial interest in identifying signal peptides and their cleavage sites [3, 12].

The standard way to identify the function of a polypeptide is by *sequence homology* determined by sequence alignment against other polypeptides of known function. The method of homology fails for signal peptides because, in spite of shared functionality, their sequence similarity is usually low [10]. Several alternative methods have been developed to overcome this difficulty.

The earliest approach is based on the (-3,-1) rule which states that the residues at positions -3 and -1 (relative to the cleavage site) are small and neutral whereas the residue at position -2 is usually aromatic, charged or polar [21]. However, the prediction accuracy using this simple rule is low, 64% for eukaryotic proteins and 47% for prokaryotic proteins [22]. In 1986, von Heijne introduced the concept of weight matrices for signal peptide identification and cleavage site prediction. Weight matrices are calculated from position specific amino acid frequencies when the signal peptides are aligned to their cleavage sites. To locate the cleavage site within a new sequence, a sliding window is moved along the sequence and the sum of the weighted residues serves as an indicator for a cleavage site at the window centre. One of the first attempts to tackle the problem of cleavage site prediction with machine learning algorithms employed a neural network

whose topology and weights were adapted using an evolutionary strategy and seven physicochemical features to encode the amino acids [17]. However the method did not achieve the accuracy of the simpler weight matrix method. More recently, Nielsen [12] developed another neural network approach, SignalP, which uses two Multilayer Perceptrons trained by backpropagation. The first network has an asymmetric input window around a hypothetical cleavage site and outputs the validity of it. The second network has a symmetric input window around a residue and classifies the residue as belonging to a signal peptide or not. The outputs of both networks are combined, yielding an accuracy of 79%, 85% and 92% for three different data sets [1]. In a subsequent study, Nielsen used hidden Markov models (HMM) for the same task but the results were not as good as the neural network approach [1,13]. SignalP is currently considered to be the benchmark algorithm for the signal peptide cleavage prediction.

Ladunga [10] has applied a software package, PHYSEAN, designed for protein classification, to the cleavage site prediction task. When the amino acids were encoded by a set of 126 normalized physicochemical features, PHYSEAN outperformed SignalP (version 1.2) by 12% but using a different data set than that of Nielsen. None of the studies described above examined the effect of different amino acid encodings on prediction accuracy. It is therefore an open question which method of encoding is appropriate for the cleavage site prediction task.

Some machine learning methods, such as the HMM [13] and the Bayes classifier [3] accommodate symbolic input and do not require numerical encoding of the amino acids. For example, Vert [20] has developed a new class of string kernels for support vector machines (SVM) that can evaluate amino acid sequences directly. Neural networks, on the other hand, do require some form of numerical encoding. The typical numerical encoding is the orthonormal, but we demonstrate in this paper that this encoding is sub-optimal. We explore several other encodings including a new encoding called BLOMAP and compare their performance for the cleavage site prediction task using Nielsen's data sets from 1999.

2. Encodings

The most frequently used encoding is the orthonormal, also known as distributed encoding, sparse encoding or encoding with indicator vectors. Each letter l_i of the amino acid alphabet $\mathrm{A} = \{l_1, l_2, \ldots, l_{20}\} = \{\mathrm{A}, \mathrm{R}, \ldots, \mathrm{V}\}$ is replaced by an orthonormal vector:

$$\begin{aligned} enc : \mathrm{A} &\rightarrow \Re^N \\ l_i &\mapsto (\delta_{i,1}, \delta_{i,2}, \ldots, \delta_{i,N}) \end{aligned} \tag{1}$$

where $i, j \in \{1, \ldots, 20\}$, N=20 and δ_{ij} is the Kronecker delta symbol. A sequence of M consecutive amino acid letters can then be mapped to a vector **x** within the feature space $H := \Re^{M \times N}$ by simply concatenating the encoded amino acid letters, using the previously defined encoding relation enc :

$$\begin{aligned} map_{enc} : A^M &\rightarrow \Re^{M \times N} \\ (a_1, a_2, \ldots, a_M) &\mapsto (enc(a_1), enc(a2), \ldots, enc(a_M)) = (x_1, x_2, \ldots, x_{M \cdot N}) \end{aligned} \tag{2}$$

The orthonormal encoding has two drawbacks. First, the dimension of the feature space is twenty times the sequence length resulting in a sparsely populated feature space. Second, since the Euclidian distance between two encoded amino acids is always two, all information about similarity between amino acids is lost. Polymers with different sequences but similar physicochemical properties will not appear closer in the input space than dissimilar polymers. One common method to alleviate this disadvantage is to group similar amino acids into sub-alphabets. Amino acids have a great variety of properties such as mass, polarity, hydrophobicity, so many groupings are possible [25]. In this paper, we use the well known Exchange-group G:=({H,R,K} {D,E,N,Q} {C} {S,T,P,A,G} {M,I,L,V} {F,Y,W}) and a hydrophobicity alphabet G:=({D,E,N,Q,R,K} {C,S,T,P,G,H,Y} {A,M,I,L,V,F,W}) from Wu [24] which are encoded as

$$\begin{aligned} enc &: \mathrm{A} \to \Re^N \\ l \in G_i &\mapsto \left(\delta_{i,1}, \delta_{i,2}, \ldots, \delta_{i,N}\right) \end{aligned} \tag{3}$$

where N is the number of groups within the sub-alphabet. Encoding amino acids with sub-alphabets reduces the dimension of the input space but there is no information about the distance between groups and the feature space remains sparsely populated. In addition, it is usually uncertain which grouping scheme is appropriate for a given problem.

Another popular method is to encode amino acids directly, using a set of physicochemical features f_i :

$$\begin{aligned} enc &: \mathrm{A} \to \Re^N \\ l_i &\mapsto \left(f_{i,1}, f_{i,2}, \ldots, f_{i,N}\right) \end{aligned} \tag{4}$$

where N is the number of features. The AAindex database [8] currently contains 494 features and the question is how to select a suitable subset. We took the following seven properties: volume, mass, hydrophobicity, surface area, α-helix, β-strand and turn propensity described in [24, page 70].

Amino acids of homologous sequences which are frequently substituted by each other are regarded as similar and the relationships are described by substitution matrices, like the BLOSUM62 matrix [7]. The matrix rows $\mathbf{m}_i$ can be interpreted as feature vectors which describe and encode the similarity between amino acid [24]:

$$\begin{aligned} enc &: \mathrm{A} \to \Re^{20} \\ l_i &\mapsto \mathbf{m}_i \end{aligned} \tag{5}$$

This real value encoding expresses the similarity between amino acids more accurately than the binary encoding by sub-alphabets, but increases the dimension of the feature space by factor 20.

An extremely compact one-dimensional encoding of amino acids can be achieved by use of *scales*. A scale $S := \left(s_1, s_2, \ldots . s_{20}\right)$ defines a value for each amino acid according to some similarity measure and the encoding becomes:

$$\begin{aligned} enc &: \mathrm{A} \to \Re \\ l_i &\mapsto s_i \in S \end{aligned} \tag{6}$$

Two widely used amino acid scales are the hydropathy scales of Kyte [9] and Eisenberg [4]. But two problems remain with scales: First, the selection of an appropriate scale -

Trinquier [19] has reviewed over 40 of them - and second, complex relationships between amino acids can not be captured by a single value.

Taylor [18] classified amino acids according to their physicochemical properties and created a Venn-diagram of ten overlapping classes (see Fig. 1). In 1987 Zvelebil [25] derived a 'truth table' from Taylor's Venn-diagram which describes the membership of an amino acid to one of ten classes as a binary vector $\mathbf{v}_i := (v_{i,1}, v_{i,2}, \ldots v_{i,10})$ with $v_{i,j} \in \{0,1\}$. This representation of amino acids can be used for encoding as well:

$$\begin{aligned} enc : \mathrm{A} &\to \Re^{10} \\ l_i &\mapsto \mathbf{v}_i = (v_{i,1}, v_{i,2}, \ldots, v_{i,10}) \end{aligned} \tag{7}$$

The Zvelebil-encoding has the advantage to preserve some of the multifaceted relationship between amino acids without the high dimensionality of the orthonormal encoding. However, the restriction to binary vectors results in information loss.

It would be desirable to have an encoding which captures the important amino acid properties without excessively increasing the dimensionality of the feature space. In the next section we introduce such an encoding.

3. BLOMAP-encoding

In this section we describe a new dense encoding, named BLOMAP. A common way to measure the similarity between amino acids is by substitution matrices, which contain the substitution frequencies for amino acids in homologous sequences. Amino acids which are frequently mutually substituted are regarded to be similar. The BLOMAP-encoding utilizes a non-linear projection method to exploit the similarity information in a substitution matrix and constructs feature vectors which preserve this information optimally.

Several non-linear projection algorithms such as the Sammon-projection [16] or the FastMap-algorithm [5] are applicable but because of the small number of elements to map and its simplicity, we implemented an improved version of the Sammon-projection. Details of the algorithm are described in [11]. For the substitution matrix, we chose the common BLOSUM62 matrix [7].

The Sammon-projection maps a set of vectors from a high dimensional input space $H := \{\mathbf{x}_i \in \Re^{m_h} \mid 1 \le i \le n\}$ to a usually lower dimensional target or feature space $L := \{\mathbf{y}_i \in \Re^{m_l} \mid 1 \le i \le n\}$:

$$\begin{aligned} sammon : H &\to L \\ x_i &\mapsto y_i \end{aligned} \tag{8}$$

with $i \in \{1, \ldots, n\}$ in a way that the mapping error E is minimized. The mapping error and the algorithm itself are based on the Euclidean distances $d_{ij}^h := dist(\mathbf{x}_i, \mathbf{x}_j)$ within the input space and the distances $d(t)_{ij}^l := dist(\mathbf{y}_i, \mathbf{y}_j)$ within the target space.

H is unknown but the BLOSUM62 matrix – which is a similarity matrix – can be converted into a distance matrix to provide the required distances d_{ij}^h by

$$d_{ij}^h := \frac{1}{2^{s_{ij}/2}} \quad \text{for } i \neq j \text{ and } d_{ii}^h := 0 . \tag{9}$$

with s_{ij} are the elements (log odd ratios) of the BLOSUM62 matrix. This allows us to apply the Sammon-projection to generate feature vectors of amino acids in L which optimally preserve the similarity information given by the BLOSUM62 matrix. The last remaining question is of which dimension m_l the feature space L should be. Tests show that almost all of the distance information d_{ij}^h can be captured with five dimensions. Table 1 contains the code vectors for the BLOMAP-encoding in five dimensions.

Table 1. Code vectors for the BLOMAP-encoding in five dimensions.

Letter	BLOMAP62(5)					Letter	BLOMAP62(5)				
A	-0.57	0.39	-0.96	-0.61	-0.69	**L**	0.65	0.84	1.25	-0.99	-1.90
R	-0.40	-0.83	-0.61	1.26	-0.28	**K**	-0.64	-1.19	-0.65	0.68	-0.13
N	-0.70	-0.63	-1.47	1.02	1.06	**M**	0.76	0.05	0.06	-0.62	-1.59
D	-1.62	-0.52	-0.67	1.02	1.47	**F**	1.87	1.04	1.28	-0.61	-0.16
C	0.07	2.04	0.65	-1.13	-0.39	**P**	-1.82	-0.63	0.32	0.03	0.68
Q	-0.05	-1.50	-0.67	0.49	0.21	**S**	-0.39	-0.27	-1.51	-0.25	0.31
E	-0.64	-1.59	-0.39	0.69	1.04	**T**	-0.04	-0.30	-0.82	-1.02	-0.04
G	-0.90	0.87	-0.36	1.08	1.95	**W**	1.38	1.69	1.91	1.07	-0.05
H	0.73	-0.67	-0.42	1.13	0.99	**Y**	1.75	0.11	0.65	0.21	-0.41
I	0.59	0.79	1.44	-1.90	-0.93	**V**	-0.02	0.30	0.97	-1.55	-1.16

However, three dimensions already produce a reasonably good approximation of the distance structure, which gives us the possibility of a visual inspection. The diagram in Figure 1 compares a skyscraper-view on L for $m_l = 3$ with the Venn-diagram by Taylor [18] on the right side.

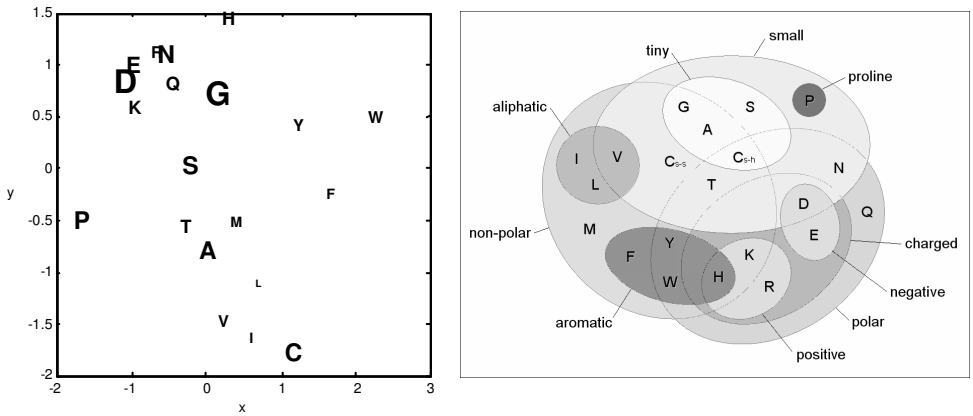

Figure 1. Left: The three-dimensional BLOMAP encodings. The size of letters indicates the third dimension of the encoding (the other two are projected onto the x- and y-axes). Right: Taylor's Venn-diagram.

The skyscraper-view in Figure 1 displays the (*x,y*)-positions of the amino acid letters in L, whereas the *z*-axis or *height* is indicated by the letter size. Amino acids are similar if their locations are close and the letter sizes are similar. Unlike principal component analysis (PCA) or any equivalent linear projection method, all axes of a Sammon-

projection are equally important. Moreover, PCA and similar techniques cannot be applied to a distance or similarity matrix.

A comparison of the diagrams confirms that the BLOMAP-encoding preserves important physicochemical relationships between amino acids. However, in addition to a pure classification, the skyscraper-view provides useful distance information. According to Betts [2], Alanine(A) and Threonine(T) are known to be indifferent amino acids and these amino acids are placed in the diagram centre. All amino acids with a unique characteristic, like Proline(P), Glycine(G), Tryptophan(W), Cysteine(C) and Histidine(H) are located at the diagram borders. The Aliphatic-group {I,L,V} and the aromatic-group {F,Y,W} - without Histidine(H) - appear as distinct clusters. And all hydrophobic amino acids can be found in the lower right corner.

4. Data and Classifiers

To evaluate the influence of amino acid encoding on cleavage site prediction, we downloaded the data suite from www.cbs.dtu.dk/ftp/hnielsen/ created by Nielsen [14]. This suite consists of three redundancy reduced data sets extracted from the SWISS-PROT sequence data base. All sequences comprise the signal peptide part of variable length and the following 30 amino acids of the mature protein.

To create a training set with labelled sequences of fixed length, a window slides over the sequences and produces a positive sample when the window centre hits the first residue of the mature protein. In all other cases a negative sample is generated. Nielson achieved the best results with asymmetric windows and we chose the same window parameters for our experiments. However, to keep the processing time reasonable and because we were only interested in the comparison of different encodings, we balanced the data sets by taking all positive samples and an equal number of randomly drawn negative samples.

Since the performance of an encoding depends also on the chosen classifier, we selected the following set of typical classifiers from the Weka 3.4 data mining package [23] which we used for our experiments:

- A Naïve Bayes (NB) classifier with distribution estimator.
- A decision tree algorithm (J48), C4.5 Revision 8.
- A k-nearest neighbour classifier (IBk), k = 5.
- A single layer perceptron (SLP), max. epochs = 2000.
- A support vector machine (SMO) with a linear kernel.

The *k*-value for the k-nearest neighbour classifier was optimized on the HIV data set [15], which is another cleavage site prediction data set. Since the results of Rögnvaldsson [15] indicate that cleavage site prediction is a linear problem and Nielsen's SignalP software [12] is based on MLPs with zero or two hidden neurons, we used a linear support vector machine and a single layer perceptron. All other classifier parameters kept their default values.

5. Results and Discussion

To evaluate the performance for the different encodings described above in combination with typical classifiers, we measured the mean error on the test set with ten fold cross validation and repeated this ten times. The bar plot in Figure 2 shows the mean test error and the lower and upper bound of the 95% confidence interval for all encodings calculated over all five classifiers and the three data sets.

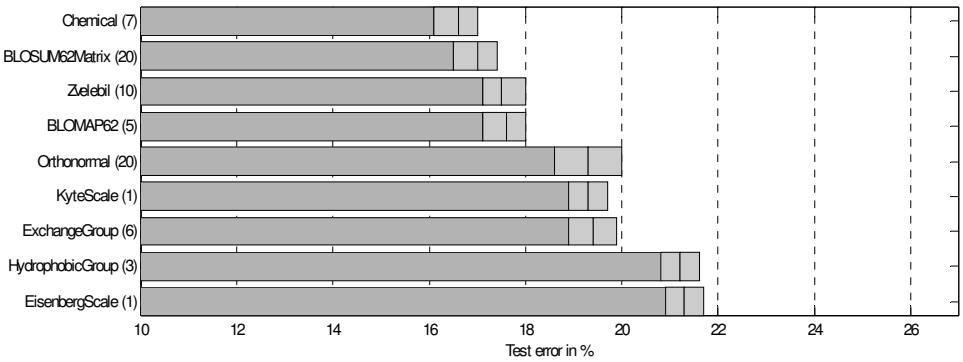

Figure 2. Mean test error and 95%-confidence intervals for all encoding with ten fold cross validation and ten repeats over all classifiers and data sets. Lower value is better. The numbers in brackets indicate the dimensionality of the encoding.

The encoding based on physicochemical properties performed best and is significantly superior to the orthonormal encoding. Also the direct encoding of the BLOSUM62 matrix, Zvelebil's truth table and the five dimensional BLOMAP-encoding outperform the orthonormal encoding.

No significant difference can be found for the orthonormal encoding, Kyte's hydropathy scale and the Exchange-group. The good performance of Kyte's one dimensional hydropathy scale is noteworthy, compared to the 20 dimensions of the orthonormal encoding. None of the sub-alphabet encodings however, achieved lower test errors than the orthonormal encoding.

In an implementation of a cleavage site predictor one would choose the classifier and encoding which performs best. Table 2 contains the mean test errors and 95%-confidence intervals of the different classifiers over all encodings and data sets.

Table 2. Mean test error and 95%-confidence intervals for the different classifiers.

NB	SMO	SLP	J48	IBk
15.4% ±0.2%	17.6% ±0.3%	19.1% ±0.4%	20.6% ±0.3%	25.5% ±0.3%

Surprisingly the Naive Bayes classifier significantly outperformed all other classifiers and the linear support vector machine comes second. An explanation for the good performance of the Naive Bayes classifier might be that the closely related Weight matrices have already proven their usefulness for cleavage site prediction [22]. The single

layer perceptron achieves a middle rank followed by the decision tree algorithm and the k-nearest neighbour classifier.

The performances for the different encodings in combination with the Naïve Bayes classifier are summarized in Figure 3.

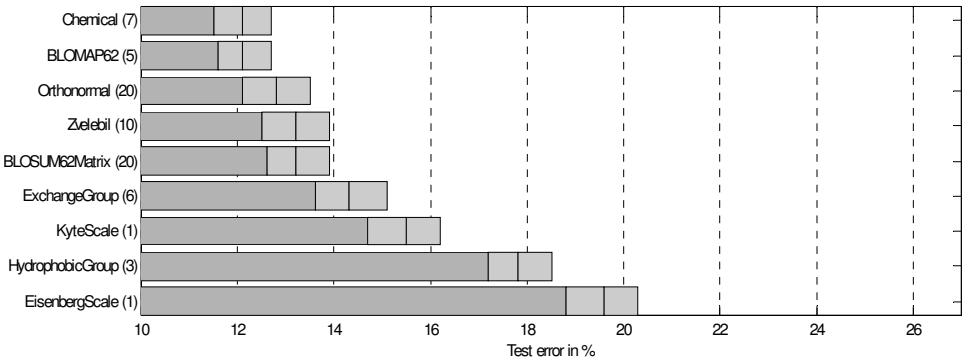

Figure 3. Mean test error and 95%-confidence intervals for the Naïve Bayes classifier with ten fold cross validation and ten repeats over all data sets. Lower value is better. The numbers in brackets indicate the dimensionality of the encoding.

The chemical and the BLOMAP-encoding are performing equally well for the Naïve Bayes classifier. However, the orthonormal encoding also achieves a low test error and Zvelebil's truth table and the direct encoding of the BLOSUM62 matrix are not significantly inferior. On the other hand, the performance of the Exchange-group and Kyte's hydropathy scale is in this case significantly lower than that of the orthonormal encoding. All other encodings generate much higher test errors.

The currently best results for cleavage site prediction are achieved by Nielsen's SignalP software [1,12] which is based on a multilayer perceptron with zero or two hidden neurons and orthonormal encoding. Therefore we were especially interested in the performance of different encodings with this type of network. The bar plot in Figure 4 contains the results for the single layer perceptron.

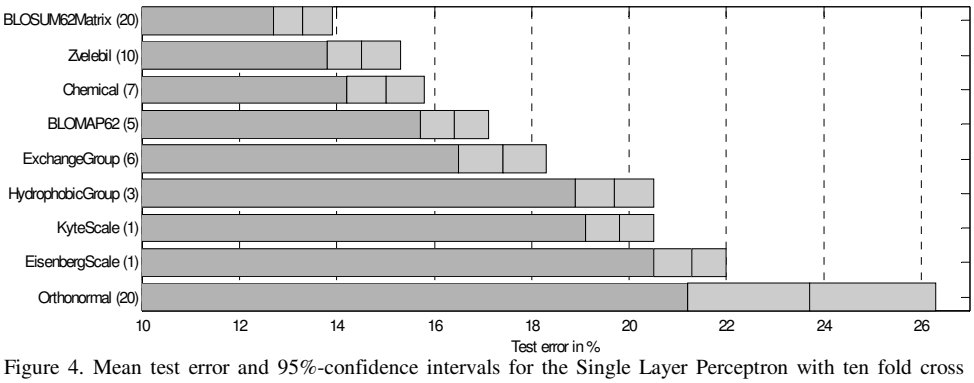

Figure 4. Mean test error and 95%-confidence intervals for the Single Layer Perceptron with ten fold cross validation and ten repeats over all data sets. Lower value is better. The numbers in brackets indicate the dimensionality of the encoding.

The bar plot shows that the orthonormal encoding in combination with a single layer perceptron is inferior to all other tested encodings for cleavage site prediction with single layer perceptrons. Note the large confidence interval, caused by performance fluctuations. Interestingly, encodings with equal dimensionality like the BLOSUM62 encoding or encodings of similar structure like the Exchange-group display stable learning. Since the network does not contain hidden neurons, the reason can not be found in the presence of local minima of the error function. Furthermore, preliminary tests (not shown here) indicate that the orthonormal encoding performs well when used in connection with multilayer perceptrons with hidden neurons. However, it does not outperform the BLOMAP-encoding with the Naïve Bayes classifier.

6. Conclusion

We have demonstrated that the encoding of amino acids has a significant influence on the accuracy of cleavage site prediction and that the commonly used orthonormal encoding should not be used in combination with a single layer perceptron for this type of task.

Since the best results to date have been achieved by SignalP, which implements single layer and multilayer perceptrons with orthonormal encoding, we expect that these results can be improved by using our new BLOMAP62-encoding and a Naïve Bayes classifier.

Compared to other standard encodings, the BLOMAP-encoding has several advantages. First of all, it simplifies the selection of a suitable encoding for a specific problem. Known substitution matrices can be utilized or problem specific matrices can be calculated. The BLOMAP62-encoding optimally preserves the similarity information contained in a substitution matrix and is scaleable to accommodate memory or time limitations. The application of the Sammon-projection on problem specific substitution matrices could lead to new insights into the metabolic relationships between amino acids.

References

1. Bendtsen, J. D., H. Nielsen, et al. (2004). "Improved prediction of signal peptides: SignalP 3.0." Journal of Molecular Biology 340: 783—795.
2. Betts, M. J. and R. B. Russell (2003). Amino acid properties and consequences of subsitutions. *In Bioinformatics for Geneticists*. M. R. Barnes and I. C. Gray. Hoboken, NJ, Wiley.
3. Chou, K.-C. (2001). "Using subsite coupling to predict signal peptides." *Protein Engineering* 14(2): 75—79.
4. Eisenberg, D., R. M. Weiss, et al. (1982). "The helical hydrophobic moment: a measure of the amphiphilicity of a helix." *Nature* 299(5881): 371—374.
5. Faloutsos, C. and K.-I. Lin (1995). "A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets." *Proc. ACM SIGMOD*: 163—174.
6. Gierasch, L. M. (1989). "Signal sequences." *Biochemistry* 28: 923—930.
7. Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." *Proc. Nat. Acad. Sci. USA 89*: 10915—10919.

8. Kawashima, S. and M. Kanehisa (2000). "AAindex: Amino Acid index database." *Nucleic Acids Research* 28: 374.
9. Kyte, J. and R. F. Doolittle (1982). "A simple method for displaying the hydropathic character of a protein." *Journal of Molecular Biology* **157**(1): 105-132.
10. Ladunga, I. (1999). "PHYSEAN: PHYsical SEquence ANalysis for the identification of protein domains on the basis of physical and chemical properties of amino acids." *Bioinformatics* 15(12): 1028—1038.
11. Maetschke, S. (2004). "A simplification of Sammon's projection method". Technical Report SM-ITEE-UQ-07-04, http://www.itee.uq.edu.au/~stefan/.
12. Nielsen, H., J. Engelbrecht, et al. (1997). "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites." *Protein Engineering* 10(1): 1—6.
13. Nielsen, H. and A. Krogh (1998). "Prediction of signal peptides and signal anchors by a hidden Markov model." *Proc Int Conf Intell Syst Mol Biol*. 9: 122—130.
14. Nielsen, H., S. Brunak, et al. (1999). "Machine learning approaches for the prediction of signal peptides and other protein sorting signals." *Protein Engineering* 12(1): 3—9.
15. Rögnvaldsson, T. and L. You (2003). Why Neural Networks Should Not be Used for HIV-1 Protease Cleavage Site Prediction. Technical Report IDE0345. *School of Information Science, Computer and Electrical Engineering*, Halmstad University: 1—8.
16. Sammon, J. W. (1969). "A nonlinear mapping for data structure analysis." *ITEE Transactions on computers* C-18(5): 401—409.
17. Schneider, G. and P. Wrede (1993). "Development of artificial neural filters for pattern recognition in protein sequences." *Journal of molecular evolution* 36: 586—595.
18. Taylor, W. R. (1986). "The classification of amino acid conservation." *Journal of theoretical biology* 119: 205—218.
19. Trinquier, G. and Y. H. Sanejouand (1998). "Which effective property of amino acids is best preserved by the genetic code." *Protein Engineering* 11: 153—169.
20. Vert, J.-P. (2002). *Support Vector Machine Prediction Of Signal Peptide Cleavage Site*. Proceedings of the Pacific Symposium on Biocomputing.
21. von Heijne, G. (1983). "Patterns of amino acids near signal-sequence cleavage sites." *European journal of biochemistry* 133(1): 17—21.
22. von Heijne, G. (1986). "A new method for predicting signal sequence cleavage sites." *Nucleic Acids Research* 14: 4683—4690.
23. Witten, I. H. and E. Frank (2000). Data Mining: Practical machine learning tools with Java implementations. San Francisco, Morgan Kaufmann.
24. Wu, C. H. and J. M. McLarty (2000). *Neural Networks and Genome Informatics*, Elsevier Science.
25. Zvelebil, M. J. J. M., G. J. Barton, et al. (1987). "Prediction of protein secondary structure and active sites using the alignment of homologous sequences." *Journal of Molecular Biology* 195:957—961.