Protein Engineering vol.15 no.4 pp.287–296, 2002

Prediction of proteasome cleavage motifs by neural networks

Can Keşmir[1,2,6], Alexander K.Nussbaum[3], Hansjörg Schild[3], Vincent Detours[4,5] and Søren Brunak[1]

[1]Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Denmark, [2]Theoretical Biology and Bioinformatics, Utrecht University, The Netherlands, [3]Institute for Cell Biology, Department of Immunology, University of Tübingen, Germany, [4]Santa Fe Institute, Santa Fe, NM and [5]Division of Theoretical Biology, and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, USA

[6]To whom correspondence should be addressed at: Theoretical Biology and Bioinformatics, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands.
E-mail: C.Keşmir@bio.uu.nl

We present a predictive method that can simulate an essential step in the antigen presentation in higher vertebrates, namely the step involving the proteasomal degradation of polypeptides into fragments which have the potential to bind to MHC Class I molecules. Proteasomal cleavage prediction algorithms published so far were trained on data from *in vitro* digestion experiments with constitutive proteasomes. As a result, they did not take into account the characteristics of the structurally modified proteasomes—often called immunoproteasomes—found in cells stimulated by γ-interferon under physiological conditions. Our algorithm has been trained not only on *in vitro* data, but also on MHC Class I ligand data, which reflect a combination of immunoproteasome and constitutive proteasome specificity. This feature, together with the use of neural networks, a non-linear classification technique, make the prediction of MHC Class I ligand boundaries more accurate: 65% of the cleavage sites and 85% of the non-cleavage sites are correctly determined. Moreover, we show that the neural networks trained on the constitutive proteasome data learns a specificity that differs from that of the networks trained on MHC Class I ligands, i.e. the specificity of the immunoproteasome is different than the constitutive proteasome. The tools developed in this study in combination with a predictor of MHC and TAP binding capacity should give a more complete prediction of the generation and presentation of peptides on MHC Class I molecules. Here we demonstrate that such an approach produces an accurate prediction of the CTL the epitopes in HIV Nef. The method is available at www.cbs.dtu.dk/services/NetChop/.

Keywords: artificial neural networks/cleavage site prediction/MHC Class I epitopes/proteasome/protein degradation

Introduction

The proteasome is a multi-subunit cytoplasmic protease that is involved both in the ubiquitin (Ub)-independent and Ub-dependent pathways of protein degradation (Rock and Goldberg, 1999). Protein degradation is a crucial step in many biological processes, including the removal of abnormal proteins, stress response, cell cycle control, cell differentiation and metabolic adaptation. In vertebrates, protein degradation has also a large influence on the immune response of the host. Cytotoxic T cells CTL recognize 8–11 amino acid long protein fragments, presented on the surface of antigen-presenting cells. There is increasing evidence that antigenic peptides result from proteasomal cleavage—in particular at the C-terminal end (Craiu *et al*., 1997; Stoltze *et al*., 1998; Paz *et al*., 1999; Mo *et al*., 1999; Altuvia and Margalit, 2000). The N-terminus is often generated with an extension by the proteasome and is later trimmed by other proteases (Mo *et al*., 1999; Stoltze *et al*., 2000).

Successful prediction of the proteasome cleavage site specificity should be valuable in the design of treatments based on CTL responses. For example, prediction could help in the choice of peptides for use in the treatment of CTL-mediated autoimmune diseases, or in vaccines inducing T-cell-mediated immunity. However, the complexity of proteasomal enzymatic specificity makes such predictions difficult. The core of the eukaryotic proteasome, 20S proteasome, is a complex consisting of 28 protein subunits, 14 of which are unique (Groll *et al*., 1997). The active sites are located in the interior of the proteasome structure. Three catalytic activities were identified, each associated with distinct subunits of the proteasome. These are chymotrypsin-2 like (ChT-L), trypsin-like (T-L) and peptidylglutamyl-peptide hydrolyzing (PGPH) activities (Cardozo *et al*., 1994; Niedermann *et al*., 1996; Heinemeyer *et al*., 1997; Cardozo and Kohanski, 1998). The stimulation with γ-interferon replaces these three catalytically active sites of the proteasome by alternative subunits (Driscoll *et al*., 1993; Gaczynska *et al*., 1993). This form of the proteasome is often referred to as the immunoproteasome. There is a continuing debate on which fraction of the MHC Class I ligands are generated by the immunoproteasome; some data suggests that immunoproteasomes generate mainly the immunodominant epitopes (Van Hall *et al*., 2000; Chen *et al*., 2001). Data-driven methods for cleavage prediction are difficult to implement because experimental data concerning cleavage sites of the proteasome are sparse. As far as *in vitro* degradation by human constitutive proteasome is concerned, the degradations of enolase (Toes *et al*., 2001) and β-casein (Emmerich *et al*., 2000) are the only examples where such experiments were performed and the generated fragments are thoroughly analyzed. Two prediction methods have been developed using these data and some additional *in vitro* peptide degradation data: PAProC (www.paproc.de) (Kuttler *et al*., 2000; Nussbaum *et al*., 2001) and MAPPP (Holzhutter *et al*.,1999; Holzhutter and Kloetzel, 2000). Since the data are limited and relate only to degradation by the constitutive proteasome, these methods may be of limited immunological relevance. Moreover, MAPPP is a linear method, and it may not capture the non-linear features of the specificity of the proteasome. Our aim is to improve these predictions by trying two different approaches: first, we train multi-layered neural networks, a non-linear classification technique, using *in vitro* degradation data. This technique is more powerful than PAProC, which uses a one-

layered network to predict proteasome cleavage. Secondly, we use naturally processed MHC Class I ligands to predict proteasomal cleavage. Since some of these ligands are generated by immunoproteasomes and some by the constitutive proteasome, such a method should predict the combined specificity of both forms of proteasomes.

The neural networks trained on MHC ligands (MHC ligand networks) were able to predict ~65% of the cleavage sites and ~85% of the non-cleavage sites in a test set composed of MHC ligands. The networks trained on the *in vitro* data (constitutive networks) showed a similar performance when tested on the degradation of peptides with the constitutive proteasome. However, when MHC ligand networks were tested on the data generated by the constitutive proteasome, or when constitutive networks were tested on the MHC Class I ligands, the performance values were very low. We also predicted the degradation of a large set of human proteins using both types of networks. The MHC ligand networks generate longer fragments than the constitutive networks. These results suggest that the two networks learn different specificities, i.e. the constitutive proteasome and the immunoproteasome have different, but overlapping specificities, as also suggested by Toes *et al*. (Toes *et al*., 2001).

The presentation of a peptide on an MHC Class I molecule involves at least three steps: degradation by the proteasome, transport to endoplasmic reticulum by TAP and binding to the MHC molecule. Therefore, a combination of the degradation prediction with TAP and MHC binding capacity should be able to give information about the abundance of a peptide being presented. We demonstrate that such a combined approach gives promising results for an HIV protein.

Material and methods

MHC Class I ligand databases

The ligand sequences associated with human MHC Class I molecules were taken from the SYFPEITHI database, a compilation of peptides eluded from MHC molecules (Rammensee *et al*., 1999), at www.uni-tuebingen.de/uni/kxi. Only peptides longer than six amino acids were included. Details of this data collection procedure are given elsewhere (Altuvia and Margalit, 2000). The database contains 229 different peptides extracted from 188 human proteins and associated with 55 human MHC Class I molecules. To prevent biases to a specific MHC binding motif, we made sure that in the final data set no more than 5% of the ligands were bound to a given MHC. In the text we referred to this data set as 'MHC ligands'. This data set is further divided into two, 85% of the sequences are used for the training and the rest are used for testing the performance of the networks.

To find out whether enlarging the data set size could improve the prediction performance, we also extracted ligands from the MHCPEP database (Brusic *et al*., 1998). The MHCPEP database (wehih.wehi.edu.au/mhcpep/) contains 13 000 peptides known to bind MHC. Among these peptides, we included only those (i) which bind to human MHC molecules, (ii) whose flanking regions were possible to reconstruct uniquely, (iii) that are only 8–11 amino acids long, and (iv) that do not originate from HIV proteins (HIV proteins are later used as a test set). This reduction resulted in 881 new ligands, giving a total of 1110 MHC Class I ligands to work on. This data set is referred to in the text as 'Enlarged MHC ligands'.

The network trained on the enlarged MHC ligands set is used to predict the cleavage of C-termini of HIV epitopes. The epitopes were compiled from the HIV Immunology Database (hiv-web.lanl.gov), which is the most comprehensive HIV epitope database for reference strains such as HXB2. The set contains 168 cleavage sites from five HIV proteins (RT, gp160, p17, p24, Nef).

To classify amino acids within a protein sequence into cleavage and non-cleavage sites one needs examples of both types of sites. Neither the MHCPEP nor the SYFPEITHI database contain negative examples, i.e. non-cleavage sites. We used several methods in order to create negative examples. The first method was to label sites within MHC ligands as non-cleavage sites. Our rationale was that the positions within an MHC ligand can only be minor cleavage sites, otherwise the peptide would not be presented on the MHC in the first place. Further, we identified the negative sites that small networks, e.g. networks with only one hidden neuron cannot learn (the large networks can learn all the sites within MHC ligands as negative sites). These sites seem to be different from the other sites within MHC ligands, and thus, they are likely to be potential cleavage sites. These sites were extracted from the training, resulting in a more consistent and 'clean' set of non-cleavage sites. The second method relies on the fact that cleavage site frequency is at the most 24% (Nussbaum *et al*., 1998) per enolase molecule. Thus, labeling random sites as non-cleaved is erroneous in maximally 24% of the cases. Random sequences with amino acid frequencies analogous to frequencies in GenBank were generated and used as non-cleavage site examples. The performance of the networks changed only slightly when different negative sites were used. The results reported here are therefore based on the first method in which any position within an epitope is considered as a non-cleavage site.

Experimental degradation data

For the prediction of cleavage by the constitutive proteasome, we used data on digests of yeast enolase (Toes *et al*., 2001) and bovine β-casein (Emmerich *et al*., 2000) using the human 20S proteasome. Toes *et al*. (Toes *et al*., 2001) extracted the proteasome from human B cells lacking immuno-subunits. This proteasome created 109 fragments from enolase, using 136 distinct cleavage sites. The mean fragment length was 7.4 amino acids. When β-casein was digested using the human 20S proteasome, 63 fragments were produced (48 distinct cleavage sites), having an average length of 18.3 amino acids and a standard deviation of 9.4 amino acids. During training of the neural networks the residues in enolase and β-casein are divided into two groups: the cleavage sites and the non-cleavage sites. The residues on the N-terminus of a verified cleavage (i.e. P1 residue) are assigned as cleavage sites, and all the other residues are assigned as non-cleavage sites.

Sequence logo

We use the Kullback and Leibler information measure to quantify the information content in the cleavage sites and the flanking regions. The purpose of this method is to quantify the contrast between a background distribution and the observed distribution for a given event. Sequence windows centered around the cleavage sites were aligned and the information content was calculated for each position i as:

$$I(i) = \sum_{L=1}^{20} p_i^L \log_2 (p_i^L / q_i^L), \quad (1)$$

where p_i^L is the probability that the amino acid L occurs at position i in a cleavage site window, q_i^L is the probability that a particular amino acid L occurs in a non-cleaved window (background distribution). This information content, expressed in bits/amino acid, was visualized using sequence logos (Schneider and Stephens, 1990).

The neural network algorithm

For this study a standard artificial feed-forward neural network model with one hidden layer of units was used. A neural network uses a network of neurons, where each neuron has multiple inputs and is connected to other neurons, and a single output which produces a non-linear response based on the weighted inputs from these neurons. Each sequence window presenting a specific feature (e.g. in our case either a cleavage window if a cleavage occurs in the middle position or a non-cleavage window) is presented repeatedly to such a network. The weights of the network are initialized randomly. After each iteration of data presentation these weights are adjusted using a standard back-propagation (a gradient descent type) algorithm. The details of this system are given in several other articles (Brunak *et al*., 1991; Baldi *et al*., 1996) and in books (Hertz *et al*., 1991; Baldi and Brunak, 2001).

Each amino acid is represented using 21 binary positions (conventional sparse encoding: Qian and Sejnowski, 1988; Hertz *et al*., 1991) in 21 input neurons. For example, alanine is represented as 100000000000000000000 and cysteine as 010000000000000000000, and so on. The last bit is used for handling incomplete windows in the initial and terminal parts of proteins.

We used sequence windows of size 3 up to 29 amino acids. The central amino acid was designated as either a cleavage or a non-cleavage site, and the actual cleavage site was located between the central residue and the following (C-terminal) residue. For example, the cleavage site L_{251} refers to the cleavage between leucine 251 and residue 252. The same number of flanking residues are used on both sites of the central residue, e.g. a window of five amino acids corresponds to a central residue and two amino acids on each site (P3P2P1P1′P2′ residues for a cleavage site; Berger and Schechter, 1970). For each window configuration, the networks made one prediction for the middle position, assigning the residue to two categories: a cleavage site or a non-cleavage site. Neural networks with 0 to 29 hidden neurons were evaluated for prediction performance. The output of the networks was a score between 0.0 and 1.0. A cleavage was assigned if the network output was larger than a threshold, which is traditionally 0.5. The results reported in this study were obtained using a threshold value of 0.7, to increase the reliability of the predicted cleavage sites. The absolute value of the threshold did not change the correlation coefficients (see below) presented here, but it influences the specificity and the sensitivity. The details of the training procedure can be found elsewhere (Brunak *et al*., 1991; Brunak and Engelbrecht, 1996).

Evaluation of network performance

We evaluated the performance of different neural networks by dividing the entire data sets into a training data set and a test data set. The performance was evaluated using a coefficient of correlation (Matthews, 1975) given by:

$$C = \frac{P_x N_x - N_{fx} P_{fx}}{\sqrt{(N_x + N_{fx})(N_x + P_x)(P_x + N_{fx})(P_x + P_{fx})}} \qquad (2)$$

where P_x is the number of true positives (experimentally verified cleavage sites which are also predicted as cleavage sites), N_x the number of true negatives (experimentally verified non-cleavage sites, predicted as non-cleavage sites), P_{fx} the number of false positives (experimentally verified non-cleavage sites, predicted as cleavage sites) and N_{fx} the number of false negatives (experimentally verified cleavage sites, predicted as non-cleavage sites). Additional performance measurements used in this paper are defined as:

$$\text{Sensitivity} = \frac{P_x}{(P_x + N_{fx})}, \quad \text{Specificity} = \frac{N_x}{(N_x + P_{fx})},$$

$$\text{PPV} = \frac{P_x}{(P_x + P_{fx})}, \quad \text{NPV} = \frac{N_x}{(N_x + N_{fx})},$$

where PPV and NPV stand for positive prediction value and negative prediction value, respectively.

Results

Cleavage inhibiting and promoting sequence motifs

The data used in this paper stem from two different sources. The first set (MHC ligands) comprises 458 cleavage sites determined by MHC Class I ligands of 188 human proteins (Altuvia and Margalit, 2000). The distribution of amino acid residues around the cleavage site for this data set is shown in logo form in Figure 1. The MHC ligand region is shown as dotted positions. Note that the C-terminus cleavage site [i.e. the P1 position, cleavage nomenclature according to Berger and Schechter (Berger and Schechter, 1970)] is included in the MHC ligand. In sequence logos, amino acid symbols are scaled according to their frequencies of occurrence relative to the background distribution. That is, if an amino acid is over-represented, it will get a large height. On the other hand, if it is under-represented, it will also get a large height, but will be given a negative value so that it can be visualized differently, e.g. as an upside down letter. If it occurs at nearly the same frequency as the background distribution, it will have a very small height. In generating this logo we used the amino acid frequencies within the MHC ligand (excluding the last position) to find the background distribution, i.e. the distribution of the amino acids that are not cleaved.

The information content is much higher around the C-terminus than N-terminus (Figure 1), as previously reported by Altuvia and Margalit (Altuvia and Margalit, 2000). This can be due to the involvement of other proteolytic processes on generating N-terminus on MHC Class I ligands (Mo *et al*., 1999; Stoltze *et al*., 2000). When we enlarged this MHC ligand data set, the basic properties of the logo given in Figure 1 did not change (data not shown).

The second data set contains *in vitro* degradation data by human 20S constitutive proteasome for two proteins: enolase (Toes *et al*., 2001) and β-casein (Emmerich *et al*., 2000). A sequence logo based on 184 distinct sites from these two proteins is shown in Figure 2. Here the most significant position is the P1 residue, followed by P2′, P2 and P3. The dominance of the hydrophobic residues (L, V, A) together with the acidic ones (D, E) at these positions is clear, whereas P seems to inhibit cleavage. Comparison of Figures 1 and 2 suggests that the nature of the *in vitro* degradation data is different from MHC Class I ligands. This can be due to the involvement of the immunoproteasome in generation of MHC

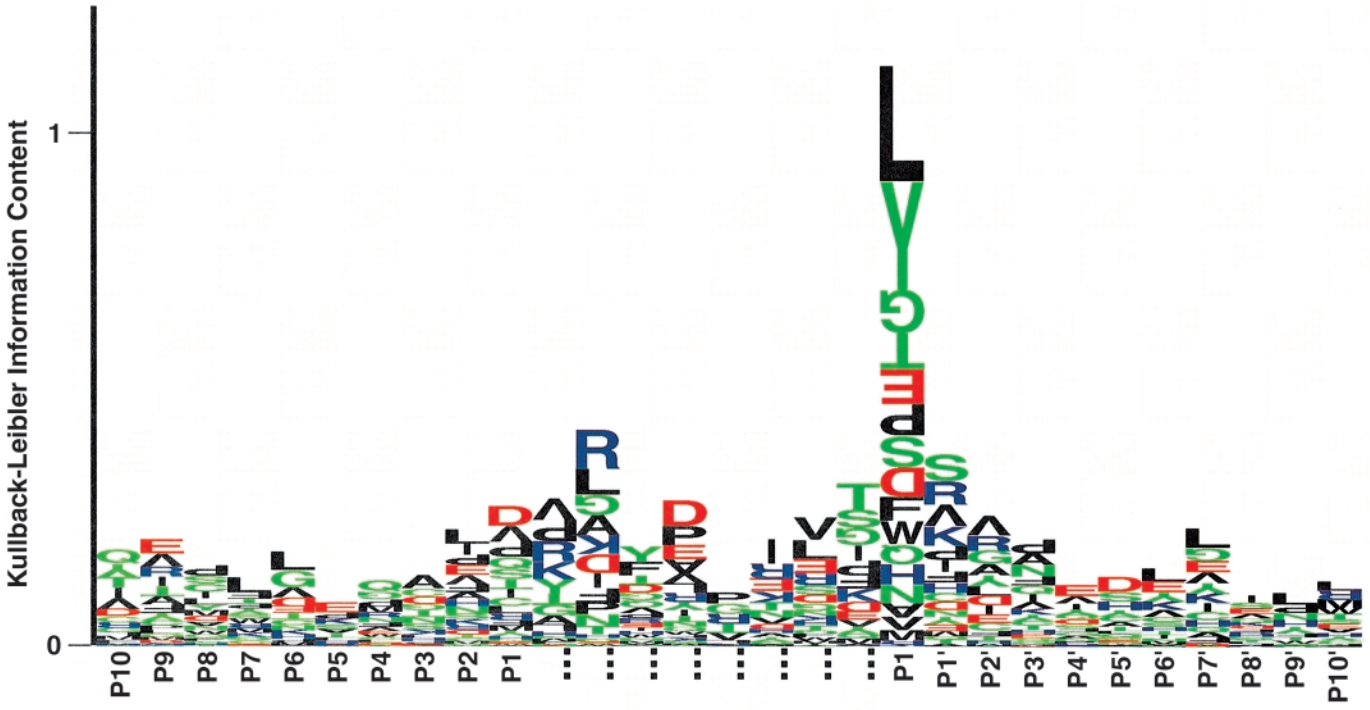

Fig. 1. Sequence logo (as described in Materials and methods) of N- and C-terminal cleavage sites for the MHC ligand database (229 unique sites for both termini). Cleavage nomenclature according to Berger and Schechter (Berger and Schechter, 1970). The level of conservation at each position is computed as the Kullback–Leibler information content. The dotted positions correspond to the MHC Class I ligand. The information content around the C-terminus is much higher than that around the N-terminus. Note that the P1 position for C-termini is the last position of the MHC Class I ligand. Amino acids are color coded according to their physicochemical characteristics. Neutral and polar, green; basic, blue; acidic, red; neutral and hydrophobic, black. Upside-down letters show the amino acids that are under-represented compared to the background distribution.

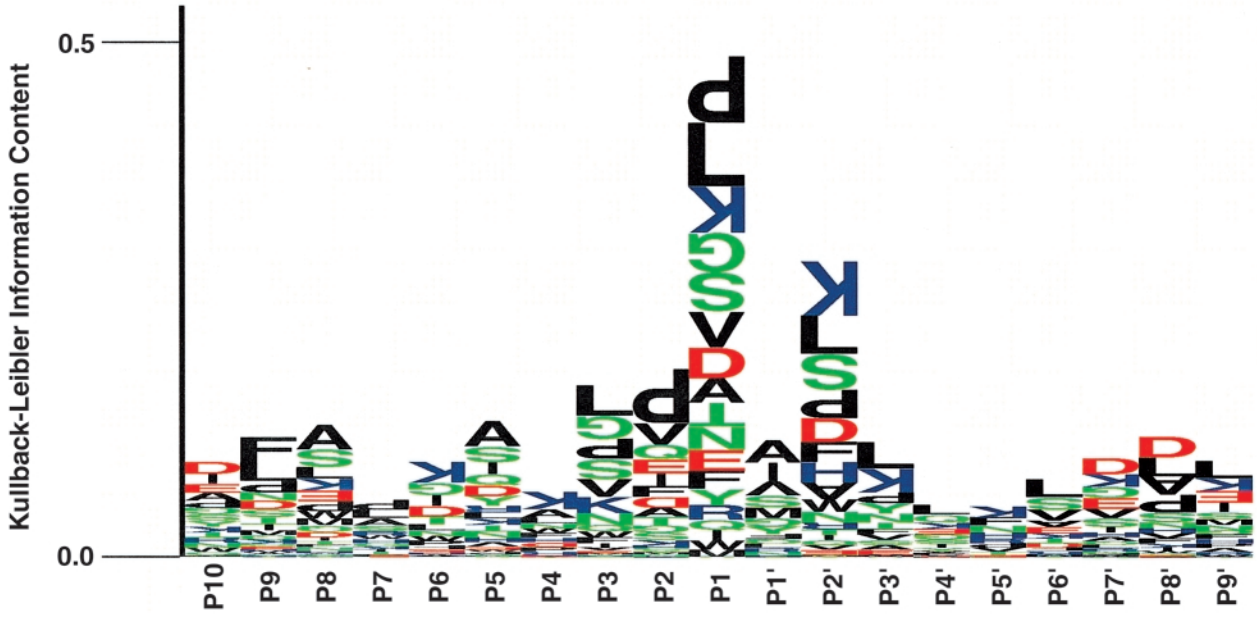

Fig. 2. Sequence logo generated using *in vitro* data on digestion of enolase and β-casein by human 20S constitutive proteasome. 184 distinct cleavage sites were used to create this logo. Color code and the method as in Figure 1.

Class I ligands. However, we did not analyze all the peptides generated by the immunoproteasome; we analyzed only the peptides that bind to MHC molecules. Therefore, this result has to be interpreted with caution.

Sequence features used for discrimination by the network can be extracted by inspecting the weights of individual neurons. In order to enlarge our analysis of cleavage promoting and inhibiting motifs, we analyzed the weights of a linear network trained on the constitutive proteasome data. This network had a seven-residue window and one hidden neuron. In the P1 position large hydrophobic residues (F, L and polar Y) promote cleavage prediction by the network. Proline at P1 and P2 is strictly cleavage inhibiting, whereas at P4 it is cleavage promoting as suggested earlier (Nussbaum *et al.*,

Table I. Cleavage motifs of human constitutive proteasomes

Position	Positive effect on cleavage	Negative effect on cleavage
P1	F, L, Y	P, G, T, N, K
P2	Q, Y, V	P, D
P3	V	G, Q
P4	P, T	D, K
P2′	H	K, S, R, E, P

Cleavage characteristics of human constitutive proteasomes extracted from the analysis of the weights of the artificial neural network. This is a network with one hidden neuron trained on degradation of enolase by human constitutive proteasome and it uses a seven-residue window, giving three-residue flanking regions on each site of the cleavage site.

1998; Shimbara *et al*., 1998). Glycine seems to be cleavage inhibiting when present at positions P1 and P3. The P2′ position may have as much influence as P2; charged residues at P2′, e.g. K, R or E, are cleavage inhibiting. In the P1′ position both experimental results and theoretical studies suggest a preference for small, β-turn promoting amino acids for cleavage (Altuvia and Margalit, 2000; Kuttler *et al*., 2000); however, in our analysis we could not identify this feature. For M, W and C, it was not possible to draw any conclusions since these amino acids have a very low frequency in enolase and β-casein. These results are summarized in Table I. Interestingly, these characteristics are very similar to the ones suggested earlier for the yeast proteasome (Kuttler *et al*., 2000).

Predictive performance of the neural networks

Two networks were trained using the MHC Class I ligands data set: one for the N-termini cleavage site (and its flanking region) and one for the C-termini cleavage site (and its flanking region). The performance of the N-termini network was lower in all the test sets, this is why in Table II, we report only the performance of the C-termini network on the test set. The method is able to predict most of the assigned non-cleavage sites, but has a somewhat poorer performance on the assigned cleavage sites. The final network that was used to obtain these results was one with a 19-residue window and 29 hidden neurons. The networks with small windows (e.g. one with a seven-residue window) have a lower predictive performance, although the difference is not very large. Interestingly, the inclusion of the constitutive proteasome data in our training increased the performance of the networks (Table II, second row). This implies that MHC Class I ligands are not produced solely by the immunoproteasome, and that the use of degradation data from the constitutive proteasome can improve the prediction of these ligands. In an attempt to improve our predictions still further we enlarged the training set of MHC Class I ligands 3-fold by including ligands from the MHCPEP database as well as the ligands used for measuring the performance of the above networks (see Materials and methods). The networks trained on this enlarged data set were used to predict the exact C-termini of MHC Class I epitopes in HIV proteins (Table II, third row). On this data set these networks performed much better than the other methods available (i.e. PAProC and MAPPP mentioned above have a correlation coefficient of ~0.1 on this data set, unpublished results).

For the constitutive proteasome data we measured the performance of the trained networks on five peptides discussed in the literature, which are degraded by human proteasome (Table III). A network trained on the degradation data from enolase and casein can predict 68% of the experimentally verified cleavage sites (Table II). To make the comparison of our results with earlier studies, and to give an idea of what errors the network makes, we printed the full cleavage map of these peptides in Table III. For these peptides our network performed just as well as the best predictor of proteasome cleavage published so far (Kuttler *et al*., 2000). Note that for the peptide data, networks having a small window size (e.g. seven residues), perform best, whereas the large window networks predict MHC Class I ligands best. The networks trained on *in vitro* data predict many cleavage sites within MHC Class I ligands, i.e. these networks predict that many of the MHC ligands are unlikely to remain intact (data not shown). This is partially because the predicted cleavage frequency is higher when *in vitro* degradation is used as a training set. Nevertheless, this observation suggests that the constitutive proteasome might generate fewer MHC Class I ligands than the immunoproteasome.

When predicting the proteasome specificity, one should obviously take the 'cleavage frequency (cleavage strength)' into account. One way of incorporating this additional measure is to use the relative abundance of the specific cleavage, which is available as the initial yield during Edman degradation [for enolase, see Toes *et al*. (Toes *et al*., 2001) and for β-casein Emmerich *et al*. (Emmerich *et al*., 2000)]. Such an approach increases the sensitivity of the constitutive networks, although not significantly (data not shown). This suggests that the prediction performance can be improved as more quantitative data concerning cleavage sites become available.

Networks trained on MHC Class I ligands predict longer fragment length

The predictive ability of the networks trained on MHC Class I ligands can be evaluated further by comparing the predicted fragment length distribution with known data. We estimated the fragment distribution for 4037 human proteins from SWISSPROT (version 38) (Bairoch and Apweiler, 2000). The calculation was based on the cleavage prediction by the network trained on MHC Class I ligands. Results are shown in Figure 3A. We used two approaches to estimate the fragment length distribution. First, we assumed that fragments were not overlapping, i.e. the probability that each predicted site will occur is one. Then, the fragment length distribution is the same as the distribution of the distance between two adjacent predicted cleavage sites. This is plotted as the solid bars in Figure 3A. However, it is known that the cleavage process is highly stochastic [overlapping fragments are very often found in the experimental systems (Nussbaum *et al*., 1998)]. Thus, each predicted cleavage site will be used with a certain probability by the proteasome and some fragments may overlap. To include this effect we used the activity of output neurons (which varies between 0 and 1) as the probability that a cleavage will actually occur at a predicted site. In this way one can repeat say 1000 independent cleavage 'simulations' allowing each cleavage to occur with a probability based on neural network predictions. The fragment distribution obtained after 1000 independent simulated cleavages of human proteins are shown as dotted bars in Figure 3A. When each cleavage occurs only with a certain probability, the frequency of longer peptides is increased.

Kisselev *et al*. (Kisselev *et al*., 1999) analyzed the degradation of three proteins, ovalbumin, casein and insulin-like

Table II. Predictive performance of neural networks

Trained on	Tested on	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Correlation coefficient
MHC ligands	MHC ligands	80	88	44	97	0.53
MHC ligands and 20S	MHC ligands	72	92	53	96	0.56
Enlarged MHC ligands	HIV proteins	66	74	50	85	0.37
20S	Peptides	68	84	70	83	0.53

Values in the table have been rounded to the nearest integer. PPV, positive prediction value reflects the reliability of the positively predicted sites; NPV, negative prediction value reflects the reliability of the negatively predicted sites. For the definition of performance measurements see Materials and methods. In all cases the train and test data sets are independent, i.e. none of the sequences used for the training is included in the test set. The compilation of the data sets are explained in Materials and Methods. 20S stands for the degradation data available by the constitutive proteasome (enolase and β-casein). The peptides are shown in Table III.

Table III. Peptides used for testing the performance of the constitutive proteasome cleavage prediction

Protein name	Data source	Cleavage map	P_x	N_{fx}	P_{fx}
pp89	*Kuckelkorn et al.* (1995)	*RLM*Y⇓D⇓MY⇓PHFMPTNL⇓GPSE⇓K⇓RV*WMS*		–	
	NN	*RLM*Y⇓D⇓M⇓Y⇓PHF⇓M⇓PTNL⇓GPSEKR⇓V*WMS*	4	2	4
OVA	Niedermann *et al.* (1997)	*YVS*GLEQL⇓E⇓SIINF⇓E⇓KL⇓TE⇓*WTS*		–	
	NN	*YVS*GL⇓EQL⇓ESIINF⇓EKLT⇓E⇓*WTS*	3	3	2
OVA	Niedermann *et al.* (1997)	*ALA*M⇓VY⇓L⇓G⇓A⇓KDSTRTQ⇓INKVVR⇓F⇓DKL⇓PGF⇓GD⇓*SIE*		–	
	NN	*ALA*M⇓V⇓Y⇓L⇓GA⇓KD⇓STRTQ⇓INKVV⇓RFDKL⇓PGF⇓GD⇓*SIE*	8	3	3
Nef	Lucchiari-Hartz *et al.* (2000)	*DWQ*N⇓Y⇓TPGPGVR⇓Y⇓PL⇓TF⇓GW⇓CY⇓KL⇓V⇓PVE*PDK*		–	
	NN	*DWQ*N⇓Y⇓TPGPGV⇓R⇓Y⇓PLTF⇓GW⇓CY⇓KLV⇓PVE⇓*PDK*	8	2	2
RU1	Morel *et al*. (2000)	*TGS*TAV⇓PYGSF⇓KH⇓V⇓DT⇓*RLQ*		–	
	NN	*TGS*TAV⇓PYGSF⇓KHV⇓DT*RLQ*	3	2	–

The predictions (given in the rows where data source is indicated as NN) are made by a network (seven-residue window and four hidden neurons), trained on enolase and β-casein data. The references in the table refer to the articles where we collected the data. We included only studies using the human proteasome. The first and the last three residues of each sequence cannot be predicted, since the network needs three-residue flanking. These positions are shown in italics. The threshold used for predictions was 0.5. P_x is true positives, N_{fx} is false negatives (missed cleavage sites), P_{fx} is false positives (wrongly predicted cleavage sites). The arrows indicate the predicted or experimentally verified cleavage sites. Cleavage sites that were found very rarely are not included in the table.

growth factor, with mammalian 26S proteasome *in vitro* and found that (i) 10–15% of peptide bonds are cleaved, (ii) only 15% of peptide products are 8 to 9 amino acids long, (iii) mean peptide length is less than eight amino acids (7.6), and (iv) two thirds of peptides generated are shorter than eight residues. Using the results given in Figure 3A, we found that, in total, 11% of all the peptide bonds were cleaved. 8 to 9Mer peptides made up 13.6% (16.4% when we include chance of overlap, see above) of all the peptides generated. The mean length was 8.9 amino acids (10.2 for overlapping peptides), which is larger than the mean length reported by Kisselev *et al*. (Kisselev *et al*., 1999) but in agreement with the data of Toes *et al*. (Toes *et al*., 2001) for the immunoproteasome. Moreover, we found that 40% of peptides were shorter than eight residues; in other words, the networks trained on MHC ligands tended to predict longer fragments (Figure 3B). When we used the networks trained on the constitutive proteasome data (*in vitro* degradation data), the fragment distribution became closer to the one reported by Kisselev *et al*. (Kisselev *et al*., 1999) (Figure 3B). Since our results are averaged over more than 4000 proteins, the agreement between the predictions and the experimental data is striking.

The main difference between two training sets, MHC Class I ligands and *in vitro* degradation using the constitutive proteasome, is the involvement of the immunoproteasome in the former set. Thus, the above results suggest that the specificity of the immunoproteasome is different from that of the constitutive proteasome. This has been suggested before (Cardozo and Kohanski, 1998; Toes *et al*., 2001; Van den Eynde and Morel, 2001), e.g. the immunoproteasome cleaves more often after hydrophobic amino acid residues, but less often after acidic and aromatic residues (Cardozo and Kohanski, 1998). Moreover, our results suggest that longer peptides can be generated by the immunoproteasome (Figure 3B). This result is in agreement with Toes *et al*. (Toes *et al*., 2001) data, where the average fragment length generated by the immunoproteasome is 8.6 amino acids, and it is 7.4 amino acids for the constitutive proteasome.

Note that the networks are trained only on the specificity of the cleavage sites, not on the optimal length of the fragments generated.

Combination of proteasome cleavage prediction and data on TAP and MHC binding on HIV Nef epitopes

The generation and presentation of peptides on MHC Class I molecules, the availability of responsive T cells, and immunoregulatory effects can all have an influence on whether immune responses are evoked against a particular epitope (Yewdell and Bennink, 1999). As a result, typically one, or a few, potential epitopes elicit a strong CTL response upon immunization with complete antigens (Yewdell and Bennink, 1999). For example, among 51 potential MHC binding peptides in the nucleoprotein and glycoprotein of lymphocytic choriomeningitis virus, only three generate a strong primary immune response (Van der Most *et al*., 1998). A possible explanation for this is that although some of the peptides have a high binding capacity to MHC, they are very unlikely to be generated by the proteasome or transported by TAP into the endoplasmic reticulum and thus they do not evoke a CTL response.

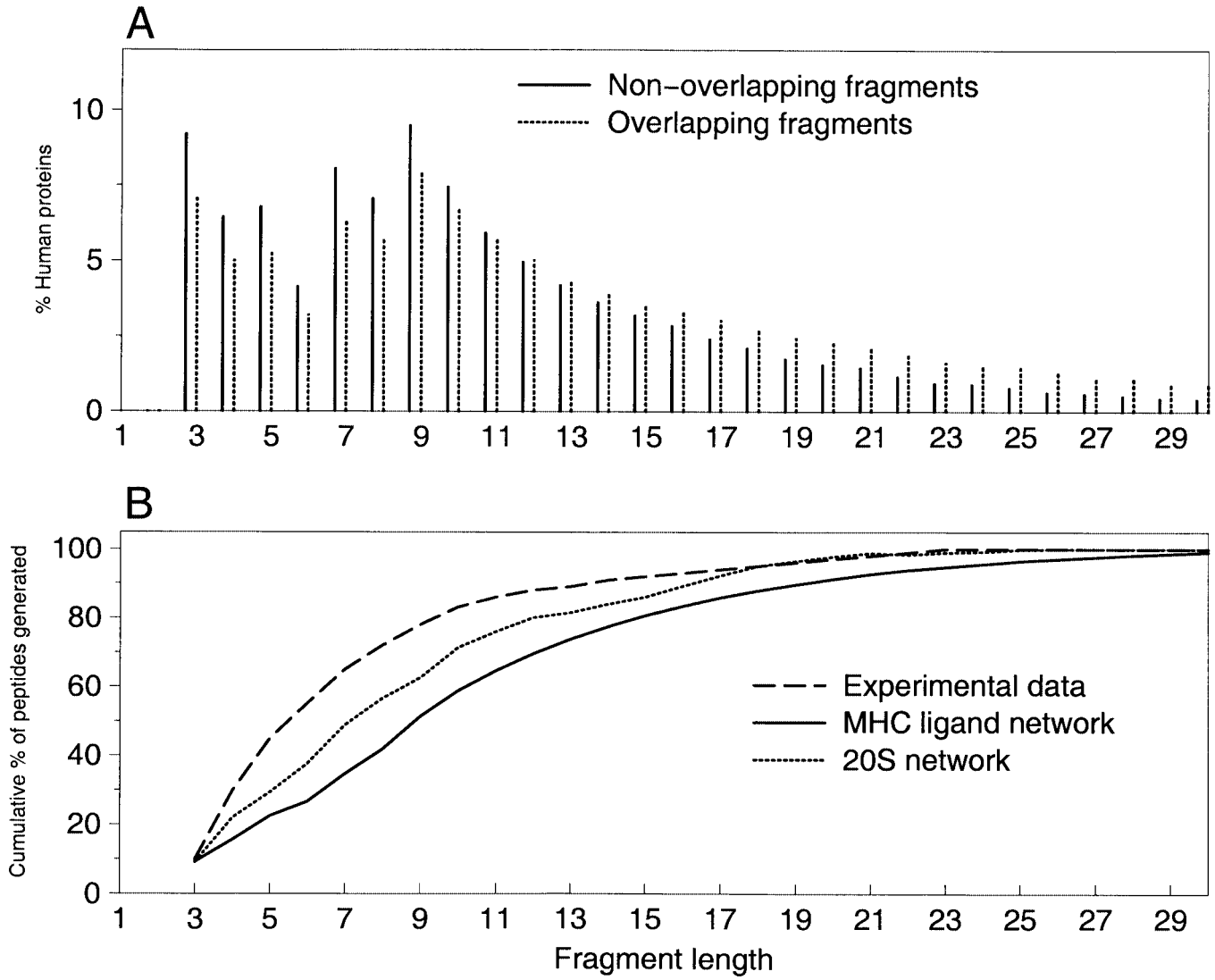

Fig. 3. The fragment length distribution of more than 4000 human proteins from SwissProt (version 38) according to the predicted degradation by the proteasome. To generate this graph we used the network trained on C-terminal cleavage of MHC Class I ligands. (**A**) Solid lines give the estimated fragment distribution based on the distance between two adjacent predicted cleavage sites, i.e. if we assume that every predicted cleavage site is realized. If the probability of a predicted cleavage site to occur is included (dotted lines, see text for details), it is possible to obtain longer peptides. (**B**) Comparison of the predicted fragment distribution with experimental data of Kisselev *et al*. (Kisselev *et al*., 1999) (dashed lines, generated only by the degradation of three proteins). The solid line shows the predicted fragment length of the human proteins by the network trained on the MHC Class I ligands. The dotted line shows the same for the network trained on the constitutive proteasome data. The predictions are made for 4037 human proteins.

Lucchiari-Hartz *et al*. (Lucchiari-Hartz *et al*., 2000) tested this hypothesis by measuring TAP and MHC affinities of five epitopes from the HIV Nef protein (Table IV). We extended their analysis by calculating the probability of a peptide being generated, *P*, by the proteasome. The generation probability of a peptide is determined by two events. First, it has to be cleaved precisely on the C-terminus, and secondly, the rest of the peptide has to remain intact after proteasomal degradation, at least to an extent that allows enough intact peptide to be loaded onto MHC Class I molecule. For each of the peptides discussed in Lucchiari-Hartz *et al*. (Lucchiari-Hartz *et al*., 2000), we calculated P_c, the probability that the C-terminus would be generated correctly, and P_{con}, the probability of not having a major cleavage within the peptide. If we assume that the output of the network is a good measure of the cleavage probability, then $P_{con} = \Pi_{O_i > 0.7}(1 - O_i)$ and $P_c = O_N$, where N is the length of the peptide and O_i is the output of the network for position i. In defining P_{con} we took into account only the sites where a cleavage was predicted, i.e. $O_i > 0.7$. The threshold of 0.7 is used for all the results reported in this study. As there is some evidence that the N-terminus is generated by different proteolytic processes (Craiu *et al*., 1997; Stoltze *et al*., 1998; Mo *et al*., 1999), we did not take into account the probability of generating the N-terminus correctly. The probability of an epitope being generated, *P*, is thus defined as $P = P_c \times P_{con}$. Finally, to combine the effects of all three steps, i.e. degradation, transportation and MHC Class I binding, we define the quality of presentation of a peptide as $Q = P / (A_{TAP} \times A_{MHC})$ where A_{TAP} and A_{MHC} are binding affinities to TAP and MHC Class I molecules, respectively. Please note that higher affinity is reflected in terms of lower A_{TAP} and A_{MHC} values. In other words, peptides with a high probability of being generated and with a high affinity to both TAP and MHC Class I molecules, should get a large Q value.

The results are given in Table IV. The observed number of MHC Class I ligands found in the cell surface and the quality

Table IV. Qualification of antigenicity for HIV Nef protein

Peptide	Sequence	A_{MHC}	P	A_{TAP}	Q ($\times 10^8$)	No. of MHC Class I ligands per cell
Nef136–145	PLTFGWCYKL	295	0.0075	49	50	85
Nef136–146	PLTFGWCYKLV	75	0.001	17	90	125
Nef128–135	TPGPGVRY	30	0.7	160	14583	3600
Nef128–137	TPGPGVRYPL	25	0.0075	195	154	840
Nef135–143	YPLTFGWCY	18	0.057	304	1020	80

Epitopes from HIV-Nef protein were tested for their affinity to TAP and MHC Class I (Lucchiari-Hartz *et al*., 2000). The first two epitopes are HLA-A2 restricted, whereas the others are HLA-B7 restricted. The binding affinity for TAP, A_{TAP}, and MHC Class I, A_{MHC}, and number of MHC Class I ligands per cell values given in this table are experimental values and are adopted from Lucchiari-Hartz *et al*. (Lucchiari-Hartz *et al*., 2000). P is calculated on the basis of our predictions and is a measure of the combined probability of cleavage at the C-terminus and the peptide being conserved (see text for exact definition). The quality of presentation, Q, is defined as $Q = P / (A_{TAP} \times A_{MHC})$. Higher values of Q indicate a larger chance of being presented. There is a good correlation between the Q value and the observed number of MHC Class I ligands per cell.

parameter Q correlate very well for the first three epitopes, whereas for the last two epitopes the correlation is weaker. The above formula used to estimate the probability of a peptide being preserved by the proteasome, P_{con}, is rather simple, which might explain why the correlation for the last two epitopes is not perfect. At the moment we are working on different ways of defining P_{con}.

Taken together, our data indicate that neural network prediction of proteasomal cleavages, in combination with data on MHC Class I binding and TAP transport efficiency, has the power to accelerate the identification of CTL epitopes.

Discussion

Obtaining a better insight into the specificity of the proteasome is an important step in our understanding of many cellular processes, ranging from metabolic adaptation to the regulation of immune responses. We have presented a computational approach whereby the problem can be tackled in two ways: The first way is to predict the specificity of the immunoproteasome partially by using MHC ligand data (which contains only a subset of true fragments created by the immunoproteasome, see below). The second way is to predict the specificity of the constitutive proteasome. These two predictions may, when combined, lead to a more reliable prediction of MHC Class I ligands. Although our specific performance on the available test set sequences can be improved (Table II), the predictions we made using a large human protein database are in agreement with the available experimental data (Figure 3). Moreover, we showed that our predictions for both degradation with constitutive proteasome (Table III) and generation of MHC Class I ligands from viral proteins (Table IV) are in good agreement with experimental findings. Another important result of our analysis is that the networks trained on the MHC Class I ligands and on the constitutive proteasome degradation data learn different, but overlapping specificities. Since the immunoproteasomes are involved in generation of MHC ligands, this result suggests that the specificities of the immunoproteasome and the constitutive proteasome are different, but nevertheless overlap, as was also recently shown by Toes *et al*. (Toes *et al*., 2001). It has been suggested earlier that the flanking regions might play an important role in determining the cleavage site (Del Val *et al*., 1991; Cardozo *et al*., 1994; Nussbaum *et al*., 1998; Theobald *et al*., 1998; Altuvia and Margalit, 2000). Looking at our network architecture, we also suggest that long flanking regions (up to nine amino acids) can influence the cleavage, as the best test performance is obtained with networks having large windows. Finally, we showed that a combination of our prediction methods with TAP and MHC affinity yields a good estimate of how abundantly a peptide can be presented by an antigen-presenting cell (see results for HIV-Nef in Table IV).

Some problems arise with regard to the use of the MHC ligand database to predict the specificity of the proteasome. For instance, many N-termini of MHC ligands seem to be generated by non-proteasomal pathways (Craiu *et al*., 1997; Stoltze *et al*., 1998; Mo *et al*., 1999; Paz *et al*., 1999; Zhou *et al*., 1999; Stoltze *et al*., 2000). Even for the C-termini, it is not possible to rule out the possibility that some exopeptidases might be involved in the post-trimming of precursor peptides generated by proteasomes. Furthermore, there is no direct evidence that MHC ligands are made only by the immunoproteasomes or by the constitutive proteasome. Therefore, a prediction scheme based on MHC ligands will model the combined, systemic specificity of the degradation. Moreover, the C-termini of MHC Class I ligands rarely contain any acidic and basic amino acids. However, the proteasome has been shown to have the enzymatic activities which allow cleavage of peptide bonds to occur immediately after basic and acidic amino acids (Nussbaum *et al*., 1998; Toes *et al*., 2001). Therefore, the use of the MHC ligand database would induce a bias towards other enzymatic activities other than trypsin-like and post-acidic (PGPH) activities. Despite all this, our results regarding the prediction of HIV-Nef epitopes demonstrate that such an approach can lead to good qualitative epitope prediction.

In an earlier theoretical study it was suggested that some side-chain properties of the flanking amino acid residues can be cleavage-determining (Holzhutter *et al*., 1999). We elaborated this idea by testing 450 side-chain properties available in the AAIndex database (Nakai *et al*., 1988). We used the classical Kolmogorov–Smirnoff (Kolmogorov, 1941) test to rank the side-chain properties according to their ability to discriminate a cleavage site from a non-cleavage site. In addition to the free energy of transfer and the volume [as suggested by Holzhutter *et al*. (Holzhutter *et al*., 1999)], several measures of hydrophobicity and other side-chain properties, related to the protein secondary structure, turned out to be possible candidates for discriminating cleavage sites from non-cleavage sites. The majority of the discriminating properties were found for the P1 residue, although some positions like P2, P1′ and P2′ are also important. We used up to 30 of the most significant side-chain properties (common to both MHC ligands and constitutive data) with or without the amino acid sequence for the prediction of cleavage sites. Both of these

approaches resulted in a poorer performance than reported in Table II.

In protein degradation, ubiquitination probably plays the largest role (Yewdell *et al*., 1999). However, once ubiqutinated, the number of predicted cleavage sites within a protein can be used as a measure of resistance to degradation. Interest has focused on the degradation of prion protein and its mutants for many years, as this protein is associated with many neurodegenerative diseases (Kretzschmar, 1999). The human prion protein, PrP, and especially its pathogenesis-associated mutant, PrP145 (a mutant having a stop codon at position 145), are predicted to be easily degraded by our networks. This result together with the experimental evidence (Zanusso *et al*., 1999) suggest that there is hardly any correlation between the degree of degradability and pathogenicity of the prion protein. Further, our networks do not predict that a polyalanine tract will be cleaved by the proteasome. This is an interesting result, since expansions of polyalanine tracts might cause diseases associated with malformation, e.g. synpolyactyly (Goodman *et al*., 1997), cleidocranial dysplasia (Mundlos *et al*., 1997) and oculaopharangeal muscular dystrophy (Brais *et al*., 1998). Another class of triplet repeat disorders is associated with polyglutamine tracts (Koshy and Zoghbi, 1997). We found that these tracts are also resistant to degradation by proteasome.

The results reported in this study show that combination of proteasomal cleavage prediction with data on TAP and MHC affinity yields to a good estimate of epitopes in proteins (see results for HIV-Nef in Table IV). As this combination efficiently identifies CTL epitopes, the combined prediction of these steps in antigen processing would probably also make the search for CTL epitopes quicker. This is very promising for future epitope prediction tools. The methods have been made publicly available at www.cbs.dtu.dk/services/NetChop. Users are encouraged to feedback any experimental confirmation or falsification of the predictions. Any new information regarding verified cleavage sites will also be most welcome. Both types of feedback can be used to retrain the networks to increase performance.

Acknowledgements

The initial MHC ligand database we used in this study was prepared by Yael Altuvia. An early version of this study has benefited greatly from discussions with Jan Hansen and Ramneek Gupta. We thank Søren Buus for his critical reading of the manuscript and his ideas concerning immunological applications. Hanah Margalit's critical comments have improved the manuscript considerably. Claus Andersen has been very helpful in preparing sequence logos.

References

Altuvia,Y. and Margalit,H. (2000) *J. Mol. Biol.*, **295**, 879–890.
Bairoch,A. and Apweiler,R. (2000) *Nucleic Acids Res*., **28**, 45–48.
Baldi,P. and Brunak,S. (2001) *Bioinformatics: The Machine Learning Approach*, 2nd edn. MIT Press, Cambridge, MA.
Baldi,P., Brunak,S., Chauvin,Y. and Krogh,A. (1996) *J. Mol. Biol.*, **263**, 503–510.
Berger,A. and Schechter,I. (1970) *Phil. Trans. R. Soc. Lond. B Biol. Sci.*, **257**, 249–264.
Brais,B., Bouchard,J.P., Xie,Y.G., Rochefort,D.L., Chretien,N., Tome,F.M., Lafreniere,R.G., Rommens,J.M., Uyama,E., Nohira,O. *et al.* (1998) *Nat. Genet.*, **18**, 164–167.
Brunak,S. and Engelbrecht,J. (1996) *Proteins*, **25**, 237–252.
Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) *J. Mol. Biol.*, **220**, 49–65.
Brusic,V., Rudy,G. and Harrison,L.C. (1998) *Nucleic Acids Res.*, **26**, 368–371.
Cardozo,C. and Kohanski,R.A. (1998) *J. Biol. Chem.*, **273**, 16764–16770.
Cardozo,C., Vinitsky,A., Michaud,C. and Orlowski,M. (1994) *Biochemistry*, **33**, 6483–6489.
Chen,W., Norbury,C.C., Cho,Y., Yewdell,J.W. and Bennink,J.R. (2001) *J. Exp. Med.*, **193**, 1319–1326.
Craiu,A., Akopian,T., Goldberg,A. and Rock,K.L. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 10850–10855.
Del Val,M., Schlicht,H.J., Ruppert,T., Reddehase,M.J. and Koszinowski,U.H. (1991) *Cell*, **66**, 1145–1153.
Driscoll,J., Brown,M.G., Finley,D. and Monaco,J.J. (1993) *Nature*, **365**, 262–264.
Emmerich,N.P., Nussbaum,A.K., Stevanovic,S., Priemer,M., Toes,R.E., Rammensee,H.G. and Schild,H. (2000) *J. Biol. Chem.*, **275**, 21140–21148.
Gaczynska,M., Rock,K.L. and Goldberg,A.L. (1993) *Nature*, **365**, 264–267.
Goodman,F.R., Mundlos,S., Muragaki,Y., Donnai,D., Giovannucci-Uzielli,M.L., Lapi,E., Majewski,F., McGaughran,J., McKeown,C., Reardon,W. *et al.* (1997) *Proc. Natl Acad. Sci. USA*, **94**, 7458–7463.
Groll,M., Ditzel,L., Lowe,J., Stock,D., Bochtler,M., Bartunik,H.D. and Huber,R. (1997) *Nature*, **386**, 463–471.
Heinemeyer,W., Fischer,M., Krimmer,T., Stachon,U. and Wolf,D.H. (1997) *J. Biol. Chem.*, **272**, 25200–25209.
Hertz,J., Krogh,A. and Palmer,R. (1991) *Introduction to the Theory of Neural Computation. Studies in the Sciences of Complexity.* Addison-Wesley, Santa Fe Institute.
Holzhutter,H.G. and Kloetzel,P.M. (2000) *Biophys. J.*, **79**, 1196–1205.
Holzhutter,H.G., Frommel,C. and Kloetzel,P.M. (1999) *J. Mol. Biol.*, **286**, 1251–1265.
Kisselev,A.F., Akopian,T.N., Woo,K.M. and Goldberg,A.L. (1999) *J. Biol. Chem.*, **274**, 3363–3371.
Kolmogorov,A. (1941) *Ann. Math. Stat.*, **12**, 461–463.
Koshy,B.T. and Zoghbi,H.Y. (1997) *Brain Pathol.*, **7**, 927–942.
Kretzschmar,H.A. (1999) *Eur. Arch. Psychiatry Clin. Neurosci.*, **249**, 56–63.
Kuckelkorn,U., Frentzel,S., Kraft,R., Kostka,S., Groettrup,M. and Kloetzel,P.M. (1995) *Eur. J. Immunol.*, **25**, 2605–2611.
Kuttler,C., Nussbaum,A.K., Dick,T.P., Rammensee,H.G., Schild,H. and Hadeler,K.P. (2000) *J. Mol. Biol.*, **298**, 417–429.
Lucchiari-Hartz,M., Van Endert,P.M., Lauvau,G., Maier,R., Meyerhans,A., Mann,D., Eichmann,K. and Niedermann,G. (2000) *J. Exp. Med.*, **191**, 239–252.
Matthews,B.W. (1975) *Biochim. Biophys. Acta*, **405**, 442–451.
Mo,X.Y., Cascio,P., Lemerise,K., Goldberg,A.L. and Rock,K. (1999) *J. Immunol.*, **163**, 5851–5859.
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. *et al.* (2000) *Immunity*, **12**, 107–117.
Mundlos,S., Otto,F., Mundlos,C., Mulliken,J.B., Aylsworth,A.S., Albright,S., Lindhout,D., Cole,W.G., Henn,W., Knoll,J.H. *et al.* (1997) *Cell*, **89**, 773–779.
Nakai,K., Kidera,A. and Kanehisa,M. (1988) *Protein Eng.*, **2**, 93–100.
Niedermann,G., King,G., Butz,S., Birsner,U., Grimm,R., Shabanowitz,J., Hunt,D.F. and Eichmann,K. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 8572–8577.
Niedermann,G., Grimm,R., Geier,E., Maurer,M., Realini,C., Gartmann,C., Soll,J., Omura,S., Rechsteiner,M.C., Baumeister,W. *et al.* (1997) *J. Exp. Med.*, **186**, 209–220.
Nussbaum,A.K., Dick,T.P., Keilholz,W., Schirle,M., Stevanovic,S., Dietz,K., Heinemeyer,W., Groll,M., Wolf,D.H., Huber,R. *et al.* (1998) *Proc. Natl Acad. Sci. USA*, **95**, 12504–12509.
Nussbaum,A.K., Kuttler,C., Hadeler,K.P., Rammensee,H.G. and Schild,H. (2001) *Immunogenetics*, **53**, 87–94.
Paz,P., Brouwenstijn,N., Perry,R. and Shastri,N. (1999) *Immunity*, **11**, 241–251.
Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.
Rammensee,H., Bachmann,J., Emmerich,N.P., Bachor,O.A. and Stevanovic,S. (1999) *Immunogenetics*, **50**, 213–219.
Rock,K.L. and Goldberg,A.L. (1999) *Annu. Rev. Immunol.*, **17**, 739–779.
Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.
Shimbara,N., Ogawa,K., Hidaka,Y., Nakajima,H., Yamasaki,N., Niwa,S., Tanahashi,N. and Tanaka,K. (1998) *J. Biol. Chem.*, **273**, 23062–23071.
Stoltze,L., Dick,T.P., Deeg,M., Pommerl,B., Rammensee,H.G. and Schild,H. (1998) *Eur. J. Immunol.*, **28**, 4029–4036.
Stoltze,L., Schirle,M., Schwarz,G., Schroter,C., Thompson,M.W., Hersh,L.B., Kalbacher,H., Stevanovic,S., Rammensee,H.G. and Schild,H. (2000) *Nat. Immunol.*, **1**, 413–418.
Theobald,M., Ruppert,T., Kuckelkorn,U., Hernandez,J., Haussler,A., Ferreira,E.A., Liewer,U., Biggs,J., Levine,A.J., Huber,C. *et al.* (1998) *J. Exp. Med.*, **188**, 1017–1028.
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) *J. Exp. Med.*, **194**, 1–12.
Van den Eynde,B.J. and Morel,S. (2001) *Curr. Opin. Immunol.*, **13**, 147–153.

Van der Most,R.G., Murali-Krishna,K., Whitton,J.L., Oseroff,C., Alexander,J., Southwood,S., Sidney,J., Chesnut,R.W., Sette,A. and Ahmed,R. (1998) *Virology*, **240**, 158–167.
Van Hall,T., Sijts,A., Camps,M., Offringa,R., Melief,C., Kloetzel,P.M. and Ossendorp,F. (2000) *J. Exp. Med.*, **192**, 483–494.
Yewdell,J.W. and Bennink,J.R. (1999) *Annu. Rev. Immunol.*, **17**, 51–88.
Yewdell,J., Anton,L.C., Bacik,I., Schubert,U., Snyder,H.L. and Bennink,J.R. (1999) *Immunol. Rev.*, **172**, 97–108.
Zanusso,G., Petersen,R.B., Jin,T., Jing,Y., Kanoush,R., Ferrari,S., Gambetti,P. and Singh,N. (1999) *J. Biol. Chem.*, **274**, 23396–23404.
Zhou,A., Webb,G., Zhu,X. and Steiner,D.F. (1999) *J. Biol. Chem.*, **274**, 20745–20748.

Received May 29, 2001; revised December 14, 2001; accepted January 4, 2002