S. Buus
S.L. Lauemøller
P. Worning
C. Kesmir
T. Frimurer
S. Corbet
A. Fomsgaard
J. Hilden
A. Holm
S. Brunak

Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach

Key words:
Artificial Neural Network (ANN); MHC class I; predictions; Query by Committee (QBC); specificity

Acknowledgments:
This work was supported by the Danish MRC (grant 22-07-0272), the 5th Framework Programme of the European Commission (grant QLGT-1999-00173), the NIH (grant AI49213-, and the Danish National Research Foundation.

Abstract: We have generated Artificial Neural Networks (ANN) capable of performing sensitive, quantitative predictions of peptide binding to the MHC class I molecule, HLA-A*0204. We have shown that such quantitative ANN are superior to conventional classification ANN, that have been trained to predict binding *vs* non-binding peptides. Furthermore, quantitative ANN allowed a straightforward application of a 'Query by Committee' (QBC) principle whereby particularly information-rich peptides could be identified and subsequently tested experimentally. Iterative training based on QBC-selected peptides considerably increased the sensitivity without compromising the efficiency of the prediction. This suggests a general, rational and unbiased approach to the development of high quality predictions of epitopes restricted to this and other HLA molecules. Due to their quantitative nature, such predictions will cover a wide range of MHC-binding affinities of immunological interest, and they can be readily integrated with predictions of other events involved in generating immunogenic epitopes. These predictions have the capacity to perform rapid proteome-wide searches for epitopes. Finally, it is an example of an iterative feedback loop whereby advanced, computational bioinformatics optimize experimental strategy, and vice versa.

Authors' affiliations:

S. Buus[1],
S.L. Lauemøller[1],
P. Worning[2],
C. Kesmir[2],
T. Frimurer[2],
S. Corbet[3],
A. Fomsgaard[3],
J. Hilden[4],
A. Holm[5],
S. Brunak[2]

[1]Division of Experimental Immunology, Institute of Medical Microbiology and Immunology, University of Copenhagen,

[2]Center for Biological Sequence Analysis, The Technical University of Denmark,

[3]Department of Virology, State Serum Institute,

[4]Department of Biostatistics, University of Copenhagen,

[5]Research Center for Medical Biotechnology, Chemistry Department, Royal Veterinary and Agricultural University, Denmark

Correspondence to:
Søren Buus
Institute of Medical Microbiology and Immunology
Panum 18.3
Blegdamsvej 3
DK-2200 Copenhagen N
Denmark
Tel: +45 3532 7885
Fax: +45 3532 7853
e-mail: S.Buus@immi.ku.dk

Received 17 February 2003, revised 16 May 2003, accepted for publication 23 May 2003

Tissue Antigens. ISSN 0001-2815

Tissue Antigens 2003 **62:** 378–384
Printed in Denmark.

Proteomes are extremely diverse and can be used to ascertain the identity of any organism. This is true even at the level of oligopeptides. Indeed, the immune system has chosen peptides as one of its prime targets. It follows that proteomes can be translated into immunogens once it is known how the immune system generates and handles peptides (1). One of the most selective events is that of peptide binding to MHC. It is therefore important to establish accurate descriptions and predictions of peptide binding to the most common MHC haplotypes.

The function of MHC class I molecules (MHC-I) is to sample intracellularly processed peptides, transport them to the cell surface and display them to cytotoxic T cells (CTL) (reviewed in 2, 3). It has been estimated that MHC-I can bind approximately

0.5% of the universe of 9-mer peptides (conversely, more than 99% is ignored) making this one of the most selective events of antigen presentation (3). However, even 0.5% of the universe of 9-mer peptides is still a sizable number of different peptides (the universe of 9-mer peptides encompasses 5.12×10^{11} members; 0.5% of this is 2.56×10^{9}). Such broad peptide binding specificity is achieved through the recognition of so-called 'motifs' representing important requirements needed for binding such as the presence and proper spacing of certain amino acids within the peptide sequence (4–9). The most important of these are known as primary anchor residues. In general, there are two to three primary anchor positions and together they constitute a 'simple motif' (reviewed in 9). However, other features such as secondary anchors and disfavored residues, adding up to an 'extended motif', are also important in defining peptide–MHC interaction (8). The most elaborate extended motifs consist of detailed statistical matrices representing the frequency of every amino acid in every position (10, 11). Predictions of peptide-MHC binding are usually performed as motif searches, which are far from perfect (12, 13). A simple motif search has a modest sensitivity; actually, it misses 70% of all binders. An extended motif search has a better sensitivity although it still misses about 30% of all binders (12); and the improvement comes at the cost of a large increase in the number of false positives (12). A likely reason for this lack of efficiency of motif searches is the necessary, but incorrect, assumption that the effect of each amino acid is independent of the sequence context. Although these assumptions may merit considerable justification as an approximation (10), further improvements in predictions may require that the entire peptide is considered including any sequence specific correlated effects (14). In addition, it is highly desirable that future prediction schemes are quantitative (15). This would allow the identification of high, intermediate, and low affinity binders, all of which are of potential biological interest. Finally, quantitative predictions would facilitate the integration of predictions of MHC binding with predictions of other events involved in antigen presentation (1).

Many other methods or refinements have been suggested including quantitative matrices (11), hidden Markov models (16), and rule-based models using binding motifs (17), however, it is generally not possible to incorporate correlated effects with any of these approaches. These can be incorporated with molecular modeling (18), but this is computer and labor intensive, and not amenable to high throughput analysis. ANN have gained increasing popularity as an efficient way to store and extract information from complex data (19). It combines the analysis of correlated effects with high throughput, and promising results have been obtained with ANN-driven predictions of peptide–MHC interactions (20–24). Briefly, ANN are trained to recognize an input (*in casu*, a specific peptide sequence) associated with a given output (the corresponding MHC binding affinity). Once trained, they recognize the complicated peptide patterns, including the correlated or non-linear effects, compatible with binding. Of utmost importance for the success of such a data-driven approach is the generation of representative and information-rich data for training. Here, we describe an efficient approach for obtaining such data and demonstrate the ability of ANN to perform quantitative predictions of peptide binding to MHC. The MHC class I molecule, HLA-A*0204, is used to exemplify this approach, however, we expect that it can be generalized to all other peptide-binding HLA-molecules.

Materials and methods

Peptide-MHC class I binding assay

Human MHC-I molecules, HLA-A*0204, were affinity-purified as previously described (25, 26). Peptides were purchased from Schaefer-N, Copenhagen, or synthesized using a standard FMOC-protection strategy (27). The purity of the peptides was verified by HPCL (>80%) and the identity by mass spectrometry. MHC-I molecules were incubated for 48 h at 18°C with increasing concentrations of test peptide and a fixed concentration (about 2 nM) of radiolabeled indicator peptide in the presence of 3 μM human β_2m as previously described. Binding was examined by Sephadex G50 spun column gel filtration under conditions, where the IC_{50} approximated the K_D (31).

Development of Artificial Neural Networks

The neural networks were of the standard feed-forward type (32). Details of sequence encoding, error function, etc. may be found elsewhere (28). The predictive performance was monitored using the Pearson correlation coefficient during training and testing of the networks. Training was terminated using early stopping (32).

Redundant sequence data sets were made non-redundant using the first version of the Hobohm procedure (30). In a sequential scan of the data in a database, this method discards sequences where the similarity in a pair-wise alignment exceeds a given scoring threshold. The resulting non-redundant data set has therefore no pairs of sequences with a similarity exceeding that threshold. When creating a non-redundant version of SWISS-PROT, a threshold of 25%

sequence identity was used. When computing the similarity between 9-mer peptides, the information content in a sequence logo (29) of binders was used to weight the positions in the calculation of the similarity.

Results and discussion

Generating representative sets of quantitative peptide-MHC binding data (primary data selection)

On average, 0.5% of the universe of 9-mer peptides binds MHC with high affinity (3). Consequently, using a random selection strategy, one would have to test 20,000 peptides experimentally to get a mere 100 independent examples of peptide-MHC binding. This is an untenable proposition and mandates a preselection strategy. To test such a selection strategy we used a biochemical peptide-HLA-A*0204 binding assay, which was already available to us (11). Due to cost and time constraints, all relevant data that were already available were included (peptides from proteins of particular interest such as p53, HSP70, HIV proteins). These were combined with data selected by screening the entire SWISS-PROT database for binders using specificity matrices developed by a previously described combinatorial peptide library approach (11). This resulted in the identification of more than 100,000 peptides, which were sequence similarity reduced based on the Hobohm algorithm (30, see Methods). A number of peptides were selected, synthesized, verified by HPLC and mass spectrometry, and only those that contained peptide of the expected mass and purity were included in the subsequent binding analysis. Thus, the primary selection and synthesis led to the generation of a panel of about 400 9-mer peptides. These were tested for binding to HLA-A*0204 using an *in vitro* biochemical peptide-MHC class I binding assay (31), which is capable of measuring the equilibrium dissociation constant, K_D, over a range of almost five decades (from 1 nM to 50,000 nM). The distribution of peptide binding affinities was: 1% very good binders ($K_D < 5$ nM), 11% good binders (< 50 nM), 14% intermediate binders (< 500 nM), 10% low affinity binders (< 5000 nM), 19% very low affinity binders (< 50,000 nM), and 44% non-binders (> 50,000 nM). Thus, all ranges of binding, even intermediate and low affinity binding, is going to contribute to the predictions presented here. Accurate quantitative predictions covering a range of binding affinities would be of immunological interest.

Quantitative artificial neural networks (ANN); a novel strategy of balanced training in a continuum

When applied to sequence analysis, ANN have typically been used for classification purposes. Previous attempts to use ANN to predict peptide binding to MHC have sought to identify binders *vs* non-binders as defined by different thresholds (20, 21, 24) (in some cases, peptides are classified into additional categories including intermediate and low affinity binding (22, 23)). From an information theoretic perspective, a measured binding affinity contains more information than a binary binding/non-binding resolution. Thus, a quantitative training approach should extract more information from the available experimental data than conventional classification approaches would do. To better represent the quantitative differences of peptide–MHC interactions throughout the measurable range, the actual binding values were transformed logarithmically prior to ANN development. This means that a small error in determining a good binder (e.g., predicted 9 nM *vs* observed 12 nM) will be considered as grave as a large error in determining a poor binder (predicted 9000 nM *vs* observed 12,000 nM). Thus, our approach will attempt to emphasize different binding ranges equally well, i.e., each example of intermediate binding will contribute as much to the ANN training as each example of high affinity binding.

The different binding ranges might not be represented at the same frequency in the available data (e.g., in our case there are few examples of very good binders and many examples of non-binders). In a conventional classification ANN approach, variations in the size of the different categories can be offset by representing each category at the same frequency during training (32). In the case of MHC binding, any random selection of peptides will be heavily skewed towards non-binders; even with our preselection strategy, there is some skewing towards non-binders. We have therefore devised a novel ANN approach, which performs a balanced training in a continuum. The available data set was randomly subdivided into seven parts of roughly the same size. This allowed for training of seven different ANN and cross-validation of the predictive performance. Furthermore, each training set was distributed into two to five bins according to the observed binding affinity. Each interval of binding affinity was represented with the same frequency during training (i.e., in every training cycle, all peptides from the least represented bin were included, while a similar number were randomly chosen from each over-represented bin). This forced the ANN to consider the prediction of binders, intermediate binders and non-binders equally important. The architectures of the ANN contained 180 input neurons (one for each of the 20 natural amino acids in each position of a 9-mer peptide), between two and 10 hidden neurons, and one output neuron.

The seven unique ANN constructed to predict the binding of 9-mer peptides to HLA-A*0204 were used to cross-validate the approach (Fig. 1). A highly significant ($P < 0.001$) correlation between the logarithms of the predicted and observed binding

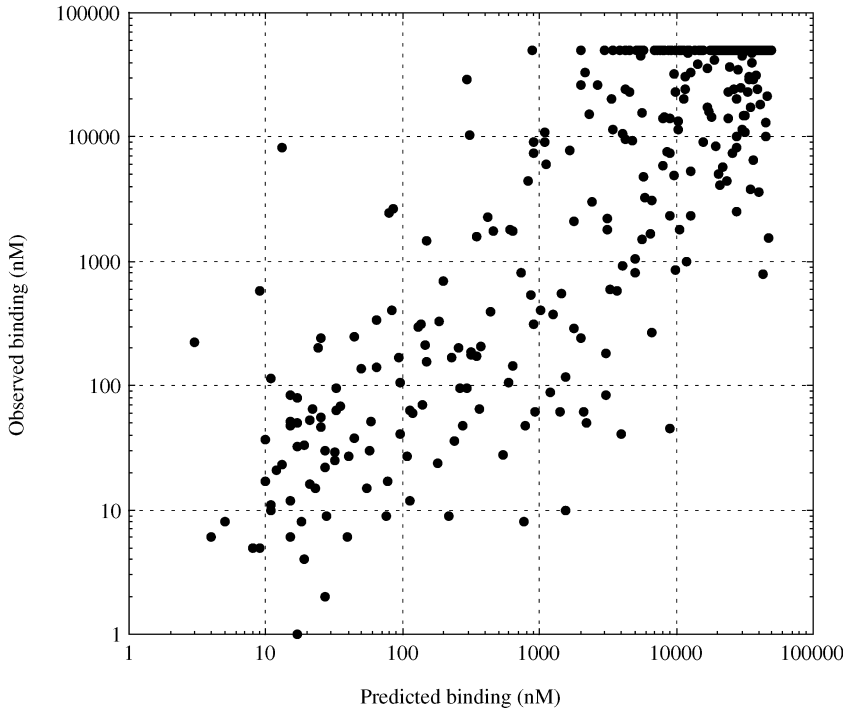

Fig. 1. ANN can perform quantitative predictions of peptide–MHC–I interaction. The binding affinity was measured in a biochemical assay (31) and expressed as the logarithm of the equilibrium dissociation constant (K_D (nM)). Subsequently, first generation ANN were trained to quantitatively predict the logarithm of the affinity of peptide binding to HLA-A*0204 using a cross-validation approach. This allowed the affinity of every peptide to be predicted by an ANN, which had not been trained on the peptide in question. The logarithm of the predicted binding *vs* the logarithm of the observed binding was plotted and analyzed by linear regression. The regression line was $y = 0.99x - 0.02$ ($n = 397$, $C_{Pearson} = 0.87$, $P < 0.001$).

was found by linear regression analysis. The regression line was close to the expected $y = x$ demonstrating that ANN indeed can be trained to predict binding *quantitatively*. For comparison purposes, ANN of the conventional classification type (i.e., such as those of Gulukota et al. (21)) were generated, and their output was fitted and calibrated in the best way possible as a work-around to obtain quantitative predictions (data not shown). As expected, the quantitatively trained ANN ($C_{Pearson} = 0.87$) were significantly ($P < 0.01$) more accurate (i.e., able to predict the experimental value) and more precise (i.e., reproducible) than the classification trained ANN ($C_{Pearson} = 0.73$). This supports our contention that a measured affinity contains more information than a binary binding/non-binding encoding. To our knowledge this is the first attempt to use quantitative data to train an ANN to predict peptide binding to HLA molecules, that is, we train the actual affinity values. In contrast, others have used a binary binding/non-binding encoding (20, 21, 24), or a more elaborate grading of binding (a 'staircase' encoding) (22, 23) for training purposes.

Unfortunately, it is not possible to validate our ANN-driven server against any of the other reported ANN-driven predictions of peptide–MHC interactions as none of these have been made available publicly and none have addressed the specificity of HLA-A*0204. Strictly speaking, only another HLA-A*0204 prediction can be compared to the present ANN-driven HLA-A*0204 prediction. We have previously used positional scanning combinatorial peptide libraries (PSCPL) to generate a quantitative HLA-A*0204-specific peptide binding matrix, and shown that it can be used to predict peptide binding. Reassuringly, the ANN-driven prediction ($C_{Pearson} = 0.87$) described here outperformed this matrix-based HLA-A*0204 prediction ($C_{Pearson} = 0.85$). In an attempt to perform an independent validation of the present ANN prediction, we compared it to predictions of the closely related HLA-A*0201 (a single methionine to arginine substitution at position 97 distinguishes HLA-A*0201 from HLA-A*0204). Two publicly available HLA-A*0201 predictions are in frequent use: BIMAS at http://bimas.dcrt.nih.gov/molbio/hla_bind/and SYFPEITHI at http://www.syfpeithi.de. Best possible fits of the BIMAS and SYFPEITHI HLA-A*0201 predictions (as in Udaka et al. (33)) were compared to our ANN-driven HLA-A*0204 prediction. Our ANN-driven prediction outperformed both matrix-driven predictions (BIMAS $C_{Pearson} = 0.83$ and SYFPEITHI $C_{Pearson} = 0.81$). Thus, the precision of the ANN-driven prediction is superior to comparable matrix-driven predictions. We attribute this to the ability of ANN-, but not matrix-, driven methods to incorporate correlated effects. Another notable advantage of the ANN-driven prediction is its ability to predict the exact binding affinity value, whereas the BIMAS prediction is somewhat arbitrary (some of the values are even assigned) and the SYFPEITHI prediction is completely arbitrary (all values are assigned). Thus, the accuracy of the ANN prediction is considerably better than that of the two competing predictions.

ANN development through successive generations; an iterative process

Although we attempted to select the primary training sets in an unbiased and representative way, the selection process was constrained by the existing data and further shaped by PSCPL-generated matrix predictions (11), i.e., it only involved an infinitely small part of the universe of 9-mer peptides. To counteract these problems, we devised an iterative approach for ANN training and development (34). We applied the seven ANN to all the possible peptides from SWISS-PROT and calculated the standard deviation of the predicted binding of each peptide. We suggest that a low SD indicates that the underlying pattern is well represented in the training set and that the prediction model successfully generalizes for this particular peptide. Conversely, a high SD indicates that the underlying pattern is poorly represented and/or that the model generalizes poorly. In any event, the latter peptides should be experimentally tested and included in subsequent training. Such a selection approach is similar to what has been named 'Query by Committee' (35). We included two additional requirements in our next generation data selection: (i) the previous generation of ANN should be in disagreement (i.e., the SD should be high), and (ii) at least one of the seven ANN should predict the peptide in question to be a good binder. To assure that the selected peptides were representative, a large cohort of peptides was selected according to QBC and subsequently similarity reduced to a more manageable number. Finally, 65 QBC-selected peptides were synthesized and tested for binding to HLA-A*0204. This new data set was added to each of the original seven training sets, and the ANN training repeated. In parallel, a similar sized panel of control peptides, which had not been selected by QBC, was synthesized, tested and added as described above. This yielded two different 2nd generation ANN, $2^{\circ}\mathrm{ANN}^{(\mathrm{QBC})}$ and $2^{\circ}\mathrm{ANN}^{(\mathrm{non-QBC})}$, respectively, which were cross-validated on the same data as for 1°ANN cross-validation. The QBC effect was most pronounced with respect to the ability to quantitatively predict very high affinity binders (<50 nM), where the $2^{\circ}\mathrm{ANN}^{(\mathrm{QBC})}$ ($C_{\mathrm{Pearson}}=0.43$) was significantly more precise ($P<0.01$) compared to the $2^{\circ}\mathrm{ANN}^{(\mathrm{non-QBC})}$ ($C_{\mathrm{Pearson}}=0.36$). Note that these results were obtained by merely expanding the training sets from 340 to 405 peptides.

An intriguing effect of the QBC principle was found when a complete scanning of the more than 3 million 9-mer peptides extracted from a non-redundant subset of SWISS-PROT was performed. Averaging the predictions of the seven networks in the ensemble, we found that the 1°ANN and the $2^{\circ}\mathrm{ANN}^{(\mathrm{non-QBC})}$ predicted roughly the same amount $\approx$0.25% of all peptides as being very high affinity binders. In contrast, the $2^{\circ}\mathrm{ANN}^{(\mathrm{QBC})}$ predicted not only these $\approx$0.25%, but additionally $\approx$0.17%, of all peptides as being very high affinity binders. Thus, the QBC approach was unique in the sense that it suggested the existence of many more high affinity binders. Previous methods have been missing at least one out of three possible binders (12), and it is tempting to speculate that the QBC approach might account for these missing binders. In that case, this gain in sensitivity could be achieved at minimal expense in terms of the number of false positives. Thus, the QBC principle appears to be an efficient way to improve both the quality and the coverage of ANN.

While this work was in progress (34), Udaka et al. Reported a QBC-like approach to predict binding to MHC molecules (33). Although there are similarities between their and our approaches, there are also significant differences. They used a hidden Markov model (HMM) as the predicting algorithm, whereas we used an ANN approach. Whenever one considers correlated effect – as for peptide-MHC binding – ANN will have an advantage over HMM, which can only consider independent contributions. As the predicting algorithm determines the outcome of the QBC procedure, one should in this case expect an ANN-driven QBC procedure to be better than a HMM-driven procedure. Furthermore, we used an entirely different strategy for selecting information-rich peptides for expansion of the data set: (i) when searching for peptides we used SWISS-PROT, as it is clear that there is no need for improvement of the prediction method in the part of sequence space not 'visited by nature'. In contrast, many of the randomly generated peptides used by Udaka et al. will not improve a method used for scanning of real-world pathogens, because such sequences are never found in real proteins; (ii) we require that at least one committee member predicts the Swiss-Prot peptide is a strong binder as there is no point in improving the method for prediction of non-binders, as epitope scanning will aim for the detection of intermediate to strong binders, and (iii) we used a novel balanced training strategy that gives much better prediction for all binding intervals (without using extra peptides).

To illustrate the 2°ANN(QBC), we used the database of MHC binding peptides maintained by Rammensee and coworkers (www.syfpeithi.de). HLA-A*0204 is very poorly represented in this database, but the closely related HLA-A*0201 can be extracted. Using the HLA-A*0204 predicting 2°ANN(QBC), 23% and 68% of 187 HLA-A*0201-restricted 9-mer peptides were predicted to be high and intermediate affinity binders, respectively. For comparison, only 0.42% and 2.8% of randomly selected peptides are predicted to be high and intermediate affinity binders to HLA-A*0204, respectively (Fig. 2) ($P<0.001$).

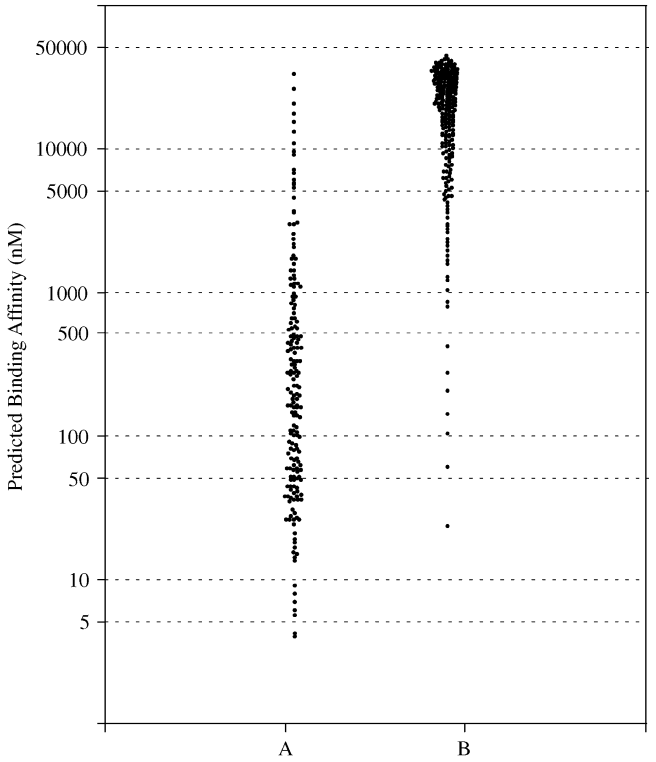

Fig. 2. Prediction of binding of natural peptides *vs* random peptides. Second generation ANN prediction of the binding affinity of (A) 187 natural HLA-A*0201-eluted 9-mer peptides extracted from the Rammensee database (www.syfpeithi.de) or (B) 230 randomly chosen 9-mer peptides extracted from SWISS-PROT.

A proteome-wide quantitative prediction of peptide–MHC interaction

ANN have the capacity to handle entire proteomes in a reasonably short period of CPU time. A prospective experiment was performed to test this on a small scale where we searched for immunogenic epitopes from HIV (Corbet et al. in press). The 81 full-length HIV genomes, and the many more full length HIV genes available in public databases ultimo 1999, were translated into proteins and scanned for the presence of HLA-A*0204 binding peptides (this corresponded to the scanning of >750.000 9-mer peptides) by the 1° ANN. The peptides were sorted according to their degree of conservation (assuming that a conserved epitope is a better vaccine candidate than an epitope only present in one HIV isolate), and according to their predicted binding. Of the 56 HIV-specific, HLA-A*02xx-restricted (the different members of the HLA-A2 supertype have largely similar specificities) epitopes, which were already known at that time (Los Alamos HIV epitopes database + MHCPEP), 46 (or 82%) had a predicted binding affinity below 500 nM. Thus, using 500 nM as a cut-off we would only have missed 18% of the known epitopes. In contrast, we gained more than 100 hitherto unknown epitope candidates, which met both the requirements of conservation and predicted MHC binding. Of these new putative HIV epitopes, 52 were synthesized and tested for binding; 36 (or 69%) were verified as binders, and the majority of these were recognized in HLA-A2 + HIV-1 patients (Corbet et al. in press). Thus, we can project that already the 1° ANN may have more than doubled the number of known HIV-derived, HLA-A*02xx binders and epitopes, and we can expect that further gains will be achieved with the 2° ANN(QBC).

The human MHC project

Taken together, our results suggest that quantitative ANN-driven predictions of peptide–MHC interactions can be generated. Such ANN promise unprecedented high sensitivity (ability to identify positives) and high specificity (ability to reject negatives) predictions. Our results also suggest a rational iterative approach to generate such ANN and demonstrate the utility of a selective sampling method such as the QBC. It therefore seems reasonable to suggest that this technology could be used to describe and predict human MHC specificities systematically. There are several hundred MHC-I alleles in the human population (36). One should start with the most common MHC-I; eventually they should all be included. As evolutionary selected variants, we expect that they will differ with respect to their peptide binding specificity and that the corresponding predictions can lead to a more detailed functional description of MHC-I polymorphism.

The long-term goal should be to integrate MHC-I predictions with those of other steps involved in antigen processing and presentation including peptide generation by proteasome digestion (37) (www.cbs.dtu.dk/services/NetChop/), peptide transport by the TAP complex (38), etc., in effect, simulating the biology of the immune system (1). These bioinformatics tools should be linked to genome/proteome databases enabling genome/proteome-wide searches for epitopes of immunological interest. This would allow scientists and clinicians to examine any organism or protein of interest for the presence of potentially immunogenic epitopes and should provide a rational approach to vaccine development and immunotherapy.

Finally, it should be noted that the iterative QBC approach is an example of how a mutual feedback between computational bioinformatics and experiments (wet biochemistry) can guide each other to achieve optimal performance with the least amount of time and resources invested. A successful integration of computation and experiment might enable the evaluation of every single member of an otherwise unmanageably large sequence space.

The ANN generated in this paper are publicly available at www.cbs.dtu.dk/services/NetMHC/

References

1. Lauemoller SL, Kesmir C, Corbet S et al. Identifying cytotoxic T cell epitopes from genomic and proteomic information: "The human MHC project". *Rev Immunogenetics* 2001: **2**: 477–91.
2. Rock KL, Goldberg AL. Degradation of cell proteins and the generation of MHC class I-presented peptides. *Annu Rev Immunol* 1999: **17**: 739–79.
3. Yewdell JW, Bennink JR. Immunodominance in MHC class I restricted T lymphocyte responses. *Annu Rev Immunol* 1999: **17**: 51–88.
4. Sette A, Buus S, Colon SM, Smith JA, Miles C, Grey HM. Structural characteristics of an antigen required for its interaction with Ia and recognition by T cells. *Nature* 1987: **328**: 395–9.
5. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee H-G. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 1991: **351**: 290–6.
6. Jardetzky TS, Lane WS, Robinson RA, Madden DR, Wiley DC. Identification of self peptides bound to purified HLA-B27. *Nature* 1991: **353**: 326–9.
7. Sette A, Buus S, Apella E et al. Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc Natl Acad Sci U S A* 1989: **86**: 3296–300.
8. Ruppert J, Sidney J, Celis E, Kubo RT, Sette A. Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell* 1993: **74**: 929–37.
9. Rammensee H-G, Friede T, Stevanonic S. MHC ligands and peptide motifs: first listing. *Immunogenetics* 1995: **41**: 178–228.
10. Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*, 1994: **152**: 163–75.
11. Stryhn A, Pedersen LØ, Romme T, Holm CB, Holm A, Buus S. Peptide binding specificity of major histocompatibility complex class I resolved into an array of apparently independent sub-specificities: quantitation by peptide libraries and improved prediction of binding. *European J Immunol* 1996: **26**: 1911–8.
12. Kast WM, Brandt RMP, Sidney J et al. Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16, E6 and E7 proteins. *J Immunol* 1994: **152**: 3904–12.
13. Andersen MH, Tan L, Sondergaard I, Zeuthen J, Elliott T, Haurum JS. Poor correspondence between predicted and experimental binding of peptides to class I MHC molecules. *Tissue Antigens* 2000: **55**: 519–31.
14. Fremont DH, Stura EA, Matsumura M, Peterson PA, Wilson IA. Crystal structure of an H-2Kb-ovalbumin peptide complex reveals the interplay of primary and secondary anchor positions in the major histocompatibility complex binding groove. *Proc Natl Acad Sci U S A*, 1995: **92**: 2479–83.
15. Yewdell JW, Bennink JR. Cut and trim: generating MHC class I peptide ligands. *Curr Opin Immunol* 2001: **13**: 13–8.
16. Mamitsuka H. Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins* 1988: **33**: 460–74.
17. Nijman HW, Houbiers JGA, Vierboom MPM et al. Identification of peptide sequences that potentially trigger HLA-A2.1 restricted cytotoxic T lymphocytes. *European J Ournal Immunol* 1993: **23**: 1215–9.
18. Rognan D, Scapozza L, Folkers G, Daser A. Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes. *Biochemistry* 1994: **33**: 11476–85.
19. Baldi P, Brunak S. *Bioinformatics. The Machine Learning Approach*. Cambridge: MIT Press, 1998: 105–125.
20. Adams HP, Koziol JA. Prediction of binding to MHC class I molecules. *J Immunol Meth* 1995: **185**: 181–90.
21. Gulukota K, Sidney J, Sette A, De Lisi C. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J Mol Biol* 1997: **267**: 1258–67.
22. Honeyman MC, Brusic V, Harrison LC. Strategies for identifying and predicting islet autoantigen T-cell epitopes in insulin-dependent diabetes mellitus. *Ann Med* 1997: **29**: 401–4.
23. Honeyman MC, Brusic V, Stone NL, Harrison LC. Neural network-based prediction of candidate T-cell epitopes. *Nature Biotechnol* 1998: **16**: 966–9.
24. Milik M, Sauer D, Brunmark AP et al. Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nat Biotechnol* 1998: **16**: 753–6.
25. Olsen AC, Pedersen LØ, Hansen AS et al. A quantitative assay to measure the interaction between immunogenic peptides and MHC class I molecules. *European J Immunol* 1994: **24**: 385–92.
26. Pedersen LØ, Hansen AS, Olsen AC, Gerwien J, Nissen MH, Buus S. The interaction between beta 2-microglobulin (β_2m) and purified class I major histocompatibility (MHC) molecules. *Scand J Immunol* 1994: **39**: 64–72.
27. Meldal M, Bisgaard Holm C, Bojesen G, Havsteen Jacobsen M, Holm A. Multiple column peptide synthesis. Part 2. *Int J Peptide Protein Research* 1993: **41**: 250–60.
28. Blom N, Hansen J, Blaas D, Brunak S. Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. *Protein Sci* 1996: **5**: 2203–16.
29. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucl Acids Res* 1990: **18**: 6097–100.
30. Hobohm U, Scharf M, Scheider R, Sander C. Selection of representative protein data sets. *Protein Sci* 1992: **1**: 409–17.
31. Buus S, Stryhn A, Winther K, Kirkby N, Pedersen LØ. Receptor–ligand interactions measured by an improved spun column chromatography technique. A high efficiency and high throughput size separation method. *Biochim Biophys Acta* 1995: **1243**: 453–60.
32. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 1994: **19**: 55–72.
33. Udaka K, Mamitsuka H, Nakaseko Y, Abe N. Empirical evaluation of a dynamic experiment design method for prediction of MHC class I-binding peptides. *J Immunol* 2002: **169**: 5744–53.
34. Fomsgaard A, Brunak S, Buus A, Corbet SL, Lauemoller SL, Hansen JE. *HIV Peptides and Nucleic Acids Encoding Them for Diagnosis and Control of HIV Infection.* Denmark: State Serum Institute, 1999: patent WO 01/55 177.
35. Seung H, Opper M, Sompolinsky H. *Query by committee. Fifth Annual ACM Workshop on Computational Learning Theory*. San Mateo, CA: Morgan Kaufmann, 1992: 287–94.
36. Marsh SG. Nomenclature for factors of the HLA system, update June 2000. WHO Nomenclature Committee for factor of HLA System [In Process Citation]. *Tissue Antigens*, 2000: **56**: 289–90.
37. Kesmir C, Nussbaum AK, Schild H, Detours V, Brunak S. Prediction of proteasome cleavage motifs by neural networks. *Protein Engineering* 2003, in press.
38. Brusic V, van Endert P, Zeleznikow J, Daniel S, Hammer J, Petrovsky N. A neural network model approach to the study of human TAP transporter. *In Silico Biol* 1999: **1**: 109–21.