Chapter 14

Predicting Immunogenicity: An Integrative Approach

The genome era provides opportunities to study the immune system from a systems biology perspective as discussed in chapter 1. We now have not only the sequence information that sheds light on the immunological diversity among individuals in a population but also advanced techniques that allow us to obtain a better estimate of the kinetics and specificity of an immune response. In this chapter we will give an example of such systems biology approaches to immunology: prediction of immunogenic regions for cytotoxic T cells. A very similar study is published by Larsen et al. [2005].

Reliable prediction of immunogenic peptides may be useful for many applications, e.g., for rational vaccine design. Many attempts have been made to predict the outcome of the steps involved in antigen presentation. As we have described earlier in the book, a number of methods have been developed that very reliably predict the binding affinity of peptides to the different MHC-I alleles [Brusic et al., 1994, Buus et al., 2003, Nielsen et al., 2003, 2004]. Likewise, a method has been developed that predicts the efficiency by which peptides of arbitrary length can be transported by TAP [Peters et al., 2003a]. Several methods have also been developed that aim at predicting the proteasomal cleavage pattern of proteins (see chapter 7 for details).

Can predictions of proteasomal cleavage patterns and TAP transport efficiency contribute to an improved identification of epitopes compared to that obtained when using only predictions of MHC-I affinity? Peters et al. [2003a] have shown that combining MHC-I affinity predictions with prediction of TAP transport efficiency leads to improved identification of CTL epitopes. This analysis can be extended to address, for a large set of different HLA alleles,

if a combined prediction method mimicking the MHC I pathway can improve prediction of epitopes. The following analysis includes epitopes from close to 70 different MHC alleles from different MHC-I supertypes [Sette and Sidney, 1999, Lund et al., 2004]. The proteasomal cleavage event were modeled by prediction algorithms as described in chapter 7.

To validate the integrative method, a data set (SYF) containing 152 9mer epitopes restricted to more than 70 different HLA alleles extracted from the SYFPEITHI database (http://syfpeithi.bmi-heidelberg.com/) are uesd. The majority of these peptides have successfully passed the steps involved in antigen presentation. The set of negative peptides (peptides that will not be presented by the MHC class I pathway) were defined as all 9mer peptides contained in the protein sequences from which the epitopes originated, except those annotated as epitopes in either the complete SYFPEITHI or Los Alamos HIV databases (www.hiv.lanl.gov/immunology. When using this definition of epitopes/nonepitopes one has to take into account that some 9mers will falsely be classified as nonepitopes because the SYFPEITHI and Los Alamos HIV databases are incomplete. Since the HLA molecules have a very specific peptide binding repertoire, this false-negative proportion will be very small. In a protein of 200 amino acids, one expects to have one binding, and approximately 199 nonbinding peptides [Yewdell and Bennink, 1999]. The potential number of false negatives is hence orders of magnitude smaller than the actual number of negatives.

14.1 Combination of MHC and Proteasome Predictions

To examine whether predictions of proteasomal cleavage can contribute to the classification of peptides into epitopes/nonepitopes independently of the predicted MHC-I binding affinity, one option is to perform a sort/split experiment: two groups of peptides with approximately equal predicted MHC-I affinity, but different predicted proteasomal cleavage, is generated. All 9mer peptides in each protein is individually sorted according to their predicted MHC-I affinity. Looking at two peptides at a time from the top of the sorted list, they are then split into two groups and the peptide with highest predicted proteasomal cleavage value is put in group H, whereas the peptide with the lowest is put in group L. Figure 14.1 shows, for four different methods predicting proteasomal cleavage, how the number of epitopes in the H group deviates from the expected number (50%).

To test if the number of epitopes is significantly different in group H as compared to group L, the binomial distribution is applied. Under the null hypothesis, the epitopes have an equal chance of falling into either group, $\pi_0 = 0.5$. If n is the total number of epitopes, the expected number of epitopes

in either group is $\pi_0\, n$. If r is the observed number of epitopes in one of the groups, the departure from the expected number can be expressed by the z-score [Armitage et al., 2004]:

$$z = \frac{r - n\pi_0}{\sqrt{n\pi(1-\pi)}}. \tag{14.1}$$

The nullhypothesis is rejected at $p = .05$ if $z > 1.96$, at $p = .01$ if $z > 2.58$, and at $p = .001$ if $z > 3.29$.

All four proteasomal cleavage methods the number of epitopes is significantly higher in group H than in group L. The method with the poorest performance is that of NetChop 20S with a p-value just below .01. The other three methods all separate the H from the L group with p-values below or close to .001. For NetChop 2.0, for example, 34% or 72% more epitopes are found in the H group. Figure 14.1 also shows that the predicted cleavage patterns of the internal amino acids add very little extra information to the predicted MHC-I affinity. When using NetChop 2.0 or NetChop 3.0 to study the predicted cleavage at position 1, only 38% and 39%, respectively, of the epitopes are located in group H. This may indicate that peptides with a high predicted proteasomal cleavage value at this position are rarely epitopes. If, however, the NetChop 20S or NetChop 20S-3.0 network is used, this scenario is reversed.

Applying the bootstrap [Press et al., 1992] method you find that the NetChop 20S method performs significantly worse than the other methods ($p < .05$ in all three comparisons). The difference in predictive performance between the other methods is, however, statistically insignificant ($p > .05$ in all cases). Thus, this analysis demonstrates that only the methods based on *in vivo* cleavage data can improve the identification of epitopes in combination with the predicted MHC-I affinity.

14.2 Independent Contributions from TAP and Proteasome Predictions

To address the question of whether proteasomal cleavage and TAP transport efficiency can contribute independently to the identification of epitopes a sort/split experiment sorting on TAP transport efficiency and splitting on proteasomal cleavage was conducted. When examining if cleavage predictions can contribute to the identification of epitopes independently of the predicted TAP transport efficiency, two groups of peptides with close to equal TAP transport efficiency, but different predicted proteasomal cleavage, were generated using the same method as described in the previous section. In this experiment the two groups H and L thus have similar TAP transport efficiency, but very different predicted proteasomal cleavage values. The result of the analysis is

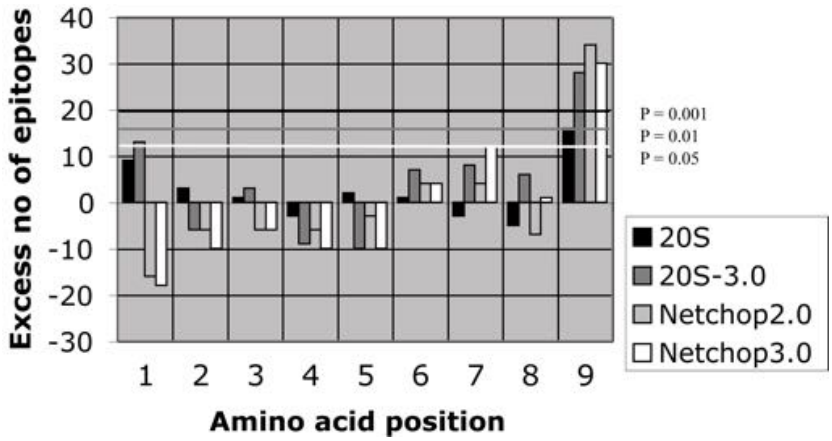

Figure 14.1: Sort/split experiment conducted sorting on predicted MHC-I affinity, splitting on predicted proteasomal cleavage. Two groups with close to equal MHC-I affinity, but with different predicted proteasomal cleavage. In total, the two groups contain 152 epitopes. The figure shows the number of epitopes in group H deviating from the expected number of 76 (50%) L. 1-9: position 1-9 of the peptide (9 is the C-terminal end). Four different methods have been used for predicting proteasomal cleavage: NetChop 20S, NetChop 20S-3.0, NetChop2.0, and NetChop3.0. Also shown are lines indicating levels of significance estimated as described in the text.

shown in figure 14.2, where NetChop 3.0 has been used for the proteasomal cleavage predictions. The figure shows how the number of epitopes in the H group deviates from the expected number (50%). In combination with TAP transport efficiency only, the predicted C-terminal cleavage can contribute significantly to the identification of the epitopes. There is an excess number of 30 epitopes between the H and L groups, corresponding to 70%. This result demonstrates that not all TAP transported peptides are cleaved equally well by the proteasome. Between two groups of peptides with equal TAP transport efficiencies, epitopes are found predominantly in the group with high proteasomal C-terminal cleavage.

Next a sort/split experiment sorting on MHC-I affinity and splitting on TAP transport efficiency is conducted to investigate if TAP transport efficiency and MHC-I binding can contribute independently to the identification of epitopes. In the experiment, most epitopes (66%, $p < .001$) fall into the group with high TAP transport efficiency. Among peptides with similar MHC-I affinity, peptides with high TAP transport efficiency are thus most likely to be epitopes.

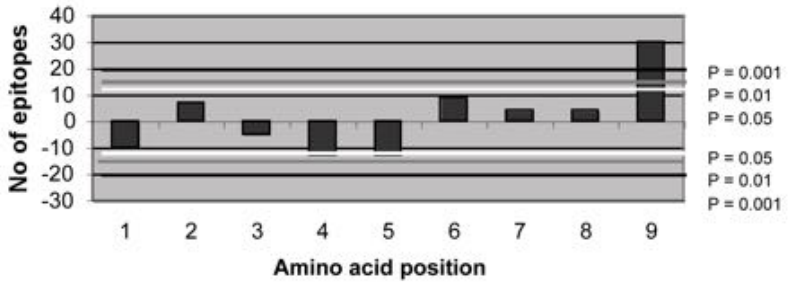

Figure 14.2: Sort/split experiment conducted sorting on predicted TAP transport efficiency, splitting on predicted proteasomal cleavage. Proteasomal cleavage is predicted using the method of NetChop 3.0. Two groups with close to equal predicted TAP transport efficiency, but with different predicted proteasomal cleavage. In total, the two groups contain 152 epitopes. The figure shows the number of epitopes in group H deviating from the expected number of 76 (50%). 1-9: position 1-9 of the peptide (9 is the C-terminal end). Also shown are lines indicating levels of significance.

14.3 Combinations of MHC, TAP, and Proteasome Predictions

A combined prediction score for MHC-I affinity, proteasomal C-terminal cleavage, and TAP transport efficiency can be defined as a weighted sum of the three individual prediction scores. We use an MHC-I affinity rescaled prediction values; TAP prediction method of [Peters et al., 2003a], and the NetChop 2.0 and 3.0 predictors described in chapter 7.

Two nonparametric performance measures are used to evaluate the performance of the combined methods. One measure is the conventional A_{ROC} value (the area under the receiver operator characteristics [ROC] curve) [Swets, 1988]. In this measure, all overlapping 9mer peptides in the SYF data set were sorted according to the prediction score. The epitopes define the positive set, whereas the negative set is made from all other 9mers, excluding 9mers present in the SYFPEITHI or the Los Alamos databases. In a typical calculation, the positive set contains 152 peptides, and the negative set more than 92,000 peptides.

The ROC curve is plotted from the sensitivity and 1-specificity values calculated by varying the cut-off value (separating the predicted positive from the predicted negative) from high to low. The A_{ROC} value is 0.5 for a random prediction method and 1.0 for a perfect method. Even though commonly used,

the A_{ROC} measure is not easy to interpret intuitively. A second performance measure with a clear and intuitive interpretation is a rank measure: for each protein in the benchmark, all 9mer peptides are sorted based on the prediction score. A given protein may appear more than once in the benchmark if it contains more than one epitope. The rank value for the protein is calculated as the number of nonepitopes with a score higher than that of the corresponding epitope. From these rank values a rank curve showing the accumulative fraction of proteins with a rank value below a certain value was constructed. From the rank curve one can then extract information on how large a fraction of the proteins will have the epitope within a rank of, e.g., 25. Finally, a single performance measure (A_{RANK}) as the area under the rank curve integrated from rank zero up to rank 100 was defined. A perfect prediction method will have all the epitopes as rank 1, and thus an A_{RANK} value of 1.0, whereas a poor method will have the epitopes well below rank 100 and hence an A_{RANK} value of 0.0. Examples of a ROC and a rank curve are shown in figure 14.3. For both the A_{ROC} and A_{RANK} performance measures, one should be aware that some 9mers will falsely be classified as nonepitopes because the SYFPEITHI and Los Alamos HIV databases are incomplete.

The SYF data set is used to estimate the set of weights where the A_{RANK} and A_{ROC} values are optimal. Next the optimal combined prediction scheme is applied to an HIV data set of 69 epitopes derived from the Los Alamos HIV database to estimate the performance gain on an independent evaluation data set.

The optimal combined method is found to have relative weights on C-terminal cleavage and TAP transport efficiency of 0.15 and 0.115, respectively. In figure 14.3, we show examples of ROC and rank curves for the SYF data set. The figure shows the performance curves for five different prediction scoring schemes: Comb, MHC, TAP, NetChop 2.0, and NetChop 3.0. Here, the Comb method is the combined method with relative weight on TAP and NetChop 3.0 of 0.115 and 0.15, respectively, while others are single predictions. In figure 14.4, we give the details of the performance measures for the different methods and their combinations.

The curves shown in figure 14.3 clearly highlight the problematic aspects of using the A_{ROC} performance measure when dealing with highly unbalanced data sets. The A_{ROC} values for the NetChop 3.0 and TAP prediction methods are close to identical (see figure 14.4). However, looking at the ROC curves for each method, it is clear that the NetChop 3.0 method provides the most useful predictions. The region of the ROC curve where the TAP predictor performs best falls in a highly nonrelevant region of the specificity. The two curves cross at a false-positive ratio of 0.4. This value corresponds to 40% false-positive predictions, and having an improved prediction method only in this specificity range is clearly irrelevant. For the rank curves this problem is not present, and

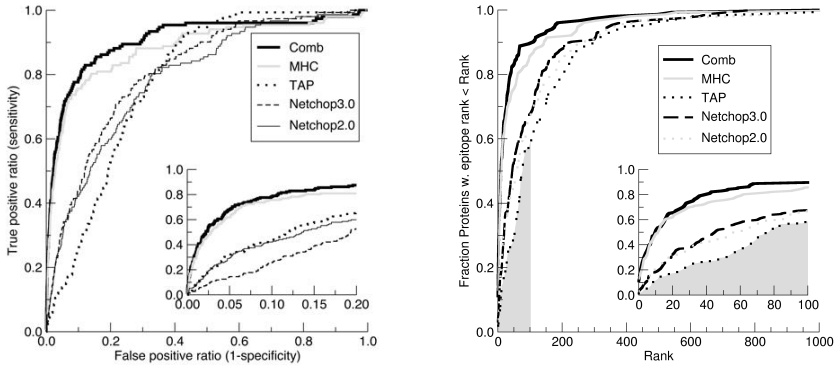

Figure 14.3: ROC and rank performance curves for different prediction methods. Left: the ROC curves. Right: rank curves. A_{RANK} is the area under the rank curve (highlighted as the shaded area under the TAP curve) as described in the text. Predictions are made on the SYF data set. The different prediction methods are; Comb: optimal combined method with relative weight on C-terminal cleavage and TAP transport efficiency of 0.15 and 0.115, respectively; MHC: MHC-I affinity; TAP: TAP transport efficiency; NetChop 3.0: C-terminal cleavage by NetChop; NetChop 2.0: C-terminal cleavage by NetChop 2.0. The inserts to the figures show high specificity/high rank, part of the corresponding curves.

we can directly identify the most relevant method from the integrated A_{RANK} value.

The results shown in figures 14.3 and 14.4 demonstrate that the combined method integrating prediction of proteasomal cleavage, TAP transport, and MHC affinity has the highest performance in terms of both the A_{ROC} and A_{RANK} values. The individual method with the poorest performance is that of NetChop 20S, followed by NetChop 20S-3.0, TAP, the NetChop 2.0 and NetChop 3.0 methods, and MHC-I affinity.

What is also clear from the results shown is that the combined method has a predictive performance superior to that of both MHC-I affinity alone and any method integrating prediction of MHC-I affinity with TAP transport efficiency or C-terminal proteasomal cleavage. The performance values for MHC, MHC+TAP, MHC+NetChop 3.0, and the combined method are 0.88, 0.90, 0.90, 0.91 for A_{ROC} and 0.70, 0.75, 0.73, 0.76 for A_{RANK} respectively. Comparing the performance values for the combined method to that of MHC-I, MHC-I+TAP, and MHC-I+NetChop 3.0, we find the following bootstrap hypothesis test values: <0.01, <0.01, <0.01 and 0.025, <0.01, <0.01 for A_{ROC} and

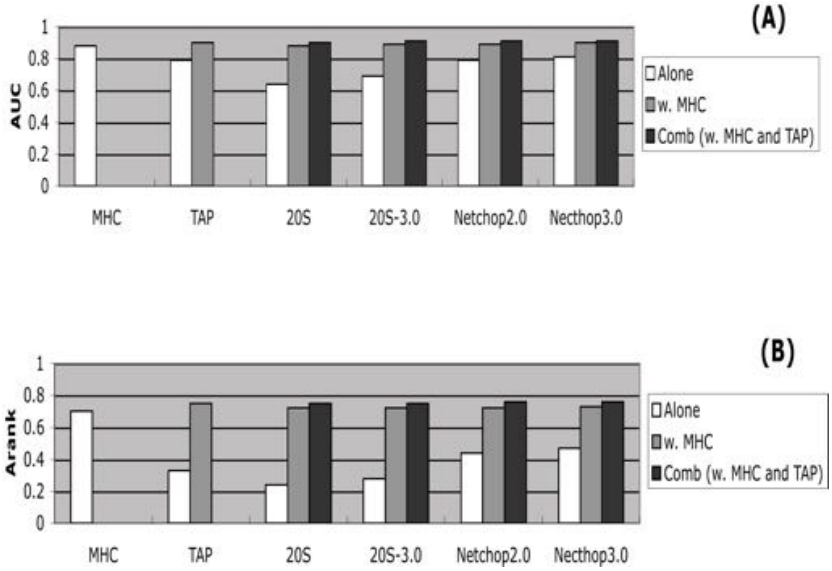

Figure 14.4: Predictive performance for different prediction methods: A_{ROC} (upper panel) and A_{RANK} (lower panel). Predictions are made on the SYF data set. The figure shows for each prediction method the performance measures for each method on its own, the optimal performance in combination with MHC affinity predictions, and the optimal performance in combination with TAP transport efficiency and MHC affinity predictions.

A_{RANK}, respectively. However, we see no significant difference between the combined methods integrating predictions from any of the three proteasomal cleavage prediction methods: NetChop 20S-3.0, NetChop 2.0, or NetChop 3.0. This analysis indeed shows that the combined method performs significantly better than all other methods in the comparison.

It is striking to observe that in combination with MHC-I affinity the TAP predictor provides more additional information useful for epitope identification than any of the NetChop predictors. The MHC-I+TAP predictor has A_{RANK} and A_{ROC} values of 0.75 and 0.90, respectively, whereas the values for MHC-I+NetChop 3.0 are 0.73 and 0.89. Using the bootstrap experiment, we find that these values are significantly different ($p < .05$).

Another interesting finding is that even though the different NetChop predictors, except NetChop 20S, individually have very different predictive performance, they achieve the same predictive performance when combined with MHC-I affinity predictions. In combination with MHC-I affinity predictions, NetChop 20S-3.0, NetChop 2.0, and NetChop 3.0 all have performance val-

ues close to 0.90 and 0.73 for A_{ROC} and A_{RANK}, respectively, and the individual performance differences are statistically significant. Finally, we also found that the NetChop 20S-3.0 and TAP predictors can be combined in a constructive manner with a predictive performance significantly higher than that of the individual predictors. This is, however, not the case for the NetChop 3.0 predictor. Here the combination with TAP only leads to a minor and insignificant improvement in the predictive performance (data not shown). This analysis suggests that the NetChop predictor trained on epitope data does indeed predict a combination of MHC-I affinity, TAP transport efficiency, and proteasomal cleavage rather than just proteasomal cleavage. As an individual prediction method for epitope recognition, the NetChop method trained on epitope data clearly outperforms the methods trained on in vitro degradation data. However, when combined with MHC-I affinity and TAP transport efficiency predictions both the epitope and in vitro trained methods achieve similar performance.

A direct measure of the performance gain when comparing the combined method to that of MHC-I affinity prediction alone is the rank number needed in order to identify 75% of the epitopes in the benchmark. For the MHC-I affinity predictions alone this rank number is 55, meaning that in a set of proteins one will have to test 55 peptides from each protein in order to identify 75% of the epitopes. For the combined method this number has dropped to 30. Even though this number is still high, the performance gain is clearly notable.

14.4 Validation on HIV Data Set

The above analysis can be done for indivual pathogens, like HIV. The results of such an analysis are shown in figure 14.5 and confirm the findings from the SYF data set. The combined method has a performance superior to that of all the individual methods. The TAP transport predictor has the poorest performance, followed by that of NetChop 3.0. Estimating the rank number needed in order to identify 75% of the epitopes in the benchmark, we find values of 52 and 30 for the MHC-I predictor alone and the combined method, respectively. These numbers thus confirm the values found when using the SYF data set. A direct implication of this performance gain is a twofold reduction in the experimental efforts needed to identify 75% of the epitopes in a large set of proteins.

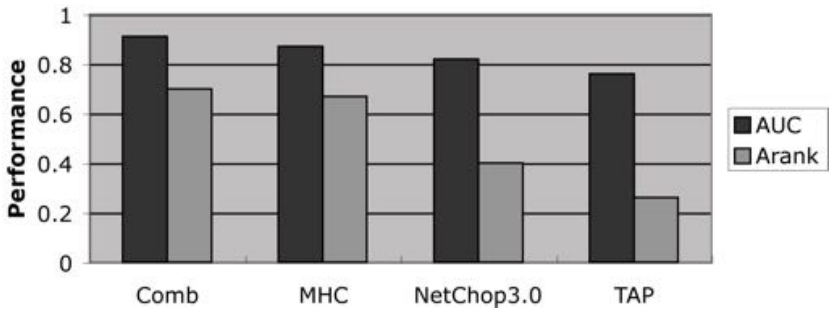

Figure 14.5: Performance for different prediction methods. Predictions are made on the HIV data set. The figure shows the predictive performance for the three individual prediction methods of MHC, C-terminal cleavage (NetChop 3.0), and TAP, as well as the combined method (Comb) with relative weight on C-terminal cleavage and TAP transport efficiency of 0.15 and 0.115, respectively.

14.5 Perspectives on Data Integration

In this chapter, we have demonstrated how an integrative approach combining predictions of the proteasomal cleavage, TAP transport efficiency, and MHC-I affinity can lead to improved CTL epitope recognition.

Other groups have previously combined different prediction methods: Hakenberg et al. [2003] developed a bioinformatical tool for prediction of CTL epitopes by combining prediction of proteasomal cleavage and MHC affinity. On a very small data set of only five epitopes from HIV-1 Nef, Kesmir et al. [2002] showed that combining predictions of proteasomal cleavage with measured TAP and MHC-I binding affinity correlates well with the observed number of MHC-I ligands presented on the cell. In another study, Peters et al. [2003a] improved identification of epitopes by combining predictions of binding affinities to the HLA-A*0201 allele with predictions of TAP transport efficiency. They also combined HLA-A*0201 affinity predictions with predictions of C-terminal cleavages by NetChop 20S, but this led to a less accurate identification of epitopes. What is novel about the analysis given in this chapter is the broad set of MHC-I (70 different alleles) specificities used. This allows us to (1) draw more general and well-founded conclusions about how to integrate the different steps in the class I pathway in an optimal manner, and (2) derive a prediction method that is broadly applicable to the identification of CTL epitopes.

Concern has previously been raised that the NetChop methods, which have been trained on natural MHC-I ligand data, do not only predict proteasomal cleavage but rather a combination of cleavage, TAP transport, and affinity to

the average MHC-I allele [Peters et al., 2003a]. We find that when predicting CTL epitopes, the NetChop method trained on epitope data outperforms the methods trained on in vitro degradation data. However, in combination with MHC-I affinity and TAP transport efficiency predictions, both methods trained on *in vitro* digest data and MHC ligands, respectively, show similar performance. This leads to the conclusion that the high performance of the NetChop method trained on epitope data does not come from more accurate prediction of the proteasomal cleavage but rather from indirect integration of TAP transport efficiency and MHC-I affinity. However, this observation also leaves promise for future improvements to CTL epitope predictions, since it should be possible to improve at least the proteasomal cleavage prediction accuracy by developing a method describing the differences between the immuno proteasome and the constitutive proteasome cleavage specificities, and thereby improve the accuracy of the integrative method.