BIOINFORMATICS ORIGINAL PAPER

Sequence analysis

Precise score for the prediction of peptides cleaved by the proteasome

Ido Ginodi, Tal Vider-Shalit, Lea Tsaban and Yoram Louzoun*

Department of Mathematics and Statistics, Bar-Ilan University, Ramat-Gan, Israel, 52900

Received on August 1, 2007; revised on December 8, 2007; accepted on December 11, 2007

Advance Access publication January 23, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: An 8–10mer can become a cytotoxic T lymphocyte epitope only if it is cleaved by the proteasome, transported by TAP and presented by MHC-I molecules. Thus most of the epitopes presented to cytotoxic T cells in the context of MHC-I molecules are products of intracellular proteasomal cleavage. These products are not random, as peptide production is a function of the precise sequence of the proteins processed by the proteasome.

Results: We have developed a score for the probability that a given peptide results from proteasomal cleavage. High scoring peptides are those that are cleaved in their extremities and not in their center, while low scoring peptides are either cleaved in their centers or not cleaved in their extremities. The current work differs from most previous works, in that it determines the production probability of an entire peptide, rather than trying to predict specific cleavage sites. We further present different score functions for the constitutive and the immunoproteasome. Our results were validated to have low error levels against multiple epitope databases. We provide here a novel computational tool and a website to use it—http://peptibase.cs.biu.ac.il/PepCleave_II/ to assess the probability that a given peptide results from proteasomal cleavage.

Contact: louzouy@math.biu.ac.il

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Proteins are degraded into short peptides during cell activity. These peptides play numerous roles, including their presentation to cytotoxic T lymphocyte (CTL) (Rock and Goldberg, 1999). Two major paths of intracellular protein degradation exist within eukaryotic cells: the lysosomal pathway and the ubiquitin-proteasome pathway (Hershko and Ciechanover, 1992). Exogenous proteins (e.g. those entering the cell via endocytosis) as well as membrane proteins are usually degraded through the lysosomal pathway, while most of the proteins in the cytoplasm are degraded by the ubiquitin-proteasome system. In mammalian cells, 80–90% of the proteins are subject to proteasomal degradation (Gronostajski et al., 1985; Lee and Goldberg, 1998). The proteasome processes aged, damaged or no longer needed cytosolic proteins, breaking them into 3–20 AA (amino acids) peptides. In most cases, the degraded proteins are previously tagged by a ubiquitin molecule (Ciechanover, 1998; Cox, 1996). Many of the resulting peptides are further degraded by aminopeptidases and other cellular mechanisms, such as nibbling (Kisselev et al., 1999), mainly at the N-terminal (Beninga et al., 1998). Some of these short peptides (mainly those 8–10 AA long) are subsequently translocated by the transporter associated with antigen processing (TAP) channels to the endoplasmic reticulum (ER), where a small part of them are loaded on major histocompatibility complex class-I molecules (MHC-I) and presented to CTL (Rammensee et al., 1993). Proteasomal proteolysis seems to take place in the cell nucleus as well (Rockel et al., 2005).

The proteasome is constructed as a complex of a core enzymatic chamber, the 20s proteasome, and a ‘cup’ of regulatory particles (RP). The 20s proteasome in its basic form (sometimes referred to as a ‘constitutive’ proteasome) is a 700 kDa molecule, composed of two 14 subunit copies, each organized in a structure of four stacked heptametrical rings. Out of the 14 subunits, only 3 have an active site. Following secretion of Interferon-γ (IFN-γ) or tumor necrosis factor-α (TNF-α), the active subunits of the 20s proteasome (β = 1[6, Y], β = 2[MC14, LMP9, Z] and β = 5[MB1, X]) are replaced by their immune counterparts (β = 1[LMP2], β = 2[LMECL1] and β = 5[LMP7], respectively). IFN-γ may also induce another structural change—the increased expression of the 11s regulator (also known as PA28, REG) (Groettrup et al., 1995; Rock and Goldberg, 1999; Tanaka and Kashara, 1998). During an immune response (e.g. an anti-viral or anti-bacterial response), the majority of the cytosolic proteasomes are transformed into their immune form (Khan et al., 2001). 20s proteasomes endowed with one or two RP are called 26s proteasomes (Cox et al., 1996).

Different types of proteasomes are known to produce different cleavage repertoires (Beninga et al., 1998; Chen et al., 2001). The immunoproteasome has been shown to be more specific than the constitutive form. This specificity fits its known involvement in the generation of antigenic peptides presented to CTL (Kesmir et al., 2003). 9mers originating from proteasomal cleavage are the main source for epitopes presented to the immune system in the context of MHC class-I (for a review see Koopmann et al., 1997). The remaining epitopes are either proteasomal cleavage products that are further trimmed, but...
or epitopes resulting from the cross talk between the lysosomal and proteasome pathways (Monu and Trombeta, 2007). In order to predict which peptides can eventually become CTL epitopes, and to understand the specificity of the proteasomal cleavage process, we built a score for the probability that a peptide is indeed a cleavage product. Many cleavage prediction algorithms have already been developed (e.g. Kesmir et al., 2002; Nussbaum, 2001; Peters et al., 2002); for a comparative review see Saxova et al. 2003. These algorithms focused on the detection of cleavage motifs within a given protein. The basic working assumption of these algorithms is the existence of such cleavage sites. The prediction is performed based on the combination of the two AA directly flanking each suspected cleavage site. However, proteasomal cleavage is a rate dependent process (Peters et al., 2002), and the production of a peptide is not solely determined by ‘cleavage sites’. For example, assume three consecutive cleavage sites: A, B and C. Often the peptide A-B is produced, as well as the peptide A-C, but not the peptide B-C. If one were only to predict peptides based on cleavage sites, the peptides A-B and B-C would have been predicted and not the peptide A-C. Thus, in order to detect MHC-I epitope candidates, the cleavage process has to be treated as stochastic. Furthermore, the probability that a given peptide is produced should be based on its length and the configuration of AA within the chain and not on each pair of AA or the region surrounding a cleavage site as proposed in the FragPredict algorithm (Holzhueter et al., 1999). We define a peptide as a cleavage product if it is cleaved in its extremities and not cleaved in its center with a high probability. On the contrary, a peptide is not a cleavage product, provided that it is either cleaved in its center, or not cleaved in at least one of its extremities. Note that since the cleavage process is rate dependent, even peptides with a low production probability may appear in practice, but those should be rare.

We introduce a linear score function, assuming a linear contribution of each AA in the peptide and in its 2-side flanking regions to the total score (that is, an exponential contribution to the production probability, as is further explained in the results section). Altvia et al. have shown that only the first flanking positions of the cleavage site make a significant contribution to the cleavage probability (Altvia and Marguil, 2000). We have therefore only used the first flanking AA. Numerical values for the parameters are estimated using a Simulated Annealing (SA) process. The different specifications of each proteasome yield different scores for the constitutive proteasome and the immunoproteasome. The proposed cleavage score can be accessed at http://peptibase.cs.biu.ac.il/PepCleaveII/.

2 METHODS

2.1 Score function

We denote each peptide by FN – P1, P2, …, Pn – FC, where FN, FC are the flanking positions to its N, C termini, respectively, and P’s are the internal residues. Assuming that each position has an independent contribution, we can roughly write:

\[ P(\text{pep. is produced}) \approx \prod_{\text{positions}} p_{\text{position}} \]

where the positions are FN, P1, …, Pn and FC. We can write

\[ p_{\text{position}} = e^{s_{\text{position}}} \]

and obtain.

\[ P(\text{pep. is produced}) \approx \sum e^{s_{\text{positions}}} \]

Thus a linear score can represent a combination of independent processes. Given a peptide, we thus define a linear score as:

\[ S_{\text{linear}}(\text{FN} - P - \text{FC}) = S_{1}(\text{FN}) + S_{2}(P_{1}) + \sum_{i=2}^{k-1} S_{i}(P_{i}) + S_{i}(P_{n}) + S_{i}(\text{FC}) \]

A peptide with a high \( S_{\text{linear}}(\text{pep}) \) has a high probability of being produced, while a peptide with a low \( S_{\text{linear}}(\text{pep}) \) has a low production probability. This score can be roughly treated as the logarithm of the production probability of a given peptide.

The values \( S_{i}’s \) were learnt on in vitro cleavage experiments, rather than from presented CTL epitopes. We assumed possible different values for \( S_{1}, S_{2}, S_{n} \) to allow for conformation effects. It seems reasonable not to assume identical cleavage behavior in the center and extremities of a peptide, since proteasomal cleavage was shown to be a length-position-rate dependent process.

2.2 Learning set

Products of proteasomal cleavage were extracted from reports on in vitro studies (mainly from Boes et al., 1994; Eggers et al., 1995; Emmerich et al., 2000; Groettrup et al., 1995; Lebiovitz et al., 1994; Nandi et al., 1997; Niedermann et al., 1997; Niedermann et al., 1996; Rivett, 1985; Tenzer et al., 2004; Toes et al., 2001; Wenzel et al., 1994). Duplicates were removed. The data were separated according to the cleaving proteasome type (if reported). The resulting peptides constitute the positive-data learning sets.

The negative learning sets include peptides that were randomized according to the PROWL AA distribution (PROWL). Random AAs were generated using a generic randomization algorithm (Matsumoto and Kurita, 1994). Peptide lengths were distributed uniformly between 6 and 13 AAs. A description of the data sets is available in Supplemental Table 1.

2.3 Learning algorithm for the values of \( S_{1} - S_{5} \)

Optimal values for \( S_{1} - S_{5} \) were learnt using a Simulated Annealing process (Kirkpatrick et al., 1983). The initial configuration of the learnt variables (20 AA × 5 components = 100 variables) was randomized with initial values between −0.5 to 0.5. The Initial temperature was set to \( T_{0}=15 \), and was decreased exponentially \((T_{i} = \lambda \cdot T_{i-1}, \lambda = 0.9)\) in constant intervals. Each mutation was a random change of a magnitude of up to 0.01. The probability of an uphill movement in temperature \( T \) (i.e. accepting a new configuration that is not better than the current one) was set as:

\[ P(\text{uphill}/T) = \exp \left( \frac{\Delta E}{T} \right) = \exp \left( \frac{\Delta S_{\text{conf}}}{T} \right) \]

The score was normalized after each iteration by a member-wise division by a factor of \( 100/\sum_{i=1}^{20} S_{i}^{2} \).

The evaluation function for the process was chosen to be,

\[ f(\text{configuration}) = \frac{1}{G + B} \left( G \frac{\text{pos}^{+}}{\text{positive set size}} + B \frac{\text{neg}^{-}}{\text{negative set size}} \right) \]

where \( \text{pos}^{+} \) and \( \text{neg}^{-} \) are the number of peptides from the positive and negative training sets which the algorithm has managed to learn. The coefficients \( G \) and \( B \) (both are non-negative integers) enable one to
overweight one group or the other. The ratio \([G:B]\) was varied in different learning processes.

The process was executed several times to ensure that the minimum attained could not be significantly improved. We have executed the learning process separately for the constitutive proteasome and for the immunoproteasome, according to the considerations introduced above. The characteristics of the processes are shown in Supplementary Table 2. Finally, we reduced the parameter values to the lowest absolute values that do not affect the success rate over the training sets.

2.4 Validation

Five validation sets were used (Supplementary Table 1, lower section). The positive validation sets contain peptides from the SYFPEITHI (Rammensee et al., 1999) and IEDB (Peters et al., 2005; Sette et al., 2005) databases. The first two positive validation sets were all naturally processed epitopes within the SYFPEITHI database, and MHC eluted epitopes in the IEDB database. In order to produce the third positive validation set, we used blast on the SYFPEITHI human epitopes versus all the genes in the ENSEMBL human genome to locate their origin, and chose only the epitopes with a unique alignment. We extracted the flanking positions of these epitopes, and used the epitopes and their flanking region’s set as a validation for the flanking scores. Note that most MHC eluted epitopes in the IEDB database had no clear curated source. This database may thus be less reliable.

The negative validation set was composed of 1,666 peptides randomized like those of the negative learning sets. These peptides’ lengths were uniformly distributed between 6 and 13 AAs. The last negative validation set was established by randomizing 1,666 peptides (with the same AA distribution), restricting lengths to 8-9 AA (as in the SYFPEITHI set). Peptides from the validation sets were not included in the training sets. For each validation set, we assessed each score using measures of sensitivity (representing a score’s ability to indicate cleavage), specificity (representing the ability to indicate non-cleaved peptides) and a correlation score calculated by (Saxova et al., 2003):

\[
\text{Sen}(\text{Scr}) = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]

\[
\text{Spec}(\text{Scr}) = \frac{\text{TN}}{\text{TN} + \text{FP}}
\]

\[
\text{CC}(\text{Scr}) = \sqrt{\frac{\text{TN} \times \text{TP} - \text{FN} \times \text{FP}}{\text{TP} + \text{FN} \times (\text{TP} + \text{FP})}}
\]

The different validation sets can be accessed in the Pep_Cleave website peptibase.cs.biu.ac.il:64080/PepCleave_II/Validation_sets

3 RESULTS

In order to calculate the probability that a given peptide is indeed the result of proteasomal cleavage, one must take into account the internal part of the peptide and the residues flanking the cleavage site. Given an \(n\) AA long peptide, we define the positions within the peptide as \(P_l - P_n\), and the flanking positions as \(\text{FN, FC}\). A peptide is produced if the positions \(\text{FN} - P_l\) and \(P_n - \text{FC}\) are cleaved and none of the internal positions are cleaved. If the cleavage probability of each position in the proteasome was simply site dependent (as is assumed by existing algorithms), the probability to produce a peptide would be:

\[
w(P_n, \text{FC})w(\text{FN}, P_l) \prod_{l=1}^{n-1} (1 - w(P_l, P_{l+1}))
\]

where \(w(x, y)\) is the probability of having a cleavage site between the residues placed in positions \(x\) and \(y\). Such a score would predict that the majority of peptides would be very short (since \(\prod_{l=1}^{n-1} (1 - w(P_l, P_{l+1}))\) decreases with peptide length). Moreover, given three neighboring ‘cleavage sites’ (A, B, C), cleavage experiments often show that the peptides A-C and A-B are formed, but not the peptide B-C (Tenzer et al., 2004), which would be predicted taking into account only cleavage points. Actually, one often observes even more complex situations with four cleavage sites where peptides A-C and B-D are formed but not A-B, B-C or C-D (Nussbaum et al., 1998).

The opposite, more precise description would require a probability score specifically adapted to each position: \(w(P_n, FC)w(\text{FN}, P_l) \prod_{l=1}^{n-1} (1 - w(P_l, P_{l+1}))\), where each \(w_i\) is different from the others. This would require a huge learning set. Here we propose a middle way of a score depending differently on the internal residues and the C and N termini. We define a score for the probability that a given peptide \(P\) shall be produced during cleavage, \(S_{\text{cleave}}(\text{FN} - P - \text{FC})\), as a linear combination of the effect of the AAs in the different positions:

\[
S_{\text{cleave}}(\text{FN} - P - \text{FC}) = S_1(\text{FN}) + S_2(P_l) + \sum_{i=2}^{n} S_i(P_i) + S_4(\text{FC})
\]

We performed a SA algorithm in order to learn optimal values for the values of \(S_1 - S_4\) in the score function. We first learnt the most basic configuration ([BSC]—basic) of all peptides measured to be cleaved. We then extended the analysis to four more different types of scores for specific proteasomes (e.g. the immunoproteasome). The score was learnt on two learning sets. The positive training set consisted of cleavage products taken from in vitro experiments (Boes et al., 1994; Eggers et al., 1995; Emmerich et al., 2000; Groettrup et al., 1995; Leibovitz et al., 1994; Nandi et al., 1997; Niedermann et al., 1997; Niedermann et al., 1996; Rivett, 1985; Tenzer et al., 2004; Toes et al., 2001; Wenzel et al., 1994). For some experiments, the proteasome type was specified. These results were used for specific training sets, namely a set of constitutive proteasome cleavage products and a set of immunoproteasome cleavage products.

The negative training set was composed of random peptides of length uniformly distributed between 6 to 13 AA and a typical AA distribution (PROWL). We used different negative training sets (for the different scores) to match the length distribution of the peptides in the positive learning set and the fraction of peptides with flanking positions. In both the negative and positive learning sets, 83.3–96.1% of the peptides had flanking regions.

The basic score represents the probability that a peptide would be produced, independently of the specific proteasome used by the cell. This score actually represents the ‘average effect’ of all proteasome types. Given the fact that in a typical experimental setup, the precise proteasome type used by the cell is unknown, such an average score is needed. Furthermore, multiple proteasome types may be used simultaneously by the cell. Thus such a score can actually represent a biological reality (Brooks et al., 2000). The SA produces a better predictor (on both the learning and validation sets) than any existing
cleavage prediction algorithm. In contrast with existing algorithms, we do not claim to precisely predict all possible cleavage sites; rather, we claim to predict the existence of cleavage products, which we do predict precisely. In order to test the prediction of the algorithm, we used five different validation sets. Each set tested a different aspect of the score quality:

- The first validation set contained 1.66 random peptides with AA and length distribution similar to the ones used in the negative learning set. This set was used to check that the learning process can be extended to similar random sets.
- The second and third sets consisted of reported naturally processed epitopes from the SYFPEITHI and IEDB databases. These epitopes are supposed to be cleaved before presentation. This set is used as a positive validation of the internal positions of $S_{\text{cleave}}$ (FN $P$ FC) (i.e. $S_2(P_1)$, $S_3(P_3)$, $S_4(P_4)$), since these epitopes do not contain flanking AAs.
- The Fourth set is composed of naturally processed SYFPEITHI human epitopes, for which a single copy was found in the human genome. The origin of these epitopes in the human genome was located, and the flanking residues were extracted. This validation set allows us to test the score for the flanking regions. This dataset is much smaller than the previous one, and was only established to show that the inclusion of the flanking regions of the peptide improves the prediction capacity of the score.
- The fifth set is SYFPEITHI like random peptides. These peptides have a similar AA usage as the first negative validation set, but the length of the SYFPEITHI epitopes and without flanking regions. It was used to check that the score performance on the real SYFPEITHI is not merely a result of the peptide structure.

Before comparing our results with CTL epitopes, one must note that our score does not directly predict CTL epitopes, since biochemical mechanisms beyond proteasomal cleavage are involved in the epitope production process. An example for this would be nibbling, which is known to affect the N-terminus rather than the C-terminus of a peptide. Generally, post-proteasomal trimming of the N-terminus has been shown to be particularly relevant in the immunoproteasome case (Beninga et al., 1998). Our score is still a good approximation for the process leading to epitope production as can be seen from the validation results (Table 1). We are currently developing a trimming algorithm to bridge the gap between the cleavage prediction and MHC presentation.

The \[BSC\] cleavage score performs almost as well in the validation sets as in the learning sets (Table 1, second column). Its false positive and false negative levels are on average, 24 and 27%, respectively. It obviously performs better with flanking than without the flanking residues (24.5% error versus 27%), but the difference is not large. Note also that we denote any predicted random peptide as a false positive. Actually a large number of those can be real epitopes. In \textit{in vitro} experiments, at least 4-5% of the peptides are measured to be produced.

### Table 1. Absolute performance of the scores developed

<table>
<thead>
<tr>
<th>Score</th>
<th>BSC</th>
<th>GOW</th>
<th>BOW</th>
<th>CON</th>
<th>IMM</th>
<th>CON-IMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1Pos. Tr. (1127)</td>
<td>897</td>
<td>1016</td>
<td>769</td>
<td>257</td>
<td>303</td>
<td></td>
</tr>
<tr>
<td>Frac</td>
<td>0.80</td>
<td>0.90</td>
<td>0.68</td>
<td>0.91</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Syf. w. fl. (81)</td>
<td>24</td>
<td>65</td>
<td>52</td>
<td>24</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Frac</td>
<td>0.74</td>
<td>0.80</td>
<td>0.64</td>
<td>0.30</td>
<td>0.53</td>
<td>0.64</td>
</tr>
<tr>
<td>Syf w.o. fl. (558)</td>
<td>388</td>
<td>435</td>
<td>321</td>
<td>185</td>
<td>319</td>
<td>373</td>
</tr>
<tr>
<td>Frac</td>
<td>0.70</td>
<td>0.78</td>
<td>0.58</td>
<td>0.33</td>
<td>0.57</td>
<td>0.67</td>
</tr>
<tr>
<td>IEDB (913)</td>
<td>615</td>
<td>665</td>
<td>469</td>
<td>240</td>
<td>491</td>
<td>557</td>
</tr>
<tr>
<td>Frac</td>
<td>0.67</td>
<td>0.73</td>
<td>0.51</td>
<td>0.26</td>
<td>0.54</td>
<td>0.61</td>
</tr>
<tr>
<td>Neg. Tr. (5500)</td>
<td>4255</td>
<td>3566</td>
<td>4748</td>
<td>4516</td>
<td>4381</td>
<td></td>
</tr>
<tr>
<td>Frac</td>
<td>0.77</td>
<td>0.65</td>
<td>0.86</td>
<td>0.87</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>Rand (gen) (1.e6)</td>
<td>751168</td>
<td>627613</td>
<td>839635</td>
<td>861900</td>
<td>784212</td>
<td>719491</td>
</tr>
<tr>
<td>Frac</td>
<td>0.75</td>
<td>0.63</td>
<td>0.84</td>
<td>0.86</td>
<td>0.78</td>
<td>0.72</td>
</tr>
<tr>
<td>Rand (syf) (1.e6)</td>
<td>775664</td>
<td>704894</td>
<td>868492</td>
<td>893396</td>
<td>730122</td>
<td>684083</td>
</tr>
<tr>
<td>Frac</td>
<td>0.78</td>
<td>0.70</td>
<td>0.87</td>
<td>0.89</td>
<td>0.73</td>
<td>0.68</td>
</tr>
<tr>
<td>TP total</td>
<td>0.70</td>
<td>0.77</td>
<td>0.58</td>
<td>0.30</td>
<td>0.55</td>
<td>0.64</td>
</tr>
<tr>
<td>TN total</td>
<td>0.76</td>
<td>0.67</td>
<td>0.85</td>
<td>0.88</td>
<td>0.76</td>
<td>0.70</td>
</tr>
<tr>
<td>Specificity total (%)</td>
<td>72.0</td>
<td>74.3</td>
<td>66.8</td>
<td>55.5</td>
<td>62.5</td>
<td>66.1</td>
</tr>
<tr>
<td>Sensitivity total (%)</td>
<td>75.2</td>
<td>69.0</td>
<td>82.0</td>
<td>81.94</td>
<td>72.0</td>
<td>68.9</td>
</tr>
<tr>
<td>CC total</td>
<td>0.47</td>
<td>0.44</td>
<td>0.45</td>
<td>0.21</td>
<td>0.31</td>
<td>0.34</td>
</tr>
<tr>
<td>TP-flanking</td>
<td>0.74</td>
<td>0.80</td>
<td>0.64</td>
<td>0.30</td>
<td>0.53</td>
<td>0.64</td>
</tr>
<tr>
<td>TN-flanking</td>
<td>0.75</td>
<td>0.63</td>
<td>0.84</td>
<td>0.86</td>
<td>0.78</td>
<td>0.72</td>
</tr>
<tr>
<td>spe-flan (%)</td>
<td>74.3</td>
<td>76.0</td>
<td>70.1</td>
<td>55.0</td>
<td>62.5</td>
<td>66.7</td>
</tr>
<tr>
<td>sen-flan (%)</td>
<td>74.8</td>
<td>68.3</td>
<td>80.0</td>
<td>68.2</td>
<td>71.1</td>
<td>69.5</td>
</tr>
<tr>
<td>CC flanking</td>
<td>0.49</td>
<td>0.44</td>
<td>0.49</td>
<td>0.19</td>
<td>0.33</td>
<td>0.36</td>
</tr>
<tr>
<td>TP no flanking</td>
<td>0.68</td>
<td>0.75</td>
<td>0.54</td>
<td>0.30</td>
<td>0.55</td>
<td>0.64</td>
</tr>
<tr>
<td>TN no flanking</td>
<td>0.78</td>
<td>0.70</td>
<td>0.87</td>
<td>0.89</td>
<td>0.73</td>
<td>0.68</td>
</tr>
<tr>
<td>Specificity no flanking (%)</td>
<td>47.6</td>
<td>52.4</td>
<td>40.4</td>
<td>30.3</td>
<td>43.8</td>
<td>47.8</td>
</tr>
<tr>
<td>Sensitivity no flanking (%)</td>
<td>60.4</td>
<td>58.9</td>
<td>63.0</td>
<td>81.8</td>
<td>70.6</td>
<td>66.4</td>
</tr>
<tr>
<td>CC no flanking</td>
<td>0.46</td>
<td>0.46</td>
<td>0.44</td>
<td>0.24</td>
<td>0.29</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Each score was applied to both the training and validation sets. Absolute number and fraction of total set size are presented. Each score was evaluated by its sensitivity, specificity and correlation score (CC). Measures were calculated for the five validation sets: SYFPEITHI without flanking, IEDB without flanking; SYFPEITHI with flanking; general randoms; randoms in SYFPEITHI format (length of 9 AA, no flanking). The lower part of the table represents the relative performance of the scores for either the validation set containing flanking or the validation sets with no flanking regions.

The current scores were learnt using a negative validation set composed of random peptides. We have attempted to learn the cleavage algorithms using peptides from cleavage experiments, not observed to be cleaved. However, the validation results, following such training were not as good as the ones obtained from training with random peptides.

The \[BOW\] score has approximately 15% false positives, not very far from the actual fraction of peptides produced from a random sequence, while maintaining approximately 60–65% properly predicted peptides in the positive validation set (with and without flanking regions—Table 1, fourth column). The opposite was found out to be more complex. Even when increasing the penalty for false negatives, we cannot reduce their fraction to <20%. This can probably be attributed to the
fact that many epitopes in the positive validation sets are not direct cleavage products, and cannot be directly learnt from in vitro cleavage experiments.

When the specific proteasome used is known, specific scores may also be needed. Proteasomes can be divided into core proteasome (20s) without the 19s cap, or with the 19s cap (26s). The constitutive proteasome itself can change some of its active subunits in the presence of IFN-γ to produce the immunoproteasome. IFN-γ can also induce the addition of a PA28 cap. This division is further supported by the presence of hybrid proteasomes (Bose et al., 2001; Brooks et al., 2000; Rivett et al., 2001).

Limited by the size of the data sets, we did not compute a score for each possible type of proteasome (IFN-γ induced, PA28, 26s versus 20s . . .). We computed two generalized scores: one for the immunoproteasome and one for the proteasome produced in the absence of IFN-γ. Our scores can thus be defined as follows:

1. Constitutive proteasome score—[CON]. This score predicts cleavage products of the constitutive proteasome.
2. Immunoproteasome score—[IMM]. The score is to be used when cleavage products are predicted in the presence of IFN-γ.

The absolute and relative performances, sensitivities and specificities of all five scores over both training and validation sets are displayed in Table 1 and Figure 1. These results are significantly better than all previously reported scores (Saxova et al., 2003). The sensitivity of the current algorithm is equal to that of top algorithms like NetChop2 (Saxova et al., 2003), or NetChop3 and has a specificity that is almost twice as large. The values for the functions $S_I - S_2$ of the five different scores are available in the appendix (Supplementary Tables 3–7). Scaled images for the components of [BSC], [CON] and [IMM] are shown in Figure 2.

The two partial scores ([CON] and [IMM]) have lower false positives than their merged counterpart, but much larger false negatives. This can be understood intuitively. The different scores proposed here aggregate different cleavage processes (i.e. cleavage by different proteasomes) under one roof. An aggregation of a larger number of processes allows us to explain more results (thus reducing the false negatives). This comes with the price of a larger number of false positives.

To determine the common prediction capacities of [CON] and [IMM], one could treat them as two parts of the same score function. The combined score [CON][IMM] predicts a peptide to be produced if at least one of its two components is positive. The combined score admits a CC of 0.47, higher than those of the [CON] an [IMM] separately.

As our score functions appear to provide a good prediction of proteasomal cleavage probabilities, we can derive conclusions about biological aspects of cleavage. We here provide a few such examples: In all scores except for the immune proteasome, Cysteine demonstrated a significant negative contribution in practically all positions. The presence of Cysteine either inside a peptide or in its 1st position reciprocal flanking region extremely diminishes its production probability (Fig. 2). To establish this result we counted the appearance of Cysteines in the SYFPETHI database and found only 6 appearances, yielding a frequency of ~0.14% versus a natural frequency of 1.82% (MHC epitopes have 13 times less Cysteines than random epitopes). This phenomenon may be accounted by the bi-sulphide bond that Cysteine establishes. Another clear correlation is between AA volume and their probability of being located at the right hand side of a cleavage site (Fig. 3). These examples highlight the facts that the cleavage scores, although only learnt on the sequences of cleaved peptides, manage to reproduce some of the structural aspects related to cleavage.

4 DISCUSSION
The prediction of proteasomal cleavage products has many applications in immunology. Previous algorithms have tried to specify cleavage sites within a given protein, and define peptides as regions between cleavage sites (Kesmir et al., 2002; Nussbaum, 2001; Peters et al., 2002). Such algorithms cannot predict two overlapping peptides, which often does occur. Moreover, these site specific methods, at the end of the day, fail to provide good predictive capacities for the probability that a given peptide should be produced. We have here presented a score for the probability that a whole sub-peptide is produced by proteasomal cleavage, rather than trying to
locate specific cleavage motifs. Our algorithm assumes a linear contribution of each AA within the peptide and its flanking region to the score. This method yields better performance for the general probability that a given peptide is the result of proteasomal cleavage. Moreover it predicts quite precisely the production of CTL epitope precursors. As in any other prediction method, the current score has both false positives and false negatives (Type I and II errors). We have attempted to balance between the need to minimize both errors types. However, in many situations, one needs to minimize either one or the other. We have thus devised appropriate scores, where either the false positives or the false negatives are minimized. While it was possible to significantly reduce the fraction of false positives, the reduction was less significant for false negatives. This is probably due to inherent errors in the databases used, or to the fact that the presented epitopes may undergo further processing not represented in in vitro cleavage experiments.

We cannot compare directly our results with other state of the art algorithms, since the goal of the different algorithms is slightly different. In contrast with most algorithms, we do not predict cleavage sites, but a score for the entire peptide. If one uses the minimal description of an epitope as a peptide with no overlapping but different sets of peptide fragments from a model protein substrate. This is probably due to inherent errors in the databases used, or to the fact that the presented epitopes may undergo further processing not represented in in vitro cleavage experiments.

The current scores are available at http://peptibase.cs.biu.ac.il/PepCleave/. The results are incorporated into a server presenting all predicted epitopes in a given virus or host. The beta version of this server is open for general access at peptibase.cs.biu.ac.il/peptibase/.

ACKNOWLEDGEMENTS

The work of Y.L, I.G. and T.V. was covered by NIH grant: 1 R01 AI61062-01. The work of T.V. was also covered by a scholarship of Yeshaia Horowitz foundation.

Conflict of Interest: none declared.

REFERENCES


