W202–W207 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki587

Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences

Manoj Bhasin and G. P. S. Raghava*

Institute of Microbial Technology, Sector 39-A, Chandigarh, 160036, India

Received November 25, 2004; Revised January 17, 2005; Accepted May 2, 2005

ABSTRACT

This manuscript describes a support vector machine based method for the prediction of constitutive as well as immunoproteasome cleavage sites in antigenic sequences. This method achieved Matthew's correlation coefficents of 0.54 and 0.43 on *in vitro* and major histocompatibility complex ligand data, respectively. This shows that the performance of our method is comparable to that of the NetChop method, which is currently considered to be the best method for proteasome cleavage site prediction. Based on the method, a web server, Pcleavage, has also been developed. This server accepts protein sequences in any standard format and present results in a user-friendly format. The server is available for free use by all academic users at the URL http://www.imtech.res.in/raghava/pcleavage/ or http://bioinformatics.uams.edu/mirror/pcleavage/.

INTRODUCTION

The protein complex known as the proteasome is the main cellular machinery responsible for intracellular protein degradation through ubiquitin-dependent and ubiquitin-independent pathways (1). Moreover, in higher eukaryotes, the proteasome complex also performs the function of generating a pool of peptides for loading onto major histocompatibility complex (MHC) class I molecules (2). It is known that MHC class I ligands which have proteasome cleavage sites at their C-termini have a greater probability of being T-cell epitopes (3). Thus, prediction of sites for cleavage by the proteasome complex is very important for subunit vaccine design based on T-cell epitopes. The proteasome exists in two forms, constitutive proteasomes and immunoproteasomes, with the latter being involved in the generation of MHC class I ligands. In the case of the immunoproteasome, the three catalytically active sites of the constitutive proteasome are replaced by three γ-interferon stimulated subunits (4,5). The peptide fragments generated from such proteasomes are thus effectively the result of the interactions between three active sites with different catalytic specificities that result in the production of peptides fragments with complex specificities (6). Consequently, the cleavage specificities of the constitutive proteasome and the immunoproteasome are different, and these are determined by residues located at the cleavage sites and at neighboring residues, further increasing the complexity of cleavage site prediction (7,8).

The experimental analysis of the products of proteasome cleavage is a cumbersome and time-consuming task. Therefore, computational techniques provide a good alternative to model the cleavage specificities of proteasomes. Owing to the scarcity of well-analyzed data, the development of a knowledge-based method is a difficult task; however, some investigators have designed algorithm(s) with the little data in hand. At present, three software programs are publicly available: PAProC (9,10), MAPPP (7) and NetChop (11). Their performances have been evaluated on an independent dataset obtained from *in vitro* digests of Nef, SSX-2 and RUI proteins and MHC ligands. In this benchmarking, the NetChop software was found to be better than the other two methods, with Matthew's correlation coefficient (MCC) values of 0.32 and 0.16 obtained on data from *in vitro* digests and MHC ligands (1). In the present study, an attempt has been made to develop a method for predicting proteasome cleavage sites in a protein sequence. The classifier used in this study includes (i) a support vector machine (SVM) (12), (ii) parallel exemplar based learning (PEBLS) (13) and (iii) the Waikato environment for knowledge analysis (Weka) (14). The method has been trained and tested on both '*in vitro* digested data' and 'MHC class I ligand data' with loci for constitutive proteasome as well as immunoproteasome cleavage specificities.

METHODOLOGY

Training data: *in vitro*

The constitutive proteasome cleavage data for yeast enolase I and β-casein were obtained from the work of Toes *et al*. (3)

*To whom correspondence should be addressed. Tel: +91 172 2690557/2690225; Fax: +91 172 2690632/2690585; Email: raghava@imtech.res.in

and Emmerich *et al*. (6), respectively. The residue at the N-terminus of the cleavage site was assigned as the cleavage residue (P1 site) and the remaining residues as non-cleavage sites, as described by Kesmir *et al*. (11).

Training dataset: MHC ligands

In order to develop prediction methods for immunoproteasome and constitutive proteasome cleavage sites, MHC class I ligands were obtained from the MHCBN database (15). The MHCBN database has 1288 HLA-A and HLA-B restricted ligands that are either natural T-cell epitopes or natural peptides eluted from MHC molecules. These MHC ligands were processed (Figure 1) and finally a dataset of 506 ligands, from >250 proteins, was obtained.

Test dataset: independent

We obtained an independent dataset from Saxova *et al*. (1) for evaluating the performance of classifiers trained on the *in vitro* data as well as on MHC ligands. The *in vitro* data consisted of experimental digestion data for the SSX-2, HIV1-Nef and RUI proteins. The MHC ligand dataset consisted of 231 unique, non-overlapping ligands derived from 135 proteins (1).

Implementation of machine learning classifiers

SVM. The SVM was implemented using the freely downloadable software SVM_light (12). For the training of SVM classifiers on *in vitro* digested data, a window size of seven amino acids was chosen, corresponding to a central residue with three amino acids on each side. For the training of classifiers based on MHC class I ligands, a sequence window size of 3–29 amino acids was used, optimized to 19. Each window represents a specific feature; that is, either it represents a cleavage window, if the cleavage site (P1) occurs at its central position, or it represents a non-cleavage window under all other conditions. The actual cleavage site occurs between the central residue (P1) and the C-terminal residue that follows it in the sequence. The classifiers predicted the central residue to be either a cleavage or a non-cleavage site for any particular window configuration. Each amino acid was represented using 21 binary encoding positions (conventional sparse encoding). The 20 amino acids were encoded as A = 10000000000000000000, G = 01000000000000000000 and so on. The twenty-first bit was added to handle the incomplete or terminal parts of proteins.

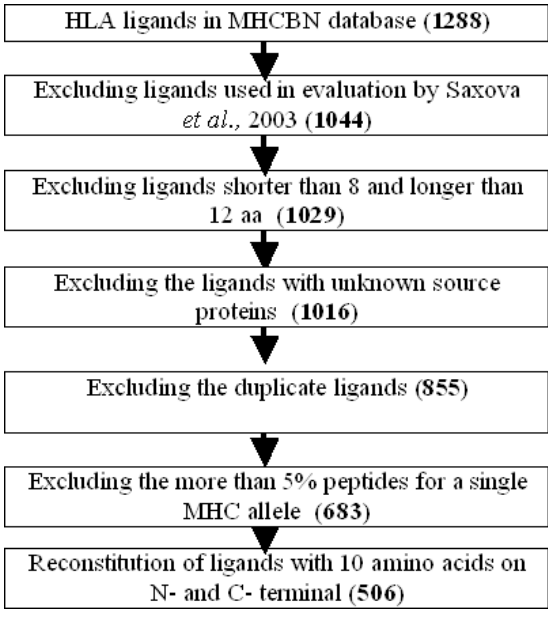

Figure 1. Diagram summarizing the compilation of the MHC class I ligand dataset used in the study.

Weka package. This is a collection of machine learning algorithms written in Java to solve real-world data-mining problems (14). In the present study, we have used three algorithms from Weka: logistic regression, naïve Bayes and J48. The data for Weka are represented in ARFF (attribute relation function format), consisting of a list of all instances, with the attribute values for the instances separated by commas. The results for Weka consist of a confusion matrix for both training and testing, showing the number of instances of each class that have been assigned to each class. Since the dataset of cleavage and non-cleavage sites is unbalanced in our study, Weka's cost-sensitive classification was used, in which the dataset was weighted according to the distribution of cleavage and non-cleavage sites and penalties were assigned to each class in the cost matrix.

PEBLS. PEBLS is a nearest-neighbor learning system designed for applications in which the instances used have symbolic feature values. It treats a set of training examples as points in a multidimensional feature space (13). Test instances are classified by finding the closest example currently contained in the feature space. The nearest neighbors are calculated by computing the distance to each object in the feature space using a modified value distance metric based on the original value distance metric of Stanfill and Waltz. These neighbors are then used to assign a classification to the test instance.

EVALUATION AND PERFORMANCE MEASURES

A 5-fold cross-validation technique was used to evaluate the performance of different classifiers. The dataset of cleavage and non-cleavage sites was divided randomly into five subsets containing equal ratios of cleavage and non-cleavage sites. The classifiers were trained on four sets and performance was assessed on the remaining fifth set. The process was repeated five times so that each set could be used for testing. The average performance of classifiers on five sets is considered to be the final performance. Threshold-dependent parameters (sensitivity, specificity, MCC and accuracy) were used to measure the performance during cross-validation as well as on the independent dataset. Detailed information about the calculation of these parameters can be obtained from the Supplementary Material.

RESULTS AND DISCUSSION

The machine learning classifiers (SVM, PEBLS and Weka) have been trained and tested on *in vitro* digested data. The detailed performance of different classifiers trained using the *in vitro* digested data is shown in Table 1. The SVM and PEBLS based classifiers were able to identify 86.4 and

Table 1. The performance of the different classifiers on *in vitro* digested and MHC ligand data

Classifiers	*In vitro* data				MHC ligand data			
	Sen	Spe	Acc	MCC	Sen	Spe	Acc	MCC
SVM								
RBF	86.4	50.7	68.6	0.42	84.3	69.0	76.7	0.54
POLY	84.6	55.6	70.0	0.43	86.2	65.4	75.8	0.53
PEBLS	57.9	62.9	60.5	0.21	25.3	96.2	88.5	0.28
Weka								
Naïve Bayes	51.3	70.9	61.6	0.23	51.4	91.7	87.3	0.39
J48.PART	50.8	69.2	60.0	0.20	41.1	88.8	83.6	0.27
Logistic	51.9	65.7	58.8	0.18	54.9	88.3	84.6	0.36

Sen: sensitivity; Spe: specificity; Acc: accuracy; MCC: Matthew's correlation coefficient.

57.9%, respectively, of experimentally proven cleavage sites. Three algorithms from the Weka package—(i) logistic regression, (ii) naïve Bayes and (iii) J48.PART (based on the PART rule learner)—were able to identify ~51% of the experimentally proven cleavage sites. These results demonstrate the superiority of the SVM based classifier over other classifiers in the prediction of cleavage sites. The classifier trained on the *in vitro* data can predict only the cleavage sites with constitutive proteasome specificity. In order to develop prediction methods for both constitutive proteasomes and immunoproteasome, we trained machine learning classifiers on the dataset of MHC class I ligands. The natural MHC class I ligands were considered to have major cleavage sites at their C-termini, whereas the rest of the positions between the N-terminus and the C-terminus had minor or no cleavage sites.

The SVM based classifier was able to recognize >84% of cleavage sites, i.e. nearly 30% higher than the other classifiers used in the study. PEBLS was able to recognize most of the non-cleavage sites but showed poor performance in recognizing the cleavage sites. Similarly, classifiers based on Weka (naïve Bayes, J48.PART, logistic regression) were able to recognize the non-cleavage sites with >88% accuracy but showed poor performance in recognizing the cleavage sites (<55%). These results may be due to the unbalanced datasets (in terms of cleavage/non-cleavage) used for training and testing.

Performance on the independent dataset

Threshold-dependent measures. We evaluated the performance of the SVM classifier on the independent dataset. Since none of the patterns of the independent dataset was used for the training and testing of the method, this was effectively an unbiased test to assess the performance of the newly developed classifiers. Table 2 summarizes the performance of SVM based classifiers trained on the *in vitro* data and the MHC class I ligand data. The SVM classifier trained on the *in vitro* digested data was able to recognize ~86% of cleavage sites and ~61% of non-cleavage sites from the independent dataset. The SVM classifier trained on MHC ligands was able to recognize ~82% of cleavage sites and ~45% of non-cleavage sites. The performance of the classifier was poor in recognizing the non-cleavage sites owing to the criteria used to obtain the prediction measure. An incorrect prediction of a non-cleavage site is one where at least one internal position of the MHC ligand had a probability of cleavage that was higher than the threshold as well as the C-terminal. Our SVM based method achieved an MCC of 0.43 for the *in vitro* data and an MCC of 0.29 for the MHC ligand data (Table 2).

Table 2. The performance of the SVM based classifiers on independent data

Methods	Sen	Spe	Acc	MCC
SVM (*in vitro* data)				
RBF	86.9	60.9	68.0	0.43
POLY	85.2	60.9	67.6	0.41
SVM (MHC ligands)				
RBF	82.3	45.0	63.9	0.29
POLY	82.7	41.1	61.9	0.26

Sen: sensitivity; Spe: specificity; Acc: accuracy; MCC: Matthew's correlation coefficient.

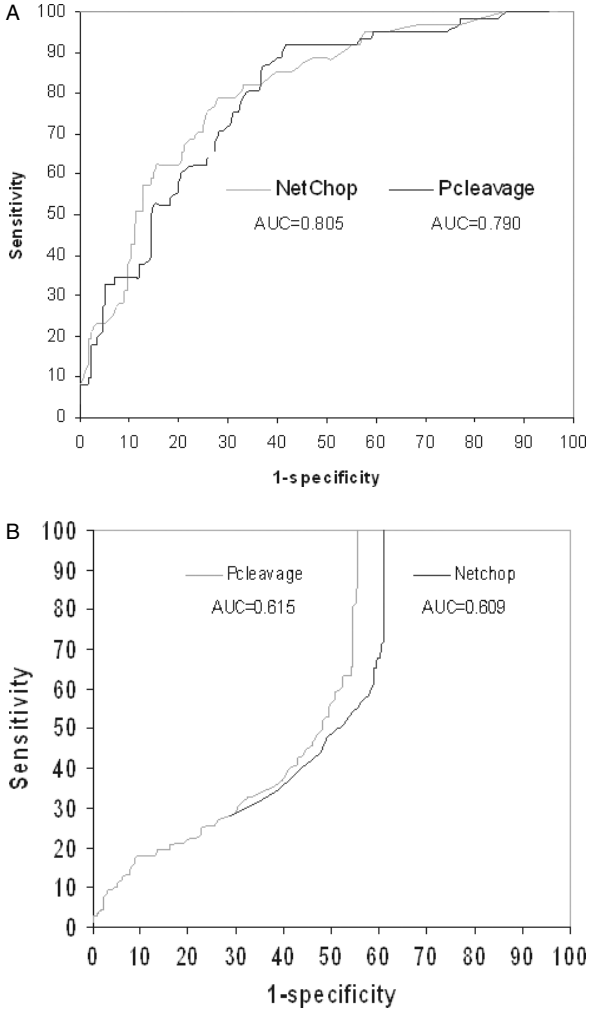

Figure 2. The threshold-independent performance of the Pcleavage and NetChop methods on the independent dataset of Saxova *et al*. (1) (**A**) on *in vitro* digested data and (**B**) on MHC ligand data.

Receiver operator characteristic plot (threshold-independent measures)

The performances of the Pcleavage method and the NetChop method were evaluated on an independent dataset (*in vitro* and

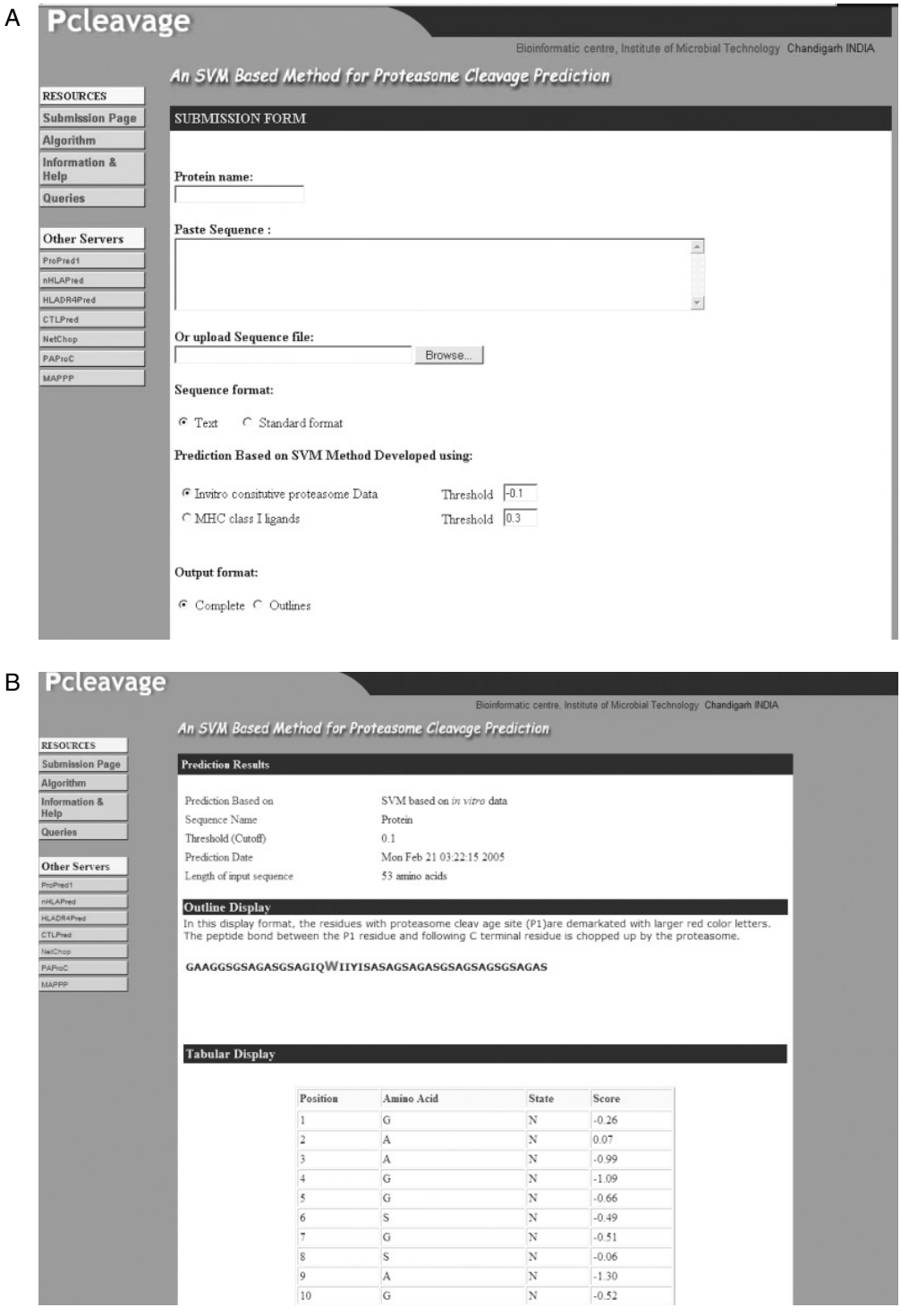

Figure 3. (**A**) Home page of the Pcleavage server. (**B**) A representative sample of Pcleavage prediction results.

MHC ligands) using threshold-independent measures involving a receiver operator characteristic (ROC) plot (Figure 2). The Pcleavage method trained on the *in vitro* digested data achieved an area under the curve (AUC) of 0.790 on the *in vitro* digested independent dataset (Figure 2A). The Pcleavage method trained on the MHC ligand data achieved an AUC of 0.615 on the independent dataset of MHC ligands (Figure 2B). We also evaluated the performance of NetChop 2.0 web version on the independent dataset. NetChop 2.0 was evaluated since it is considered to be the best for predicting *in vitro* proteasomal cleavage. This module achieved an AUC of 0.805 on the *in vitro* digested independent dataset (Figure 2A). On the other hand, the NetChop 2.0 module trained using C-terminal data based on MHC ligands achieved

an AUC of 0.609 on the independent data for MHC ligands (Figure 2B). The analysis demonstrated that the performance of our SVM based classifiers was nearly equal to the performance of the best existing method (NetChop 2.0) on independent data. Therefore, our method will complement existing methods such as NetChop in the prediction of proteasome cleavage sites.

DESCRIPTION OF THE SERVER

Pcleavage is an SVM based method developed for the prediction of constitutive proteasome and immunoproteasome cleavage sites in antigen sequences. The home page of the server is simple and intuitively designed using HTML (Figure 3A).

Input

The server can read amino acid sequences in plain or any standard format (e.g. EMBL, GCG, FASTA). The server uses the ReadSeq program to convert the format of the input sequence. This allows the antigen sequence to be uploaded from files as well as by pasting or typing the sequence for submission.

Options

The web server allows the user to select the SVM classifiers for predicting cleavage sites in a query sequence. It has two classifiers: (i) an SVM trained on the *in vitro* digested data, which is optimized to predict constitutive proteasome cleavage sites; and (ii) an SVM trained on MHC ligands, which is suitable for predicting both constitutive proteasome and immunoproteasome cleavage sites. The server also allows users to select their own cut-off threshold values instead of the default threshold.

Output

The server presents the results in a user-friendly format. The output of the server consists of a short description of the user-defined parameters (such as cut-off threshold, input sequence, SVM classifier) and a mapping of the cleavage sites on the query sequence submitted. The server displays the residues at the first position (P1) of all cleavage sites using larger, red type (Figure 3B). The peptide bond between the P1 residue and the following C-terminal residue is chopped off by the proteasome. The server also allows the complete result to be displayed instead of only outlines. The results are presented in a tabular format. The four columns of the table include the following information: (i) amino acids in the single-letter code; (ii) the position of the amino acid in the sequence; (iii) the prediction score and (iv) the predicted state. A residue is assigned a cleavage state (S) if the predicted score is greater than the threshold. On the other hand, a residue is assigned a non-cleavage state (N) if the predicted score is less than the threshold (Figure 3B).

CONCLUSION AND LIMITATIONS

The method described here is expected to complement existing methods for proteasome cleavage site prediction. The accuracy of proteasome cleavage prediction can be further enhanced by adding more data from T-cell epitopes or naturally processed MHC ligands. A promising computational tool for estimation of T-cell epitopes can be developed by combining this method with MHC and TAP binder prediction methods. However, the use of MHC ligands to develop a method for the prediction of constitutive proteasome and immunoproteasome cleavage sites is not fully correct. The C-terminal of MHC ligands represents only a subset of the cleavages that occur during *in vivo* degradation because not all the degradations result in the production of fragments which can be transferred through the TAP transporter to ER and bind to MHC molecules (5). The MHC ligands represent *in vivo* degradation better than *in vitro* digestion data, so it is advisable to use *in vivo* proteasome degraded data such as natural MHC ligands or T-cell epitopes than *in vitro* digested data (3).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Can Kesmir for his valuable suggestion on proteasome cleavages prediction. We are grateful to Drs R. K. Jain and P. Guptasarma for critically reading the manuscript. Thanks are also due to the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India, for financial assistance. This manuscript carries IMTECH communication no. 044/2003. Funding to pay the Open Access publication charges for this article was provided by DBT, Government of India.

Conflict of interest statement. None declared.

REFERENCES

1. Saxova,P., Buus,S., Brunak,S. and Kesmir,C. (2003) Predicting proteasomal cleavage sites: a comparison of available methods. *Int. Immunol.*, **15**, 781–787.
2. Rock,K.L. and Goldberg,A.L. (1999) Degradation of cell proteins and the generation of MHC class I-presented peptides. *Annu. Rev. Immunol.*, **17**, 739–779.
3. Toes,R.E., Nussbaum,A.K., Degermann,S., Schirle,M., Emmerich,N.P., Kraft,M., Laplace,C., Zwinderman,A., Dick,T.P., Muller,J. *et al.* (2001) Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J. Exp. Med.*, **194**, 1–12.
4. Uebel,S. and Tampe,R. (1999) Specificity of the proteasome and the TAP transporter. *Curr. Opin. Immunol.*, **11**, 203–208.
5. Craiu,A., Akopian,T., Goldberg,A. and Rock,K.L. (1997) Two distinct proteolytic processes in the generation of a major histocompatibility complex class-I presented peptide. *Proc. Natl Acad. Sci. USA*, **94**, 10850–10855.
6. Emmerich,N.P., Nussbaum,A.K., Stevanovic,S., Priemer,M., Toes,R.E., Rammensee,H.G. and Schild,H. (2000) The Human 26 S and 20 S proteasomes generate overlapping but different sets of peptide fragments from a model protein substrate. *J. Biol. Chem.*, **275**, 21140–21148.
7. Holzhutter,H.G., Frommel,C. and Kloetzel,P.M. (1999) A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome. *J. Mol. Biol.*, **286**, 1251–1265.
8. Altuvia,Y. and Margalit,H. (2000) Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism. *J. Mol. Biol.*, **295**, 879–890.

9. Kuttler,C., Nussbaum,A.K., Dick,T.P., Rammensee,H.G., Schild,H. and Hadeler,K.P. (2000) An algorithm for the prediction of proteasomal cleavages. *J. Mol. Biol.*, **298**, 417–429.
10. Nussbaum,A.K., Kuttler,C., Hadeler,K.P., Rammensee,H.G. and Schild,H. (2001) PAProC: a prediction algorithm for proteasomal cleavages available on the WWW. *Immunogenetics*, **53**, 87–94.
11. Kesmir,C., Nussbaum,A.K., Schild,H., Detours,V. and Brunak,S. (2002) Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.*, **15**, 287–296.
12. Joachims,T. (1999) Making large-Scale SVM Learning Practical. In Scholkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel methods—support vector learning*. MIIT Press, Cambridge, MA, London, England, pp. 169–184.
13. Cost,S. and Salzberg,S. (1993) A weighted nearest neighbor algorithm for learning with symbolic features. *Mach. Learn.*, **10**, 57–78.
14. Witten,I.H. and Frank,E. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufman, San Francisco.
15. Bhasin,M., Singh,H. and Raghava,G.P.S. (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics*, **19**, 665–666.