Selecting Informative Data for Developing Peptide-MHC Binding Predictors Using a Query by Committee Approach

Jens Kaae Christensen
jenskc@cbs.dtu.dk
Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, DK-2800 Lyngby, Denmark

Kasper Lamberth
k.lamberth@immi.ku.dk
Department of Experimental Immunology, Institute of Medical Microbiology and Immunology, University of Copenhagen, Dk-2200 Copenhagen N, Denmark

Morten Nielsen
mniel@cbs.dtu.dk
Claus Lundegaard
lunde@cbs.dtu.dk
Peder Worning
peder@cbs.dtu.dk
Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, DK-2800 Lyngby, Denmark

Sanne Lise Lauemøller
lauemoeller@get2net.dk
Søren Buus
sb@immi.ku.dk
Department of Experimental Immunology, Institute of Medical Microbiology and Immunology, University of Copenhagen, Dk-2200 Copenhagen N, Denmark

Søren Brunak
brunak@cbs.dtu.dk
Ole Lund
lund@cbs.dtu.dk
Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, DK-2800 Lyngby, Denmark

Strategies for selecting informative data points for training prediction algorithms are important, particularly when data points are difficult and costly to obtain. A Query by Committee (QBC) training strategy for selecting new data points uses the disagreement between a committee of different algorithms to suggest new data points, which most rationally...
Complement existing data, that is, they are the most informative data points. In order to evaluate this QBC approach on a real-world problem, we compared strategies for selecting new data points. We trained neural network algorithms to obtain methods to predict the binding affinity of peptides binding to the MHC class I molecule, HLA-A2. We show that the QBC strategy leads to a higher performance than a baseline strategy where new data points are selected at random from a pool of available data. Most peptides bind HLA-A2 with a low affinity, and as expected using a strategy of selecting peptides that are predicted to have high binding affinities also lead to more accurate predictors than the baseline strategy. The QBC value is shown to correlate with the measured binding affinity. This demonstrates that the different predictors can easily learn if a peptide will fail to bind, but often conflict in predicting if a peptide binds. Using a carefully constructed computational setup, we demonstrate that selecting peptides with a high QBC performs better than low QBC peptides independently from binding affinity. When predictors are trained on a very limited set of data they cannot be expected to disagree in a meaningful way and we find a data limit below which the QBC strategy fails. Finally, it should be noted that data selection strategies similar to those used here might be of use in other settings in which generation of more data is a costly process.

1 Introduction

The ongoing genome projects have provided us with the entire genomes of many of the organisms that are harmful to humans. This information may be used in rapid design of vaccines against emerging threats. The immune system does not react to entire protein sequences but rather to small peptide parts of these called epitopes. The major histocompatibility class I (MHC I) molecules sample peptides from proteins produced within a cell and transport them to the cell surface. When cytotoxic T cells from the immune system discover a peptide from a foreign organism bound to an MHC I molecule on the surface of a cell, it is activated to kill the cell, since it is likely that the cell is infected (Rock & Goldberg, 1999). The binding of peptides to MHC is the most selective step in the cellular process of preparing epitopes from proteins (Yewdell & Bennink, 1999). A number of computational methods have been developed to search for epitopes (reviewed by Schirle, Weinschenk, & Stevanovic, 2001). Together with binding assays, these methods can be used to rapidly discover epitopes that subsequently can be used in vaccine development. The simpler prediction tools rely on weight matrices, and all assume that the amino acids at each position along the peptide sequence contribute in an independent manner to the binding energy of the peptide. (Parker, Bednarek, & Coligan, 1994; Meister, Roberts, Berzofsky, & De Groot, 1995; Stryhn et al., 1996). Similar types of approaches are used by the EpiMatrix method (Schafer, Jesdale, George, Kouttab, & De Groot, 1998),
Selecting Informative Data

the BIMAS method (Parker et al., 1994) and the SYFPETHI method (Rammensee, Bachmann, Emmerich, Bachor, & Stevanovic, 1999). In contrast to this, artificial neural networks (ANN) can take correlations between different positions in the sequence into account, and neural network methods for predicting peptide-MHC binding have already been developed (Brusic, Rudy, & Harrison, 1994; Buus et al., 2003; Nielsen et al., 2003).

In a population, a given HLA molecule is found in many different sequence forms (HLA is said to be polymorphic), and depending on the sequence, the HLA molecule will bind a highly specific repertoire of peptide sequences. This extreme polymorphism of the HLA molecule and the diverse specificities of different HLA molecules are complicating factors for vaccine development. Different persons are likely to carry different HLA molecules (called alleles), and their immune system will therefore generally recognize different sets of peptides from a given organism. In order for a vaccine to have a broad coverage, it must therefore contain peptides that are recognized by many different HLA molecules. The quest to define the specificity of all HLA molecules has been coined the human MHC project (Lauemøller et al., 2000).

Sette and Sidney (1999) have defined nine supertypes that cover most of the HLA-A and -B polymorphism. Different alleles within these supertypes are believed to have similar binding specificity. To span the MHC polymorphic space, one should train predictors for alleles representing each supertype. Buus et al. (2003) have shown that neural networks trained on quantitative data have a performance superior to that of networks trained on classification data. To derive reliable predictors, one should therefore obtain suitable sets of peptides with associated binding affinity measurements for the different MHC molecules in order to train neural networks. Identifying MHC binding peptides is a slow and costly process. Since only approximately 1 in 200 peptides can bind to a given MHC (Yewdell & Bennink, 1999), many peptides, if randomly chosen, have to be tested to find a few that bind. A more efficient way of discovering new epitopes is only to test only those that have been predicted to be good epitopes. To follow this approach, reliable and accurate predictors are a prerequisite. An optimal method to develop accurate predictors involves a guided selection of peptides that subsequently have their binding affinity experimentally determined. Hereafter, the predictors are retrained using these new data. In order to minimize the number of experiments that must be performed, it is important to be able to select the data points that will increase the performance of the predictors the most. Here we compare different strategies for selecting such points in order to establish the validity of different data selection schemes.

Seung, Opper, and Sompolinsky (1992) proposed the Query by Committee (QBC) strategy for selecting new data points (queries). They argue that the Shannon information of a query can be used to guide the selection of new data points and show that this value can be estimated by the disagree-
ment among different predictors. According to the strategy, a committee of predictors is trained on the available data set, and the next query is chosen according to the principle of maximal disagreement. They show that this leads to an exponentially decreasing generalization error rather than the decrease following an inverse power law that is obtained using randomly chosen inputs. Seung et al. (1992) suggest choosing new data points so that half of an ensemble of predictors classify them as positive examples and the other half as negative examples. Here we deal with a problem in which the output is a real valued number. Instead of making a threshold defining positive and negative examples, we modify the definition of disagreement to be the standard deviation of the outputs of the predictors. Ensembles of networks were generated by training on different subsets of the data, and we sought to avoid overfitting by stopping the training when the performance was optimal on the remaining data. An alternative method would be to use Bayesian learning (Neal, 1994). We previously proposed using QBC for selecting data points with maximal complementary information content relative to data already available (Buus et al., 2003). Seung et al. (1992) showed that QBC is effective for two theoretical models; however, they did not establish whether QBC should be the method of choice in practical applications where only a limited set of data is available.

In the peptide-MHC system, ANNs tend to agree on those peptides that do not bind, and a high QBC is often associated with a high affinity (Buus et al., 2003). This is in agreement with the high-affinity peptides being highly underrepresented in the data set available for training. Thus, it can be expected that a selection strategy where novel peptides are selected based on the predicted affinity will result in increasing performance simply because it will provide more high-affinity peptides. To evaluate if high-QBC data are indeed more informative than high affinity data, we performed a carefully designed experiment where two sets of data were constructed so that the peptides have close to identical affinity but very different values of QBC. Training ANNs using these two data sets, we were able to confirm that the QBC selection strategy is superior to a random strategy.

2 Materials and Methods

2.1 Data Set. Our data set comprises 528 peptide 9-mer sequences and their corresponding affinities to the HLA A*0204 molecule (Nielsen et al., 2003). The distribution of affinities in this data set is shown in Figure 1. From this set, 399 peptides have been selected by a weight matrix method, and the set is therefore enriched by peptides with high affinity as compared to a set completely randomly selected data set by random (Buus et al., 2003). The remaining 129 peptides were selected by a combination of QBC and ANN predicted affinity.
Figure 1: Affinity profile of the complete set of available data. Number of peptides within log-transformed \((1 - (\log(\text{affinity})/\log(5\times10^5)))\) affinity intervals of 0.1.

2.2 Artificial Neural Networks. All ANNs were of a standard amino-acid-encoding, three-layer feedforward type (Qian & Sejnowski, 1988) with \(9 \times 20 = 180\) input neurons (encoding a 9mer peptide), and continuous output. Amino acids were sparsely encoded, and target affinities were presented to the networks as \(1 - (\log(\text{affinity})/\log(50,000))\). Thus, peptides with high affinity have a log-transformed value close to one and nonbinders a value close to zero (Buus et al., 2003). The training of the ANN was done in a balanced manner. In each training cycle, an equal number of peptides with an affinity below and above 200 nM was presented to the network. This forces the networks to consider no or low binders equally important as high binders. Five networks were obtained, each trained on four-fifths of the data and tested on the last fifth, thus generating an ensemble of five networks trained and tested on different subsets of the data. All networks were initialized with the same weights that were all randomly set to 0.1 or -0.1. Networks were trained using backpropagation, and the prediction performance during training was monitored using minimal error. For each training and test data set, the network training is stopped when the prediction error on the test data is minimal. In the following ANN performance, evaluations were obtained as fivefold cross-validation performances.
2.3 Basic Experimental Setup. In all experiments, we divided the complete set of available data randomly into three pools:

1. A start pool used as training and test set pool for the first-generation networks
2. A select pool from which new peptides are selected
3. An evaluation pool, with peptides that are used for performance evaluation and not used during training and testing

Neural networks are initially derived from peptides in the start pool. Subsequently, all peptides in the select pool are ranked using a specific selection. The top $N$ ranked data are added to the start pool, and the resulting (enriched) pool comprises the training and test set used to make the next-generation network. The procedure is repeated until the select pool is empty. At all iterations, the neural network performance is evaluated using the peptides in the evaluation pool. The evaluation pool comprises a constant and large fraction of all available data to preserve comparability and objectivity, respectively; all experiments should be as comparable as possible, and the more data that are included in the evaluation pool, the greater the chance is that they are actually representative of the complete pool of unknown real-world data. In this work, we test the performance of a series of six selection schemes to select data points for use in the subsequent training round:

1. QBC: Point sorted on descending standard deviation, $S$, of the outputs of the ensemble of ANNs
2. Predicted binding affinity: Points sorted on descending predicted affinity, $P$
3. Product of QBC and predicted binding affinity: Points sorted on descending $S \cdot P$
4. Random selection: Points selected at random
5. Negative QBC: Points sorted on ascending $S$
6. Negative product of QBC and the predicted binding affinity: Points sorted on ascending $S \cdot P$

In order to establish the statistical significance of one selection scheme’s being better than another, 500 experiments were performed, and we counted the number of times, $X$, that the first scheme performed the best. The number of times, $Y$, that the second scheme 2 performed best is then $500 - X$. Two selection schemes could, for instance, be high QBC versus low QBC. $X$ and $Y$ were so large they can be assumed to follow Poisson distributions and

$$z = \frac{X - Y}{\sqrt{X + Y}}$$
can be assumed to follow a gaussian distribution with a mean of zero and a standard deviation of one (Armitage, Berry, & Matthews, 2002). From the \( z \) value one can readily calculate the corresponding \( p \)-value.

3 Results

We started with start pools comprising 60 peptides, thus leaving 268 and 200 peptides in the select and evaluation pools, respectively. These experiments showed that 60 data points, when using fivefold cross-validation, were not enough for the second-iteration ANNs to benefit more from any of the biased selection procedures compared to pure random selection (data not shown). This was likely to be because 60 data points (48 in training and 12 in test) are not sufficient for the networks to learn even the most basic binding patterns (e.g., which amino acids are allowed and or disallowed at the anchor positions), and thus the networks have not been “taught” enough to obtain any useful predictive performance and disagree in a meaningful manner.

Repeating similar experiments, gradually increasing the number of peptides in the start pool, resulted in a consistent increase in the prediction accuracy of both QBC- and affinity-selection schemes over the random selection scheme. Figure 2 shows the results of 20 repeated experiments with start pools comprising of 90 peptides. All 20 experiments were extended by repeatedly adding the top 30 data points from each of the six ranking schemes (see materials and methods), iterating until the select pool was emptied. All 20 experiments were constructed so that they share the same evaluation pool, whereas the start pool and the select pools were randomly sampled from the residual 328 \((528 - 200)\) data points. The average Pearson correlation coefficient (Corr) and corresponding standard deviation over the 20 experiments are shown as a function of the number of peptides in the training and test sets. The Pearson correlation coefficient (also called the linear correlation coefficient or Pearsons \( r \)) is defined as

\[
\text{Corr} = \frac{\sum_i(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i(x_i - \bar{x})^2 \sum_i(y_i - \bar{y})^2}}
\]

for pairs of quantities \((x_i, y_i)\) with means \(\bar{x}\) and \(\bar{y}\), respectively (Press, Flannery, Teukolsky, & Vetterling, 1989). Figure 3 shows that all three positive nonrandom methods are superior to the random method above about 100-120 data points, and clearly, that the “negative” selection schemes clearly have the reverse effect. It is less clear which of the positive selection schemes perform the best.

In the following, we keep 120 data points in all starting pools, 200 data points in the evaluation pool, and 208 data points in the select pools. Figure 3 shows a scatter plot of QBC values versus the log-transformed binding affinities for a set of 208 peptides in the select pool. The figure also shows
that the ANNs tend to agree on which peptides do not bind and that a high QBC often is associated with a medium or high affinity. The running mean of the data further shows a parabolic tendency with a maximum QBC value for intermediate binding peptides. This suggesting that the ANNs have more difficulty in learning the amino acid pattern for intermediate-binding peptides compared to that of high-binding peptides. The question was therefore whether it was the QBC method itself or just the principle of selecting peptides with a high or intermediate affinity that accounted for the successful predictive performance. To answer this question, it is was necessary to unlinked high-QBC values from high measured affinity and compare two sets of peptides with close to equal binding affinities and maximal difference in QBC values. In order to obtain such sets of peptides, we sorted all peptides in the select pool according to measured affinity. Then for every two peptides in the sorted list, we distributed the one with the highest QBC to a “high-QBC” set and the other to a “low-QBC” set. In this way, two sets, each comprising 104 peptides, were obtained. Each of these
Figure 3: Log-transformed binding affinity as a function of the QBC measure in a typical experiment. The $x$-axis shows the log-transformed affinity and the $y$-axis the QBC measure (standard deviation of the predictions) of each of the 208 peptides in the select-pool. Most peptides are placed at the zero log-transformed affinity line, reflecting the overrepresentation of nonbinders in the data set. Most nonbinders and peptides with low affinity have a low QBC value, whereas peptides with an intermediate and high affinity tend to have a higher QBC. In grey is shown a running mean curve over QBC values in a window of 25 data points.

sets was then added back to the starting pool, resulting in two training sets for second-generation networks each comprising 224 (120 + 104) peptides.

The performances of the two second-generation networks were then compared in 500 separate experiments. For each experiment, the difference in mean QBC and mean performance between the low-QBC and high-QBC set was determined. In 332 out of 500 (66.4\%) experiments, the high-QBC set resulted in a higher performance than the low-QBC set. Using the approximations described in materials and methods, we find that the high-QBC selection scheme performs significantly better than the low-QBC selection scheme ($P < 0.001$). We find that the difference in mean binding affinity in the high-QBC and low-QBC sets was close to zero and that there was no correlation among the high-QBC sets between mean binding affinity and performance. This demonstrated that the results obtained are unrelated to
the binding affinity. Thus, a clear separation of QBC and binding affinity was obtained, and the efficiency of QBC data selection strategy was shown to be efficient independently of affinity. In real-life applications, the measured affinity is not known, and the peptides must be ranked using the predicted affinity. For comparison, we performed 500 experiments where sorting on predicted affinity does the splitting of the select pool. This resulted in a positive/negative ratio of 310/190 and a corresponding $p$-value $P < 0.001$, which shows that in this example, the high QBC set also gives significantly higher performance than the low QBC.

Estimating the mean performance and mean standard error of the high-QBC and low-QBC enriched sets for the 500 second-generation networks, we obtain the values $0.833 \pm 0.0025$ and $0.825 \pm 0.0025$, respectively. The mean values are separated by close to four standard deviations, again demonstrating that the QBC selection strategy is capable of defining data points of high information content independent of peptide binding affinity.

4 Discussion

A model system of MHC class 1 peptide 9-mer affinity data was used to evaluate a QBC algorithm for selection of particularly high-informative data. A criterion for the successful outcome of using an algorithm that relies on prediction methods is that a sufficient amount and quality of data are available. In the case of peptide-MHC affinity, we demonstrated that a threshold of about 100 peptides is needed to generate the first predictors below which consistency in the performance of the QBC algorithm gradually vanishes. Above the this threshold, both the QBC- and affinity-based prediction schemes were superior to a random selection scheme.

To determine whether QBC or high affinity accounted for the improved performance, we constructed pairs of peptide sets where we maximized the difference in QBC values while keeping the distribution of affinities similar. Selecting sets of data with high QBC and comparing to corresponding low-QBC sets with similar affinity distribution resulted in higher-prediction performances of the high-QBC peptides. Thus, we show by example that QBC is a superior selection scheme independent of high affinity.

Selecting data with high QBC results in higher performance of the next-generation network. The performance differences obtained here might appear insignificant. However, it is important to keep in mind that the set of peptides to choose from during our experiments was prohibitively small as compared to the possible number of 9-mer peptides: $9^{20}$ corresponding to approximately $10^{11}$ peptides. Thus, in real life situations, this difference is likely to be considerably enlarged because the networks will have the opportunity to choose the best possible most informative peptides, which are unlikely to be present in our data set.

The presented findings show that QBC can be used to speed up the improvement of methods for predicting peptide-MHC binding. Moreover,
the result here is completely general and can be used in any other field of bioinformatics or other discipline in which obtaining new data in a directed manner is an option. However, it is important to keep in mind the minimal amount of data necessary to obtain reliable results using a QBC strategy because alternative methods might perform better when only very limited amounts of data are available.

Acknowledgments

This work is a part of the IHWG. It was supported by the Danish National Research Foundation, the Danish MRC (grant 22-01-0272), the 5th Framework programme of the European Commission (grant QLRT-1999-00173), and an NIH (grant AI49213-02).

References


Received March 7, 2003; accepted May 27, 2003.