COMPARISON OF METHODS FOR PREDICTING PROTEASOME CLEAVAGE MOTIFS

1 Nikolaev S.V., 1 Afonnikov D.A., 1,2 Ivanisenko V.A., 2 Bazhan S.I., 1 Kolchanov N.A.

1 Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia, e-mail: nikolaev@bionet.nsc.ru
2 State Research Center of Virology and Biotechnology “VECTOR”, Koltsovo, Novosibirsk region, Russia

Key words: proteasome, proteolysis, proteasome cleavage motifs, antigenic epitope, MHC1, algorithm, prediction

Resume

Motivation: Proteasome proteolysis is one of the major processes of degradation of proteins with structural defects or antigens. Methods for predicting proteasome cleavage motifs are of great importance for medicine and biotechnology. At present, there are several algorithms of prediction of such motifs. In the present paper, we have chosen three of them for comparison.

The algorithm by Kesmir et al. (2002) is based on experimental data on proteasome cleavage in vitro. The other two algorithms (Altuvia, Margalit, 2000; Kesmir et al., 2002) are based on data on peptides binding to the major histocompatibility complex (MHC) class 1. In this study, we compare the predictions of proteasome cleavage motifs performed with these three algorithms.

Results: We have shown that the results of the test can be interpreted as follows: two algorithms (Kesmir et al., 2002) detect complementary features of proteasome cleavage motifs. Combination of their predictions has improved the recognition of the motifs in the test sample of fragments.

Introduction

Degradation of proteins in the cell is vital for both the regulation (e.g., rapid digestion of transcription factors) and elimination of defective proteins and antigens (Ciechanover, 1998). Proteasomal cleavage is one of the main processes of such degradation. It is performed by a protein complex, proteasome. The core of the proteasome is a multi-enzymatic complex, whose subunits are endoproteases integrated into a cylindrical aggregate consisting of four rings. Each of the rings consists of seven subunits. This allows cleavage of an enzyme next to any amino acid residue but in a certain context (Orlowski et al., 2000). It is believed that peptides resulting from this cleavage have the same C terminus as the MHC1 antigenic epitope (Altuvia, Margalit, 2000). The N terminus of the peptide can be additionally modified by aminopeptidases in the cytosol and/or cytoplasmic reticular lumen (Altuvia, Margalit, 2000).

Recent experimental (Nussbaum et al., 1998) and theoretical (Holzhutter et al., 1999; Altuvia, Margalit, 2000) studies of proteasome cleavage have brought about a number of methods for prediction of cleavage motifs. The present paper is dedicated to testing the three algorithms providing quantitative estimates of cleavage site scores.

Methods and Algorithms

Prediction algorithms


Algorithm 3 (AM). The Cleavage Scores Table calculated with the use of data on MHC1-binding peptides (Altuvia, Margalit, 2000), is available by the address http://bioinfo.md.huji.ac.il/marg/cleavage/scores.html. We used this table for predicting cleavage motifs as follows. A sequence was scanned with a window with a length equaling two amino acid residues (A, B). Cleavage site scores were selected from the Cleavage Scores Table at the cross of row A and column B and were assigned to the first residue of the window. The scores picked from the Cleavage Scores Table were normalized to the range [0, 1]. The algorithm is implemented in the form of a script in the system Matlab 5.2.0.

Comparison of Algorithms

The algorithms for predicting cleavage motifs were compared using a sample of protein fragments with the proteasome cleavage motifs determined in vitro (Holzhutter et al., 1999). The sequences of these fragments are shown in Fig. 1.
1) insulin B chain 
FvNQHLCgSHLVEALYLVCGERGFFYTPKa
2) viral peptide HBVcAg
AyrppNAPILSTpeTTVVRrGRSPrrTPs
3) viral peptide pp89 
rlMYDMYphfMptnLgpsEKrVwMs
4) OvaY 51-71
YqlINKVVRFDkLPgFGDsEa
5) OvaY 249-269 
YVsgLEqLEsINFEkLtEwts
6) Ova 239-281 
msMLvLPdeVSglEgLLEiFkLtEEwItSSnVMeeRKIkvyl
7) p53wt
TleDssgnLLgRnsFeVrVCapgrdr

Fig. 1. Sequences used for testing the algorithms for predicting cleavage motifs. Capitalized are amino acid residues next to which the proteasome proteolysis was experimentally observed in vitro (Holzhutter et al., 1999).

For all the tested sequences, proteolysis motif scores were either obtained from www servers or calculated (for algorithm 3). Then, all the sequences and their experimental and predicted cleavage motifs were pooled into one sample. Probabilities of type I and II errors were calculated in this sample for various threshold levels. The dependence curves for type I and II (false positive and false negative) errors are shown in Fig. 2.

![Graph showing false positive vs false negative errors for three prediction methods: solid line, NetChop C-term 2.0; dotted line, Altuvia and Margalit; and dot-dash line, NetChop 20 S.](image)

We verified the agreement between the predictions of cleavage motifs made by these algorithms for various positions of the test sequences. For this purpose, we constructed scattering diagrams of the predicted scores for each algorithm pair (Fig. 3).

Probabilities of type I and II errors were calculated for combined predictions. Predictions made with the threshold of 0.5 were taken into account. The following combinations were considered: (1) disjunction of predictions, (2) conjunctions of predictions, (3) disjunction of conjunctions, and (4) “voting” method (Table).
Table. Probabilities of errors of types I and II for combined predictions.

<table>
<thead>
<tr>
<th>Combination</th>
<th>Type I error</th>
<th>Type II error</th>
<th>Sum of type I and II errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. NCC2</td>
<td>0.59</td>
<td>0.24</td>
<td>0.83</td>
</tr>
<tr>
<td>2. NC20S</td>
<td>0.62</td>
<td>0.21</td>
<td>0.83</td>
</tr>
<tr>
<td>3. AM</td>
<td>0.22</td>
<td>0.72</td>
<td>0.94</td>
</tr>
<tr>
<td>4. NCC2 ∪ NC20S</td>
<td>0.41</td>
<td>0.36</td>
<td>0.77</td>
</tr>
<tr>
<td>5. NCC2 ∩ NC20S</td>
<td>0.6</td>
<td>0.24</td>
<td>0.84</td>
</tr>
<tr>
<td>6. NC20S ∩ AM</td>
<td>0.65</td>
<td>0.24</td>
<td>0.89</td>
</tr>
<tr>
<td>7. (NCC2 ∩ NC20S) ∪ (NCC2 ∩ AM)</td>
<td>0.6</td>
<td>0.24</td>
<td>0.84</td>
</tr>
<tr>
<td>8. (NC20S ∩ AM) ∪ (NCC2 ∩ AM)</td>
<td>0.65</td>
<td>0.17</td>
<td>0.82</td>
</tr>
<tr>
<td>9. (NC20S ∩ AM) ∪ (NC20S ∩ AM)</td>
<td>0.45</td>
<td>0.31</td>
<td>0.76</td>
</tr>
<tr>
<td>10. Voting method</td>
<td>0.45</td>
<td>0.31</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Results and Discussion

The plots of cleavage motif prediction errors (Fig. 2) show that the algorithms tested have approximately equal ratios between the errors of types I and II. Note that in the range of type I errors from 0.2 to 0.4, the least errors of type II occur in the case of the algorithm NC20S, trained on experimental data.

Figures 3a–c show that the cleavage motif scores predicted by different algorithms are in poor agreement. This may result from the fact that one algorithm was trained only on experimental data on in vitro proteasome proteolysis, and the two other algorithms, on the C-ends of antigenic epitopes. The last two algorithms show a certain agreement in predictions: dots tend to occur in the left upper corner of Fig. 3c. In addition, the algorithm AM yields overprediction in comparison with NCC2, probably, because the former uses positions P1 and P2 for evaluation of the scores, and the latter uses a wider window around the peptide bond examined.

Probabilities of prediction combination errors $A_i$ made by the tested algorithms are shown in Table. The following combinations were considered: (1) disjunction of prediction pairs $(A_i ∪ A_j)$, (2) conjunction of prediction pairs $(A_i ∩ A_j)$, (3) disjunction of conjunction pairs $(A_i ∪ A_j) ∪ (A_i ∩ A_j)$, and (4) voting method $(A_i ∩ A_j) ∪ (A_i ∩ A_j) ∪ (A_i ∩ A_j)$. The
probabilities of correct prediction for test fragments increase when predictions by the methods NCC2 and NC20S are joined (row 4 vs. rows 1 and 2, Table). This may indicate that the training samples for the neural networks NCC2 and NC20S have complementary features, and their weight matrices also code for complementary information. This is confirmed by the low probability of the conjunction of predictions by these networks (row 5, Table). In addition, Fig. 2 shows that mere threshold change cannot yield this ratio between errors of types I and II by none of these methods.

Conjunction of predictions made by AM with those made by NC20S and NCC2 (rows 6 and 7, respectively) somewhat reduces the prediction level but less than with algorithms 1 and 2. Disjunction of these conjunctions also increases the prediction probability (row 10, Table) and is as good and the combined prediction by the “voting” method (row 11, Table).

For all methods, the threshold level of 0.5 is taken. Disjunction of predictions by the methods NetChop C-term 2 and NetChop 20 S increases the probability of correct prediction (row 4).

Thus, our comparison has demonstrated that the available methods for predicting proteasome cleavage motifs are sensitive to the training sample. Disjunction of the predictions made by the neural networks NetChop C-term 2 and NetChop 20 S improves the motif recognition in the test sample of fragments.

Further prediction improvement will depend on both developing the models underlying the prediction methods and obtaining new experimental data. Note that methods for comparison of algorithms for predicting proteasome cleavage motifs with a limited volume of experimental information should also be developed.

Acknowledgements

The study was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376 and 01-07-90084); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); US Department of Energy (grant № 535228 CFDA 81.049); and CRDF (grant № RB0-1276).

References