International Immunology, Vol. 15, No. 7, pp. 781–787
doi: 10.1093/intimm/dxg084

Predicting proteasomal cleavage sites: a comparison of available methods

Patricia Saxová[1,2], Søren Buus[3], Søren Brunak[1] and **Can Keşmir[1,4]**

[1]Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Lyngby, Denmark
[2]Institute of Biology and Ecology, P. J. Šafárik University, Kosice, Slovakia
[3]Institute for Medical Microbiology and Immunology, University of Copenhagen, Copenhagen, Denmark
[4]Theoretical Biology/Bioinformatics, Utrecht University, Utrecht, The Netherlands

Keywords: artificial neural network, cleavage site prediction, MHC class I epitope, proteasome, protein degradation

Abstract

The proteasome plays an essential role in the immune responses of vertebrates. By degrading intercellular proteins from self and non-self, the proteasome produces the majority of the peptides that are presented to cytotoxic T cells (CTL). There is accumulating evidence that the C-terminal, in particular, of CTL epitopes is cleaved precisely by the proteasome, whereas the N-terminal is produced with an extension, and later trimmed by peptidases in the cytoplasm and in the endoplasmic reticulum. Recently, three publicly available methods have been developed for prediction of the specificity of the proteasome. Here, we compare the performance of these methods on a large set of CTL epitopes. The best method, NetChop at www.cbs.dtu.dk/Services/NetChop, can capture ~70% of the C-termini correctly. This result suggests that the predictions can still be improved, particularly if more quantitative degradation data become available.

Introduction

Proteasomes are multisubunit proteases that play a central role in the degradation of proteins in the cell (1). Some degradation products of the proteasome are taken up by the transporter associated with antigen processing (TAP) and transferred into the endoplasmic reticulum. Here they can associate with newly synthesized MHC class I molecules. Recognition of such MHC–peptide complexes on the cell surface by activated cytotoxic T lymphocytes (CTL) is essential for the cellular immune responses (2).

The proteasome has at least three different catalytic activities: trypsin-like (i.e. cleavage after basic amino acids), chemotrypsin-like (i.e. cleavage after large, hydrophobic amino acids) and peptidyl-glutamyl-peptide-hydrolyzing activity (i.e. cleavage after acidic amino acids) (3). Since the overall enzymatic activity is the result of an interaction between these catalytic subunits, the cleavage-inhibiting or -enhancing motifs are quite complex. In the presence of IFN-γ, the three catalytic subunits of the proteasomes of vertebrates are replaced by their homologous subunits to form an 'immunoproteasome' (4). The cleavage specificity of the constitutive proteasome and the immunoproteasome seems to be different (5,6), a factor that further increases the complexity of the enzymatic activity of the proteasome.

Due to the involvement of the proteasome in the generation of antigenic peptides it is of general interest to obtain additional insight into the specificity of the proteasome, and to predict which peptides will be generated from both pathogenic and human proteins. At the moment three proteasome cleavage prediction methods are publicly available on the Internet: PAProC (www.paproc.de) developed at Tübingen University (7,8), MAPPP (www.mpiib-berlin.mpg.de/MAPPP/) developed at the Max-Planck Institute in Berlin (9,10) and NetChop (www.cbs.dtu.dk/services/NetChop/) developed at the Center for Biological Sequence analysis at the Technical University of Denmark (11).

PAProC is a method for predicting cleavages by human proteasomes as well as wild-type and mutant yeast proteasomes. The influences of different amino acids at different positions are assessed using a stochastic hill-climbing algorithm (7) based on the experimentally *in vitro* verified cleavage and non-cleavage sites (8).

MAPPP is a method that combines proteasome cleavage prediction with MHC-binding prediction. FragPredict is the part of the MAPPP package that deals with the proteasome cleavage prediction. FragPredict consists of two algorithms. The first algorithm uses a statistical analysis of cleavage-

Correspondence to: C. Keşmir, Theoretical Biology/Bioinformatics, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. E-mail: C.Kesmir@bio.uu.nl

Transmitting editor: M. J. Bevan

Received 15 January 2003, *accepted* 26 February 2003

enhancing and -inhibiting amino acid motifs to predict potential proteasome cleavage sites (9). The second algorithm, which uses the results of the first algorithm as an input, predicts which fragments are most likely to be generated. This algorithm is based on a kinetic model of the 20S proteasome (10) and it takes the time-dependent degradation into account.

NetChop is a neural network-based method trained on MHC class I ligands generated by the human proteasomes. Every MHC ligand has to be generated by the proteasome, therefore the rationale behind using MHC class I ligands is that these ligands bear the closest resemblance to naturally processed *in vivo* cleavage products. However, as some of the products of the proteasome would not bind MHC molecules, MHC class I ligands represent only a subset of *in vivo* cleavage products. The MHC class I ligands used to develop NetChop were compiled from public databases (11). There are two versions of NetChop available, 1.0 and 2.0. The later version is trained with a data set that is 3 times larger.

The aim of this study is to compare the performance of the three publicly available methods mentioned above. Since there is increasing evidence that antigenic peptides result from proteasome cleavage especially at the C-terminal end [see, e.g. (12–15)], we test all the methods on a set of publicly available MHC Class I ligands. We are concerned primarily with the ability of the methods (i) to predict correctly the C-terminal of a ligand and (ii) not to predict *major* cleavage sites within the ligand. We excluded N-terminal cleavage analysis, because the majority of the T cell epitopes are trimmed at their N-terminal by other peptidases, e.g. in the endoplasmic reticulum (15).

We find that the method developed using MHC class I ligands, i.e. NetChop, predicts CTL epitope boundaries more accurately than the methods based on *in vitro* degradation data.

Methods

Performance measurement

We require that a proteasome cleavage prediction method should be able to identify the C-terminal of any natural MHC class I ligand without predicting major cleavage sites within the ligand. Thus, for each ligand we test whether (i) the proteasome cleavage prediction methods can predict the C-terminal cleavage correctly and (ii) the same methods do not predict a cleavage site within the epitope (i.e. all positions except the C-terminal residue) which is more likely than at the C-terminal.

The predictions originate from scores that are compared with a threshold and they are classified as follows:

True positive (TP): if the prediction at the C-terminal, P_c, is above the threshold.

False negative (FN): if P_c is less than the threshold.

True negative (TN): if no cleavages are predicted within the epitope (excluding the C-terminal residue) or if the predicted cleavage sites within the epitope are less likely than at the C-terminal (i.e. less than P_c and the threshold).

False positive (FP): if there is at least one predicted cleavage site within the epitope which is more likely than at the C-terminal (i.e. higher than P_c).

We use the following performance measures to compare NetChop, PAProC and MAPPP:

Sensitivity = TP/(TP + FN)

Specificity = TN/(TN + FP)

$$CC = \frac{TP \times TN - FN \times FP}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}}$$

The sensitivity gives the percentage of C-terminal cleavages that are predicted correctly and the specificity gives the percentage of epitopes with no major predicted cleavage sites (i.e. cleavage sites that are more likely than at the C-terminal) within the epitope. The correlation score, CC, is a measure of how well a method performs *both* in positive (i.e. true cleavage sites) and in negative (i.e. true non-cleavage sites) examples.

Results

Organization of test data set

We focus on the prediction of the specificity of the human proteasome, and therefore we use only peptides associated with HLA-A and HLA-B molecules from the SYFPEITHI database (16) to test various methods. In October 2001 there were 977 unique ligands associated with 120 different HLA-A and HLA-B molecules in the SYFPEITHI database. These ligands are either known T cell epitopes or are naturally processed peptides eluted from MHC molecules. We discarded ligands <8 or >12 amino acids. We also excluded ligands that had already been used for developing NetChop 1.0 or 2.0. The source protein for each ligand was searched for in the SWISSPROT database (17). When an epitope was found in several homologous proteins, homologous proteins were aligned and the most representative protein was chosen unless some additional information about the source protein could be deduced from the original paper. Only epitopes originating from human proteins or from possible human pathogens were included in the data set. The resulting set of 402 peptides contained homologous ligands. In order to prevent possible biases in the analysis, the homologous ligands were excluded using the FASTA (18) and Hobohm-1 algorithms (19). The final set used in our analysis consisted of 249 unique ligands from 135 proteins. The process is described in Fig. 1. The list of ligands is given in Appendix A. Excluding overlapping epitopes, we tested each method on 231 ligands.

Comparison of the methods predicting cleavage by the human proteasome

We use three performance measures to compare the publicly available methods for predicting proteasome cleavage. The formal definitions of these measures are given in Methods. Since there is accumulating evidence that the C-termini of MHC ligands are cleaved precisely by the proteasome, each method should be able to predict the C-terminal of HLA ligands as possible cleavage sites. The sensitivity measure gives the percentage of cleavage sites predicted at the C-terminal of 231 MHC ligands. Note that while all three methods can predict proteasome cleavage sites, only FragPredict can predict fragments generated by the proteasome. In order to

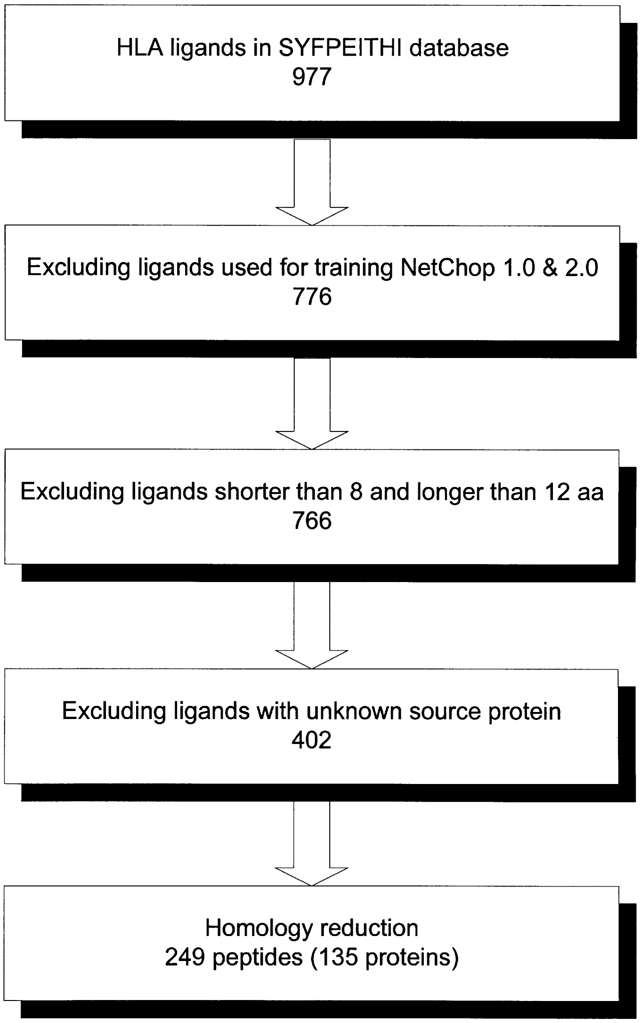
HLA ligands in SYFPEITHI database
977
Excluding ligands used for training NetChop 1.0 & 2.0
776
Excluding ligands shorter than 8 and longer than 12 aa
766
Excluding ligands with unknown source protein
402
Homology reduction
249 peptides (135 proteins)

Fig. 1. Diagram summarizing the compilation of the data set used in this study.

be able to compare the FragPredict method with the other two methods, we use only the prediction of cleavage sites from FragPredict. For FragPredict and NetChop, which produce the probability scores of cleavage for each position in a protein sequence, we used a threshold of 0.5 to classify the predictions, i.e. any position in the sequence with a predicted probability >0.5 is considered as a predicted cleavage site. PAProC does not allow the use of a threshold value for predictions; we assume that the sites with corresponding '+++', '++' and '+' values produced by this method are predicted cleavage sites. The performance measures of the methods for this data set are given in Table 1. FragPredict is able to predict most of the C-termini as cleavage sites, followed by NetChop 2.0. In contrast, PAProC and NetChop 1.0 predict much fewer of the MHC ligand C-termini residues as cleavage sites.

Table 1. The performance of three publicly available methods for the prediction of proteasomal cleavage sites deduced from natural human MHC class I ligands

Method	*N*	Sensitivity	Specificity	CC
PAProC	217	45.6	30.0	–0.25
FragPredict	231	83.5	16.5	0.00
NetChop 1.0	231	39.8	46.3	–0.14
NetChop 2.0	231	73.6	42.4	0.16

N corresponds to the number of natural MHC ligands tested. PAProC requires a flanking region (six positions to the left and four positions to the right of a cleavage site); 14 of the ligands are found at the beginning, or end, of their source protein and could therefore not be analyzed by PAProC. For each ligand, the C-terminal residue should be predicted as a cleavage site. Sensitivity shows the percentage of correct predictions out of *N* true cleavage sites. Specificity shows the percentage of *N* MHC ligands that are predicted as not containing any major cleavage sites. A threshold value of 0.5 was used to classify cleavage and non-cleavage sites. The definitions of the measures are given in Methods. Sensitivity and specificity are in percentages.

An effective prediction method should also be capable of identifying non-cleavage sites (i.e. sites that are not likely to be used by the proteasomes). When the MHC ligands are used as a test set for proteasome cleavage predictions, it is hard to define which sites are really non-cleavage sites. Many CTL epitopes contain minor cleavage sites [see, e.g. (20,21)]. Nevertheless, an epitope should not contain a major cleavage site, i.e. a cleavage site that is more likely than the cleavage site at the C-terminal. Therefore, one can assume that if a method does not predict any major cleavage sites within an epitope, it is able to classify non-cleavage sites correctly. In other words, an incorrect prediction of a non-cleavage site (i.e. a false positive) is one where at least one internal position within an epitope has a probability of cleavage higher than both the threshold *and* the probability of the cleavage at the C-terminal. Following this definition, the total number of true non-cleavage sites becomes the same as the number of epitopes. The specificity measure in Table 1 gives the percentage of the MHC ligands with no major predicted cleavage sites within the ligand. NetChop 1.0 is the most successful method in classifying non-cleavage sites, followed by NetChop 2.0 and PAProC. FragPredict predicts many major cleavage sites within ligands that would make them highly unlikely MHC ligands. The performance of this method does not change much when we use the full FragPredict package (i.e. including the fragment prediction method): 11% of MHC ligands are predicted to stay intact during the protein degradation (using the suggested value of $P > 0.9$). There are other ways of measuring the performance on non-cleavage sites and we have tried many of them, e.g. one can assume that each position within a ligand should have a cleavage probability lower than the threshold. In all cases, the ordering of the methods according to their success in classifying non-cleavage sites correctly did not change (results not shown).

The correlation coefficient (CC) is a measure of how well a method performs *both* on positive (i.e. true cleavage sites) and negative (i.e. true non-cleavage sites) examples. CC = 0 corresponds to random prediction and CC = 1.0 represents 100% correct prediction. A negative CC value means that the predictions are not correlated with the real values. Only NetChop 2.0 has a positive CC (see Table 1). This suggests that NetChop 2.0 generates the most reliable predictions.

Table 2. The performance of three publicly available methods for the prediction of proteasomal cleavage sites identified by *in vitro* degradation studies

Method	Sensitivity	Specificity	CC
PAProC	46.4	64.7	0.10
FragPredict	72.1	41.4	0.12
NetChop 1.0	34.4	91.4	0.31
NetChop 2.0	57.4	76.4	0.32

A threshold of 0.5 was used for FragPredict and NetChop to classify cleavage and non-cleavage sites

Different threshold values can be used in FragPredict and NetChop to classify positions as predicted cleavage sites or predicted non-cleavage sites. When a low threshold is used the methods predict more cleavage sites (and *vice versa* for a high threshold). We investigate the performance measurements of both methods at the standard threshold of 0.5 and at the threshold when the methods reach a maximum correlation coefficient. However, varying the threshold did not change the ranking of the methods according to their performance (results not shown).

The better performance of NetChop may be due to the fact that it was trained using MHC ligands. MHC ligand data reflect not only proteasome specificity, but they also reflect a combined specificity of the proteasome, TAP and MHC. Thus, it cannot be ruled out that NetChop captures this combined specificity and thus performs best when the C-termini of MHC ligands are used for proteasome cleavage predictions. To see if this is the case we also tested all three methods on *in vitro* degradation data generated by the human proteasome. We collected such data from the literature (see Appendix B) excluding the data used to develop PAProC and FragPredict. The results shown in Table 2 confirm that NetChop is able to capture the specificity of the proteasome better than the other methods.

Conclusion

We found that NetChop, an artificial neural network trained with MHC class I ligands, predicts the C-terminal of CTL epitopes more reliably. This is mainly because NetChop can predict the non-cleavage sites better than any of the other methods (see Table 1). There are two possible explanations for this. First, artificial neural networks are much more non-linear than the other two methods. Thus they might capture the complex specificity of the proteasome better. Second, both PAProC and FragPredict are based on very limited set of *in vitro* degradation data, whereas NetChop is trained on a larger data set, i.e. with MHC class I ligands.

The C-termini of MHC ligands represent only a subset of cleavage sites occurring during *in vivo* degradation because not all cleavages would result in protein fragments that can be transferred to the endoplasmic reticulum and can bind to an MHC class I molecule. Thus, the use of MHC ligands to develop a method that can predict proteasome cleavage has been the subject of much criticism (H. Margalit, pers. commun.). However, here we demonstrate that the C-termini of MHC ligands might even represent the specificity of the *in vivo* degradation better than the *in vitro* cleavage maps. Degradation data derived from *in vitro* experiments probably overestimate *in vivo* degradation, because the methods based on this type of data, e.g. FragPredict, predict that most of the MHC ligands in our data set will be destroyed due to major cleavage sites within the ligands.

Even the best method could predict only 73% of the C-termini of natural MHC class I ligands correctly. Moreover, only 42% of the natural MHC ligands are predicted to remain intact. The stochastic nature of degradation (22) and the differences between the immunoproteasome and the constitutive proteasome are just two of many reasons that can explain the poor performance. The use of quantitative data, i.e. concerning not only the cleavage sites used, but also how often a certain site is used, improves the prediction results significantly (C. Kesmir *et al.*, unpublished). Thus, it should be possible to improve on current prediction methods when more quantitative data become available.

In a separate study we found that NetChop 2.0 can correctly discriminate the C-termini of natural MHC ligands from the rest of the protein (results not shown). Thus, NetChop can discriminate the regions that are most likely to be presented to T cells across a protein. This creates a promising future perspective to identify the immunogenic regions in the pathogenic and the human genomes.

Acknowledgements

We thank Lars Juhl Jensen for advice on the statistical evaluation of our results. We are grateful to S. M. McNab for linguistic advice.

Abbreviations

CTL	cytotoxic T lymphocyte
TAP	transporter associated with antigen processing

References

1 Rock, K. L. and Goldberg, A. L. 1999. Degradation of MHC class I-presented peptides. *Annu. Rev. Immunol.* 17:739.

2 Rammensee, H. G., Falk, K. and Rotzschke, O. 1993. Peptides naturally presented by MHC class I molecules. *Annu. Rev. Immunol.* 11:213.

3 Uebel, S. and Tampe, R. 1999. Specificity of the proteasome and the TAP transporter. *Curr. Opin. Immunol.* 11:203.

4 Tanaka, K. and Kasahara, M. 1998. The MHC class I ligand-generating system: roles of immunoproteasomes and the interferon-γ-inducible proteasome activator PA28. *Immunol. Rev.* 163:161.

5 Van den Eynde, B. J. and Morel, S. 2001. Differential processing of class-I-restricted epitopes by the standard proteasome and the immunoproteasome. *Curr. Opin. Immunol.* 13:147.

6 Toes, R. E., Nussbaum, A. K., Degermann, S., Schirle, M., Emmerich, N. P., Kraft, M., Laplace, C., Zwinderman, A., Dick, T. P., Muller, J., Schonfisch, B., Schmid, C., Fehling, H. J., Stevanovic, S., Rammensee, H. G. and Schild, H. 2001. Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J. Exp. Med.* 194:1.

7 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.

8 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.

9 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.

10 Holzhutter, H. G. and Kloetzel, P. M. 2000. A kinetic model of vertebrate 20S proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates. *Biophys. J.* 79:1196.

11 Kesmir, C., Nussbaum, A. K., Schild, H. and Brunak, S. 2002. Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.* 15:287.

12 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.

13 Stoltze, L., Dick, T. P., Deeg, M., Pommerl, B., Rammensee, H. G. and Schild, H. 1998. Generation of the vesicular stomatitis virus nucleoprotein cytotoxic T lymphocyte epitope requires proteasome-dependent and -independent proteolytic activities. *Eur. J. Immunol.* 28:4029.

14 Paz, P., Brouwenstijn, N., Perry, R. and Shastri, N. 1999. Discrete proteolytic intermediates in the MHC class I antigen processing pathway and MHC I-dependent peptide trimming in the ER. *Immunity* 11:241.

15 Mo, X. Y., Cascio, P., Lemerise, K., Goldberg, A. L. and Rock, K. 1999. Distinct proteolytic processes generate the C and N termini of MHC class I-binding peptides. *J. Immunol.* 163:5851.

16 Rammensee, H. G., Bachmann, J., Emmerich, N. N., Bachor, O. A. and Stevanovic, S. 1999. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50:213.

17 Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28:45.

18 Pearson, W. R. and Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* 85:2444.

19 Hobohm, U., Scharf, M., Schneider, R. and Sander, C. 1992. Selection of representative protein data sets. *Protein Sci.* 1:409.

20 Lucchiari-Hartz, M., Van Endert, P. M., Lauvau, G., Maier, R., Meyerhans, A., Mann, D., Eichmann, K. and Niedermann, G. 2000. Cytotoxic T lymphocyte epitopes of HIV-1 Nef: generation of multiple definitive major histocompatibility complex class I ligands by proteasomes. *J. Exp. Med.* 191:239.

21 Morel, S., Levy, F., Burlet-Schiltz, O., Brasseur, F., Probst-Kepper, M., Peitrequin, A. L., Monsarrat, B., Van Velthoven, R., Cerottini, J. C., Boon, T., Gairin, J. E. and Van den Eynde, B. J. 2000. Processing of some antigens by the standard proteasome but not by the immunoproteasome results in poor presentation by dendritic cells. *Immunity* 12:107.

22 Nussbaum, A. K., Dick, T. P., Keilholz, W., Schirle, M., Stevanovic, S., Dietz, K., Heinemeyer, W., Groll, M., Wolf, D. H., Huber, R., Rammensee, H. G. and Schild, H. 1998. Cleavage motifs of the yeast 20S proteasome beta subunits deduced from digests of enolase 1. *Proc. Natl Acad. Sci. USA* 95:12504.

23 Ayyoub, M., Stevanovic, S., Sahin, U., Guillaume, P., Servis, C., Rimoldi, D., Valmori, D., Romero, P., Cerottini, J. C., Rammensee, H. G., Pfreundschuh, M., Speiser, D. and Levy F. 2002. Proteasome-assisted identification of an SSX-2-derived epitope recognized by tumor-reactive CTL infiltrating metastatic melanoma. *J. Immunol.* 168:1717.

Appendix A

Table 3. The list of peptides (including the flanking regions) used in our study

QVPLRPMTYK**AAVDLSHFLKEK**GGLEGLIHSQ	NEF_HV1PV	73
PAATLEEMMT**ACQGVGGPGHK**ARVLAEAMSQ	GAG_HV1BR	338
EAIRFIGRAM**ADRGLLRDI**KAKTAYEKIL	VNUC_INBAA	253
LLGMLMICSA**AENLWVTVYY**GVPVWKDATT	ENV_HV1S3	20
VLEWRFDSRL**AFHHVAREL**HPEYFKNC	NEF_HV1BR	180
MELA**ALCRWGLLL**ALLPPGAAST	ERB2_HUMAN	1
THTVPIYEGY**ALPHAILRL**DLAGRDLTDY	ACTB_HUMAN	160
TAPPAHGVTS**APDTRPAP**GSTAPPAHGV	MUC1_HUMAN	131
TALLKIEGVY**ARDETEFYL**GKRCAYVYKA	R35A_HUMAN	25
QAPSNRVMIP**ATIGTAMYK**LLKHSRVRAY	BRL1_EBV	124
VFDNKFHIIG**AVGIGIAVV**MIFGMIFSMI	CD9_HUMAN	187
LVVSFVVGGL**AVILPPLSPYFK**YSVMINKATP	NI9M_HUMAN	19
LAAGWPMGYQ**AYSSWMYSY**TDHQTTPTFV	EBN3_EBV	34
YGGISLLSEF**CRVLCCYVL**EETSVMLAKR	VIE1_HCMVA	299
GGIGRFYIQM**CTELKLSDY**EGRLIQNSLT	VNUC_IAPUE	34
CGIAVGTTIV**DADKYAVTV**ETRLIDERAA	OM1E_CHLTR	357
IGKMRYVSVR**DFKGKVLI**DIREYWMDPE	P15_HUMAN	65
EELFDFLHAR**DHCVAHKL**FNNLK	UCRH_HUMAN	69
ATLCSALYVG**DLCGSVFLV**GQLFTFSPRR	POLG_HCV1	269
DVDNASLARL**DLERKVESL**QEEIAFLKKL	VIME_HUMAN	208
VIDTLTCGFA**DLMGYIPLV**GAPLGGAARA	POLG_HCV1	122
SALSEGATPQ**DLNTMLNTV**GGHQAAMQML	GAG_HV1BR	172
YLEYRQVPDS**DPARYEFLW**GPRALAETSY	MAG1_HUMAN	248
IVKNIDDGTS**DRPYSHAL**VAGIDRYPRK	RL27_HUMAN	24
RDYFEEYGKI**DTIEIITDR**QSGKKRGFGF	ROA2_HUMAN	129
KEALLDTGAD**DTVLEEMNL**PGKWKPKMIG	POL_HV1RH	75
YARKRSAHTN**DVKQLTEVV**QKVSTESIVI	POL_HV1U4	508
IKKQLGSLVS**DYCNVLNKEF**TAGSVEITLR	BRL1_EBV	18
LVKTGTITTF**EHAHNMRVM**KFSVSPVVRV	EF2_HUMAN	479
ATLYCVHQRI**EIKDTKEAL**DKIEEEQNKS	GAG_HV1BR	82
AMKAYINKVE**ELKKKYGI**	ACBP_HUMAN	69
RGRERFEMFR**ELNEALELK**DAQAGKEPGG	P53_HUMAN	333
LVKLWYQLEK**EPIVGAETFY**VDGAASRETK	POL_HV1BR	589
AQQNNVEHKV**ETFSGVYKK**LTGKDVNFEF	RS7_HUMAN	161
QLEKEPIVGA**ETFYVDGAANR**ETKLGKAGYV	POL_HV1A2	583
GHQAAMQMLK**ETINEEAAEW**DRVHPVHAGP	GAG_HV1BR	192
VCMFLASKLK**ETSPLTAEKL**CIYTDNSIKP	CGD2_HUMAN	104
PSLRILYMTD**EVNDPSLTIK**SIGHQWYWTY	COX2_HUMAN	79
FIMESGAKGC**EVVVSGKLR**GQRAKSMKFV	RS3_HUMAN	125
EGQELSDEDD**EVYQVTVY**QAGESDTDSF	MDM2_HUMAN	264
KTTDGYLLRL**FCVGFTKKR**NNQIRKTSYA	RS3A_HUMAN	127
SLDKLKEVKE**FLGENISNFL**SLAGNTYQLT	APL1_HUMAN	232
AIKWEYVVLL**FLLLADARV**CSCLWMMLLI	POLG_HCV1	713
GPLLVLQAGF**FLLTRILTI**PQSLDSWWTS	VMSA_HPBVW	173
MPAWGAL**FLLWATAEA**TKDCPSPCTC	GPIX_HUMAN	1
EFGATVELLS**FLPSDFFPSV**RDLLDTVSAL	CORA_HPBVJ	8
SGWGSIEPEE**FLTPKKLQCV**DLHVISNDVC	KLK3_HUMAN	155
VPNSDPPRYQ**FLWGPRAYA**ETTKMKVLEF	MGB1_HUMAN	260
QQYRNWFLKE**FPRLKSKL**EDNIRRLRAL	APL1_HUMAN	119
LDIRQGPKEP**FRDYVDRFYK**TLRAEQASQE	GAG_HV1BR	282
RAFTEEGAIV**GEISPLPSL**PGHTDEDVKN	VNS1_IAMAN	148
HHNLLVCSVS**GFYPGSIEV**RWFRNGQEEK	HB2F_HUMAN	140
TVLIIKSLRS**GHDPRAQGTL**	HA2Q_HUMAN	241
KNTMMRKAIR**GHLENNPAL**EKLLPHIRGN	RLA0_HUMAN	57
DLNTMLNTVG**GHQAAMQML**KETINEEAAE	GAG_HV1BR	182
PHHERCSDSD**GLAPPQHLIRV**EGNLRVEYLD	P53_HUMAN	177
DARMQAIQNA**GLCTLVAML**EETIFWLQEI	IE63_EBV	249
IKARAACRAA**GLQDCTMLV**CGDDLVVICE	POLG_HCV1	2717
NSASILPEME**GLSEFTEYL**SESVEVPSPF	MAPB_HUMAN	215
LLVPFVQWFV**GLSPTVWLSV**IWMMWYWGPS	VMSA_HPBVJ	338
AMPHLLVGSS**GLSRYVARL**SSNSRIINHQ	DPOL_HPBVJ	443
VPVKLKPGMD**GPKVKQWPL**TEEKIKALVE	POL_HV1BR	175
VGGVYLLPRR**GPRLGVRAT**RKTSERSQPR	POLG_HCVH	31
PPVLQPIQVM**GQGGSPTAM**AASAVTQAPT	EBN4_EBV	821
RPQDVKFPGG**GQIVGGVYL**LPRRGPRLGV	POLG_HCVH	18
NRFGMDKIYE**GQVEVTGDEY**NVESIDGQPG	RL5_HUMAN	110
IGYSEKDRFQ**GRFDVKIEV**KS	ATNB_HUMAN	283
DAVKVTLGPK**GRNVVLDKS**FGSPTITKDG	CH60_YEREN	25
LIVTRIVELL**GRRGWEALKY**WWNLLQYWSQ	ENV_HV1BR	781
IFHKDLCQAQ**GVALQTMKQ**EFLINLVKQK	FETA_HUMAN	532
SAGATVGIMI**GVLVGVALI**	CEA5_HUMAN	684
EEIWEELGVM**GVYDGREHTV**YGEPRKLLTQ	MAG4_HUMAN	220
YYAMLAKTGV**HHYSGNNIEL**GTACGKYYRV	RL30_HUMAN	61
MN**HLGNVKYLV**IVFLIFFDLF	TRAP_PLAFA	1
TLPALSTGLI**HLHQNIVDV**QYLYGVGSSI	POLG_HCV1	681
WQKLETFWAK**HMWNFISGI**QYLAGLSTLP	POLG_HCV1	1754
TKILEPFRSQ**HPDIVIYQY**MDDLYVGSDL	POL_HV1U4	319
ATPPGSVTVP**HPNIEEVAL**STTGEIPFYG	POLG_HCV1	1349
RPPPGRRPFF**HPVGEADYFEY**HQEGGPDGEP	EBN1_EBV	397
EFIWMCMTVR**HRCQAIRKK**PLPIVKQRRW	EBN4_EBV	139
IKGGRHLIFC**HSKKKCDEL**AAKLVALGIN	POLG_HCV1	1385
TLIGANASFS**IALNFPGSQK**VLPDGQVIWV	PM17_HUMAN	76
GYIKGIVKDI**IHDPGRGAPL**AKVVFRDPYR	RL8_HUMAN	39
EAFSKNLKLG**IHEDSTNRRRL**SELLRYHTSQ	HS9B_HUMAN	430
FLLSLRGAGA**IKADHVSTY**AAFVQTHRPT	HA2Q_HUMAN	22
SVGLGKVLID**ILAGYGAGV**AGALVAFKIM	POLG_HCV1	1841
PAAEHRLREE**ILAKFLHWL**MSVYVVELLR	TERT_HUMAN	530
TRVESENKVV**ILDSFDPLV**AEEDEREISV	POLG_HCV1	2242
WDVLKGSRVS**ILFGHENRV**STLRVSPDGT	GBB5_HUMAN	310
RPILSPLTKG**ILGFVFTLTV**PSERGLQRRR	VMT1_IAPUE	49
PVGEIYKRWI**ILGLNKIVRMY**SPTSILDIRQ	GAG_HV1BR	256
FQNLQVIRGR**ILHNGAYSL**TLQGLGISWL	ERB2_HUMAN	425
SSIVYEAADA**ILHTPGCV**PCVREGNASR	POLG_HCV1	210
AELELAENRE**ILKEPVHGVY**YDPSKDLIAE	POL_HV1BR	466
DLTFLARSAL**ILRGSVAHK**SCLPACVYGP	VNUC_IAPUE	255
GPLCIRMDQA**IMDKNIILKA**NFSVIFDRLE	VNS1_IAPUE	113
LRGTKALTEV**IPLTEEAEL**ELAENREILK	POL_HV1A2	438
VQGACRAIRH**IPRRIRQGL**ERILL	ENV_HV1BR	838
GVVAGGGVAL**IRAASAITA**AGLKGDNEDQ	CH60_YEREN	410
GVD**IRHNKDRKV**RRKEPKSQDI	RL18_HUMAN	1
KISKGANPVE**IRRGVMLAV**DAVIAELKKQ	CH60_HUMAN	130
AIGCVRNLKQ**IVDCLTEMY**YMGTAITTCE	FAFY_HUMAN	1511
LHPDKWTVQP**IVLPEKDSW**TVNDIQKLVG	POL_HV1BR	401
KQGQGQWTYQ**IYQEPFKNLK**TGKYARTRGA	POL_HV1BR	498
LNAWVKVVEE**KAFSPEVIPMF**SALSEGATPQ	GAG_HV1BR	151
IGVGAYGTVY**KARDPHSGHFV**ALKSVRVPNG	CDK4_HUMAN	12
IPYWDWRDAE**KCDICTDEY**MGGQHPTNPN	TYRO_HUMAN	233
ANIQEFAGCK**KIFGSLAFL**PESFDGDPAS	ERB2_HUMAN	359
CGHEALTGTE**KLIETYFSK**NYQDYEYLIN	MYPR_HUMAN	34
LWDQSLKPCV**KLTPLCVTL**NCTNVNGTAV	ENV_HV1MA	110
QVRIKPGSAN**KPKDELDY**ENDIEKKICK	CSP_PLAFA	358
VAAGMNPMDL**KRGIDKAVI**AAVEELKKLS	CH60_YEREN	107
KGKGDKAQIE**KRIQEIIEQ**LDVTTSEYEK	CH60_HUMAN	359
EDQKIGIEII**KRTLKIPAM**TIAKNAGVEG	CH60_HUMAN	459
FSVPLDEDFR**KYTAFTIPSI**NNETPGIRYQ	POL_HV1BR	283
GKRTEQGKEV**LEKARGSTY**GTPRPPVPKP	EBN3_EBV	264
KYAMQLEITI**LIVIGILIL**SVILYFIFCR	E311_ADE03	20
NLPGCSFSIF**LLALLSCLTV**PASAHQVRNS	POLG_HCVH8	52
SYLKGSSGGP**LLCPAGHAV**GIFRAAVCTR	POLG_HCV1	1159
SAHFPGFGQS**LLFGYPVYV**FGDCVQGDWC	TAT_HTL1A	6
TPGTQSPFFL**LLLLTVLTV**VTGSHASST	MUC1_HUMAN	2
EVCNDQVDLY**LLMDCSGSI**RRHNWVNHAV	TRAP_PLAFA	41
LNLTTMFLLM**LLWTLVVLL**ICSSCSSCPL	LMP2_EBV	319
YKCVDRLDKV**LMIIPLINV**TFIISSDREV	VGLH_EBV	532
RDGNNEDNEK**LRKPKHKKL**KQPGDGNPDP	CSP_PLAFA	95
QFLSLQCLQA**LYVDSLFFL**RGRLDQLLRH	MAPE_HUMAN	291
MLLSVPLLLGLLGLAVAEPA	CRTC_HUMAN	1
MMRKLAILSVSSFLFVEALF	CSP_PLAFA	1
NNQGNGQGHN**MPNDPNRNV**DENANANNAV	CSP_PLAFA	290
MVDGTLLLLSSEALALTQT	HLAE_HUMAN	1
RHMQDAEMFT**NAACMALNI**WDRFDVFCTL	OM1E_CHLTR	111
ATMEELQREI**NAHEGQLVI**ARQKVRDAEK	NCAP_HANTV	2
NWMTETLLVQ**NANPDCKTI**LKALGPAATL	GAG_HV1BR	314
QPGYPWPLYG**NEGLGWAGW**LVSPRGSRPN	POLG_HCVTW	78
VGTLEEIIDD**NHAIVSTSV**GSEHYVSILS	PRS4_HUMAN	109
ALINVSANCP**NHFEGHYQY**KSIPVEDNHK	DUS1_HUMAN	202
DKTVALWDLR**NLKLKLHTF**ESHKDEIFQV	RBB7_HUMAN	294
PPWQAGILAR**NLVPMVATVQ**GQNLKYQEFF	PP65_HCMVA	485
TKILEPFRKQ**NPDIVIYQY**MDDLYVGSDL	POL_HV1BR	332
VLKIITFTKN**NQFQALLQY**ADPVSAQHAK	PTB_HUMAN	210
IWGKTPKFKL**PIQKETWETW**WTEYWQATWI	POL_HV1BR	549
TSSSPQPKKK**PLDGEYFTL**QIRGRERFEM	P53_HUMAN	312
QEQIGWMTSN**PPIPVGDIY**KRWIILGLNK	GAG_HV1MA	249
KMPATSRPTA**PPSGKGGNY**PVQQIGGNYT	GAG_SIVSP	117
MKQ**QAGIGILLA**LTTAICWGAL	YHBE_ECOLI	1
VHFKNTRETA**QAIKGMHIR**KATKYLKDVT	RL17_HUMAN	24
VQNIQGQMVH**QAISPRTLNAW**VKVVEEKAFS	GAG_HV1BR	134
DRFYKTLRAE**QASQEVKNW**MTETLLVQNA	GAG_HV1N5	297
ATRDGKLPAT**QLRRHIDLLV**GSATLCSALY	POLG_HCV1	247
QAAADTGHSS**QVSQNYPIV**QNIQGQMVHQ	GAG_HV1BR	116

SCHAASNPPA**QYSWFVNGTF**QQSTQELFIP	CEA5_HUMAN	258	AYRPPNAPIL**STLPETTVVRR**RGRSPRRRTP	CORA_HPBVJ	131
YKNRVASRKC**RAKFKQLL**QHYREVAAAK	BZLF_EBV	180	TSVPAAPPPA**STNRQSGRQ**PTPLSPPLRD	VMSA_HPBVJ	75
GISIKLQEEE**RERRDNYV**PEVSALDQEI	RS17_HUMAN	67	KTCPVQLWVD**STPPPGTRV**RAMAIYKQSQ	P53_HUMAN	139
NNTRKSIRIQ**RGPGRAFVTI**GKIGNMRQAH	ENV_HV1BR	306	NGKRLEPNWA**SVKKDLISY**GGGWRLSAQW	POLG_DEN3	1534
SAPLPPHTTE**RIETRSARH**PWRIRFGAPQ	POLS_RUBVT	254	PKMFAKGTEI**THAVVIKKL**NEILQARGKK	IF38_HUMAN	315
EGSDTITLPC**RIKQIINMW**QKVGKAMYAP	ENV_HV1H2	409	QLQAQHLSHA**THGPPVQL**PPHPSGLQPP	TLE3_HUMAN	127
LRSLCLFSYH**RLRDLLLIVTR**IVELLGRRGW	ENV_HV1BR	765	HHCKLTQVLN**THYVAPRRL**LLTGTPLQNK	SN24_HUMAN	889
GGELDRWEKI**RLRPGGKKKY**KLKHIVWASR	GAG_HV1BR	9	YPYRLWHYPC**TINYTIFK**IRMYVGGVEH	POLG_HCV1	611
MHG**RLVTLKDIV**LDLQPPDPVG	VE7_HPV11	1	VPLAHSSSAF**TITDQVPFSV**SVSQLRALDG	PM17_HUMAN	198
PPSQASSGQA**RMFPNAPYL**PSCLESQPAI	WT1_HUMAN	116	ALEGFDKADG**TLDSQVMSL**HNLVHSFLNG	TYR2_HUMAN	350
VSTVQCTHGI**RPIVSTQLL**LNGSLAEEEV	ENV_HV1A2	245	FQPLHTVMRE**TLFIGSHVV**LRELRLNVTT	VGLH_EBV	410
SGCPERLASC**RPLTDFDQGW**GPISYANGSG	POLG_HCV1	450	GCLLDRKAVG**TPAGGGFPR**RHSVTLPSSK	TISB_HUMAN	33
LNQSVEINCT**RPNNNTRKSI**RIQRGPGRAF	ENV_HV1BR	293	PGFQALSEGC**TPYDINQML**NCVGDHQAAM	GAG_HV2BE	172
QEEEEVGFPV**RPQVPLRPMTY**KAALDISHFL	NEF_HV1A2	65	QEILDLWIYH**TQGYFPDWQNYT**PGPGIRYPLT	NEF_HV1A2	111
QKIETAFLMA**RRARSLSAERY**TLFFDLVSSG	EBN4_EBV	233	EPRGSDIAGT**TSTLQEQIGW**MTNNPPIPVG	GAG_HV1BR	229
ELEVECATQL**RRFGDKLNF**RQKLLNLISK	APR_HUMAN	20	KNQVAMNPTN**TVFDAKRLIGR**RFDDAVVQSD	HS7C_HUMAN	56
GSDSPTLDNS**RRLPIFSRL**SISDD	TISB_HUMAN	315	PAGLKKKKSV**TVLDVGDAY**FSVPLDEDFR	POL_HV1BR	264
GGLEGLIHSQ**RRQDILDLWI**YHTQGYFPDW	NEF_HV1PV	95	IAKITPNNNG**TYACFVSNL**ATGRNNSIVK	CEA5_HUMAN	642
QNPVPVGNIY**RRWIQLGLQK**CVRMYNPTNI	GAG_HV2D2	250	PVSPGDQLPG**VFSDGRVAC**APVPAPAGPI	EBN3_EBV	349
RLIVFPDLGV**RVCEKMALY**DVVTKLPLAV	POLG_HCV1	2578	PGRGEPRFIA**VGYVDDTQF**VRFDSDAASQ	1A01_HUMAN	39
MRVKEKYQHLWRWGWRWGTM	ENV_HV1H2	1	INEEAAEWDR**VHPVHAGPIA**PGQMREPRGS	GAG_HV1BR	204
MRVMAPRALLLLLSGGLALT	1C11_HUMAN	1	LHGMDDPERE**VLEWRFDSRL**AFHHVARELH	NEF_HV1BR	170
QLQARILAVE**RYLKDQQLL**GIWGCSGKLI	ENV_HV1BR	580	TMVAGAVWLT**VMSNTLLSAW**ILTAGFLIFL	LMP2_EBV	432
VGNIVQSCNP**RYSIFFDY**MAIHRSLTKI	EBN3_EBV	104	TITDDVRVQE**VPKLKVCAL**RVTSRARSRI	RL18_HUMAN	84
VDDLRAIAEE**SDEEEAIVAYTL**ATAGVSSSDS	VIE1_HCMVA	368	VLDVGDAYFS**VPLDKDFRKY**TAFTIPSINN	POL_HV1A2	263
HYREVAAAKS**SENDRLRLL**LKQMCPSLDV	BZLF_EBV	199	FPSTAQAQAA**VQGPVGTDF**KPLNSTPATT	Z207_HUMAN	286
SGGDPEIVTH**SFNCGGEFF**YCNSTQLFNS	ENV_HV1H2	365	EQTRSKAGLL**VSDGGPNLY**NIRNLHIPEV	RRP1_IAPUE	581
PGYAGMLGNS**SHIPQSSSY**CSLHPHERLS	ITF2_HUMAN	236	TEARDLHCLL**VTNPHTDAW**KSHGLVEVAS	G45B_HUMAN	112
EKVTWTEAAG**SIRDGVRAY**TALHYLSHLS	QORL_HUMAN	115	KKKYKLKHIV**WASRELERF**AVNPGLLETS	GAG_HV1BR	25
GGSGTYCLNV**SLADTNSLAV**VSTQLIMPGQ	PM17_HUMAN	560	GKWSKSSVIG**WPTVRERM**RRAEPAADGV	NEF_HV1LW	3
TVKTNSVPNM**SLDQSVVEL**YTDTAFSWSV	OM1E_CHLTR	167	LAAMLRQLAQ**YHAKDPNNL**FMVRLAQGLT	PSD2_HUMAN	741
SSTQASLEID**SLFEGIDFYT**SITRARFEEL	HS71_HUMAN	276	TPPLITDYRE**YHTDTTVKF**VVKMTEEKLA	TP2A_HUMAN	950
SAGHTVSGFV**SLLAPGAKQNV**QLINTNGSWH	POLG_HCV1	391	LGFLQRTDLS**YIKSFVSDA**LGTTSIQTPW	EBN3_EBV	148
LSISSCLQQL**SLLMWITQCFL**PVFLAQPPSG	CTG1_HUMAN	147	FPVIFSKASE**YLQLVFGIEV**VEVVPISHLY	MAG2_HUMAN	147
STLPGNPAIA**SLMAFTAAV**TSPLTTSQTL	POLG_HCV1	1779	PQPPICTIDV**YMIMVKCWMI**DSECRPRFRE	ERB2_HUMAN	942
SLTSAQSGDY**SLVIVTTFV**HYANFHNYFV	VGLH_EBV	215	DVGAGVIDED**YRGNVGVVL**FNFGKEKFEV	DUT_HUMAN	183
WGVLAGIAYF**SMVGNWAKV**LVVLLLFAGV	POLG_HCV1	353	FLTLSILDRY**YTPTISRER**AVELLRKCLE	PSB2_HUMAN	137
RCALGVFRKF**SRFPEALRL**ALMLNDMELV	PSD2_HUMAN	250	KKFIRHQSDR**YVKIKRNW**RKPRGIDNRV	RL32_HUMAN	17
TMESSTLELR**SRYWAIRTR**SGGNTNQQRA	VNUC_IAPUE	373	DPASRELVVS**YVNVNMGLK**IRQLLWFHIS	CORA_HPBVO	78
AYLTLAKHTI**SSDYVIPIGTY**GQMKNGSTPM	TYRO_HUMAN	136	MSWRGRST**YYWPRPRRY**VQPPEMIGPM	GGE4_HUMAN	1
PAHLLQDDIS**SSYTTTTTI**TAPPPGVLQN	ACOD_HUMAN	2			

The peptides are shown in boldface. The SWISSPROT accession number of the proteins and the start position follow the sequence.

Appendix B

Table 4. Samples of peptide degradation by the human constitutive proteasome *in vitro*

Cleavage map	Reference
D↓WQN↓Y↓TPGPGVR↓Y↓PL↓TF↓GW↓CY↓KL↓V↓PVEPDK	20
TGSTAV↓PYGSF↓KH↓V↓DT↓RLQ	21
MNGD↓DAF↓ARR↓PTV↓G↓A↓QIPEKIQ↓K↓A↓FD↓DIAKYFSKEEWEKMKA↓SEKIFYV↓Y↓M↓KRKYEAMT↓KL↓GF↓K↓A↓TL↓PPFM↓CN↓KRA↓EDFQGNDL↓DNDPNRGNQVER↓PQM↓T↓F↓G↓RL↓QGISPKI↓MPKKPAEEGNDSEEVPEAS↓GPQNDG↓KEL↓CPPGKPTTSEKIHE↓R↓SGPKRGEHAW↓TH↓RL↓RE↓R↓KQ↓L↓VIY↓E↓EISDPEEDDE	23

Data have been collected from literature to test the performance of three publicly available methods for the prediction of proteasomal cleavage sites; an arrow represents the observed cleavage site.