An Attempt to Improve the Quantitative Epitope Prediction by Modelling Alternative Binding Modes

Filippo Aluffi-Pentini[1], Valeria De Fonzo[2] and Valerio Parisi*[,2]

[1]*Dipartimento Metodi e Modelli Matematici, Università di Roma "La Sapienza", Via A. Scarpa 16, 00161 Roma, Italy*

[2]*Dipartimento di Medicina Sperimentale, Università di Roma "La Sapienza", Viale Regina Elena 324, 00161 Roma, Italy*

Abstract: *Motivation*: A good quantitative epitope prediction, i.e. a reliable prediction of the strength of the MHC-epitope binding, is decisive in order to better understand the immune system response. The prediction is often performed by means of the scoring-matrix method that usually assumes a single binding configuration: each amino acid of the epitope binds to a specific pocket of the MHC molecule, in a way independent from other bindings.

Results: We have put forward the assumption, suggested by the allosteric Monod framework, that a number of alternative states exist, each one characterised by an interaction energy expressed by a scoring matrix. We have developed and suitably evaluated an algorithm for epitope prediction based on such assumption, and we finally discuss the results and the possible reasons why such results unexpectedly appear to be unsatisfactory.

INTRODUCTION

In the normal operation of the human immune system, the Major Histocompatibility Complex (MHC) molecules - of class I and II - interact, within a cell, with short peptides obtained from the fragmentation of various proteins and present such peptides on the cell surface for a possible recognition as foreign antigens by the receptors (TCR) exhibited from the T-cells, a necessary condition for activating the immune response.

The peptides presented by MHC molecules are derived from the fragmentation, for class I, of cytosolic proteins (possibly of a virus), while, for class II, of extracellular proteins (possibly of a parasite).

MHC class I molecules are found on almost every cell of the body (but not on red blood cells), while MHC class II molecules are found only on a few specialized cell types, e.g. macrophages, which are professional antigen-presenting cells.

In the interaction a peptide binds to an MHC molecule in a site called groove, more exactly the side chain of every amino acid (aa) of the peptide binds to a corresponding site in the groove, called pocket.

The groove interacts with a peptide tract having a length of about 9 aa.: while the geometric shape of a groove of an MHC class I molecule is such that the maximum length of the whole peptide must be bounded (9 aa or only slightly more), so that no extra tract of the peptide can hang outside the groove, such steric constraint does not exist for an MHC class II molecule.

*Address correspondence to this author at the Dipartimento di Medicina Sperimentale, Università di Roma "La Sapienza", Viale Regina Elena 324, 00161 Roma, Italy; Tel: +39 06 4991 0774; Fax: +39 338 0 9981736; E-mail: valerio.parisi@uniroma1.it

The probability that a molecule of a peptide will be presented to the TCR grows with the binding affinity (or, equivalently, with the free energy of the interaction) between the MHC molecule and the peptide. When the binding affinity to a given MHC allele is greater than a threshold value (defined by the fact that beyond that value the possible immune response is considered good), the peptide is usually named epitope. When the binding affinity is not sufficiently known, or even unknown, the peptide should be more appropriately named a prospective epitope, although, for the sake of simplicity, this terminology is not always strictly adhered to.

It would be clearly of utmost practical importance to know the affinity between a given MHC allele molecule and a given peptide molecule. However measuring all such binding affinities requires a great number of sophisticated and expensive experiments, and it is therefore only natural to resort to a computational prediction, validated by a much smaller number of key measurements.

Two types of epitope prediction are usually considered: qualitative or quantitative. The first one consists in discriminating the peptide as epitope and non-epitope, without giving any further information about the binding affinity. The quantitative type consists in predicting a figure of merit for the binding, sometimes by using probabilistic arguments, but more frequently by giving a measure of the binding affinity, expressed by means of the experimental quantity known as IC_{50}, which represents the molar concentration of the peptide such that half of the MHC molecules are bound, so that when the binding is strong the IC_{50} is small. More specifically a peptide can be considered an epitope if its IC_{50} is less then a given threshold.

The IC_{50} is by far the most widely used. Other quantities, possibly with different physical dimensions, are sometimes

1875-0362/09

considered as a measure of the binding strength, such as, for example, the EC_{50} (which many consider as equivalent to IC_{50}) or the half-life of the MHC-peptide complex, or the dissociation costant K_d. Since such quantities are directly or inversely proportional to each other (usually with a good approximation, especially in the case of small concentrations) when the logarithm of the binding affinity is used (as we shall see later on), different measurements of the binding strength involve only an inessential shift.

While there exist different types of human MHC molecules and epitopes with different lengths, in this paper we shall restrict our study to the following largely used choice: we consider only the alleles of the MHC class I molecules, and only the epitopes whose length is 9 amino acids.

We note that for the MHC class I molecules the epitope prediction appears to be theoretically simpler, thanks to the steric constraint described above, and it is easier to get many experimental data; moreover such molecules are very important since they are synthesised by almost all human cells in order to counter viruses and cancers.

We now briefly consider some of the main prediction methods.

Practical methods used to perform epitope prediction include methods based on artificial intelligence, on binding motifs, and on quantitative matrices.

1) Artificial intelligence methods, mostly based on Artificial Neural Networks (ANN), are general-purpose algorithms that look for purely empirical relations, and in some cases obtain remarkably good results; but it is very difficult to obtain from them some physical insight.

2) Methods based on binding motifs arise from the observation that, for a specific allele, the epitopes frequently exhibit the same motif, i.e. have in particular positions particular aa that are presumably the anchors of the binding. For example Parker and co-workers remarked that the HLA-A2 molecules bind with epitopes that contain Leucine or Methionine at the second position, and Valine or Leucine at the ninth position [1]. The frequency of such occurrences has suggested to simply consider as prospective epitopes those exhibiting such a motif. Usually for each allele only one motif is used, but sometimes the situation is more complex; for example, in [2] two motifs are considered for a particular allele (A*0101), in order to model two alternative peptide binding modes.

3) Methods based on quantitative matrices: while a motif simply indicates whether a given aa can occur in a given position of an epitope for a specific MHC allele, more information is given by a scoring matrix, called quantitative matrix, which for each aa and each position gives a "figure of merit" (linearly related to the binding free energy) for that aa in that position; according to a common hypothesis of independent binding of side chains (IBS), a global figure of merit for every prospective epitope can be obtained simply as the sum of the relevant partial figures of merit [1].

We now consider in more detail the methods based on quantitative matrices.

From a physical point of view each partial figure of merit is simply the interaction partial free energy of an aa with its pocket, and the IBS assumption states that such energies are mutually independent [1].

We write down the formulas for the special but frequent case where the peptide length of the prospective epitope is equal to 9, so that, in each of its nine positions, any considered peptide has one of the 20 aa. We shall call $a(k)$ the aa in position k ($k = 1, \cdots, 9$) of the peptide a. In this case the quantitative matrix, for a specific allele, has 9 rows and 20 columns: and, to give an example, if the aa are alphabetically ordered by one-letter symbol (so that the Y, for the Tyrosine, represents the last aa) the partial binding free energy of a possible Tyrosine in the third position of the peptide is just the element in row 3 and column 20.

The problem is to predict (for a specific allele) the value of IC_{50} of a peptide as a function of its nine aa. Since, from the thermodynamic point of view, the binding strength depends only on the binding free energy, according to the IBS assumption, the free energy of the epitope binding to the MHC molecule should be considered as the sum of the free energies of the bindings of each aa of the peptide with the corresponding pocket, with no mutual interactions.

In other words the IBS assumption provides a model for computing the binding strength that can be written in the form

$$S_a(M) = \sum_{i=1}^{9} M_{i,a(i)} \tag{1}$$

where the considered MHC allele is not explicitly indicated, a refers to the considered peptide, the quantitative matrix M is the table, defined above, whose element $M_{i,a(i)}$ expresses the contribution, to the free energy of the MHC-peptide binding, by the binding of the amino acid $a(i)$ to the pocket i, and $S_a(M)$ is an estimate of the quantity $A_a = \ln(IC_{50})$.

Recalling Arrhenius equation, A_a is linearly related to the binding free energy $\Delta G^\circ = -RT \ln(IC_{50})$, where T is the absolute temperature and R is the gas constant [3].

In order to use the model for epitope prediction one needs of course to know the matrix M.

If the matrix M is not known, in order to obtain a good estimate it is natural to minimize in some sense the overall "discrepancy" between the values of A_a and S_a.

We now describe the most frequently used procedure.

One defines the deviation as

$$d_a(M) = S_a(M) - A_a \tag{2}$$

and the total square deviation (relative to the given MHC allele) as

$$D(M) = \sum_a d_a^{\,2}(M) \tag{3}$$

A value M^{opt} minimizing D, i.e. implicitly defined by

$$D\left(M^{opt}\right) = \min_M \left(D(M)\right) \tag{4}$$

is considered as a good estimate of M, and the value of $S_a\left(M^{opt}\right)$ is considered as a good estimate of A_a.

We note however that in this form, as it can be easily seen, the minimization problem is degenerate: in fact if, given a matrix M, we define a matrix M^g with $M^g_{i,k} = M_{i,k} + g_i$ with 9 arbitrary g_i subject to $\sum_i^9 g_i = 0$ we have $S_a\left(M^g\right) = S_a(M)$ for each a, and therefore $D\left(M^g\right) = D(M)$.

This corresponds to $9-1=8$ degrees of freedom, and therefore, in a different parlance, the problem has ∞^8 minimizers.

The main steps of a method for quantitative epitope prediction, using a model based on IBS, are:

- compute a value M^{opt} based on experimental values of A_a for a number of a (training set)
- use M^{opt} to compute, for a number of different a, the value of $S_a\left(M^{opt}\right)$ as the prediction of the value of A_a.

As a side remark we note that the quantitative matrix M is often called an additive matrix since the score S is obtained by a suitable sum of its elements (1). A perfectly equivalent point of view, first introduced by [1], is to consider a quantitative matrix M^* of elements $M^*_{i,\alpha} = \exp\left(M_{i,\alpha}\right)$ to obtain a score S^* as a suitable product of elements of M^*,

$$S_a^*\left(M^*\right) = \prod_{i=1}^{9} M^*_{i,a(i)} \tag{5}$$

so that the quantitative matrix M^* is called a multiplicative matrix and obviously

$$S_a^*\left(M^*\right) = \exp\left(S_a(M)\right) \tag{6}$$

A small annoying complication, which however has an important effect on the basic formulas in what follows, stems from the form in which some of the experimental data are often provided, due to limitations of the adopted measurement procedures.

The measured value A_a^+ is often provided as having a definite value A_a, but sometimes is given in an indefinite form $A_a^+ < U_a$ or $A_a^+ > L_a$, i.e. as having an indefinite value belonging to an "out of range" incertitude zone defined by its boundary value (U or L).

Since usually both definite and indefinite values are present, one selects a conventional value for the deviation $d_a(M)$ as follows.

When the measure has a definite value, one obvious selects $d_a(M) = S_a(M) - A_a$ as in Eq. (2).

When the measure has an indefinite value, one selects as the value of $d_a(M)$ the smallest deviation consistent with the measurement: in other words, if S_a belongs to the same zone as A_a^+ one selects $d_a(M) = 0$. Otherwise one uses the corresponding boundary value $d_a(M) = S_a(M) - U_a$ or $d_a(M) = S_a(M) - L_a$.

Summing up, if some measurements have indefinite values, the deviation is given by

$$d_a(M) = \begin{cases} S_a(M) - A_a & \text{if} \quad A_a^+ = A_a & & \\ 0 & \text{if} \quad A_a^+ < U_a & \text{and} & S_a(M) < U_a \\ 0 & \text{if} \quad A_a^+ > L_a & \text{and} & S_a(M) > L_a \\ S_a(M) - U_a & \text{if} \quad A_a^+ < U_a & \text{and} & S_a(M) > U_a \\ S_a(M) - L_a & \text{if} \quad A_a^+ > L_a & \text{and} & S_a(M) < L_a \end{cases} \tag{7}$$

We note that if all the experimental data were provided as having a definite value, the total square deviation $D(M)$ would be a quadratic function of the elements of the matrix M. However since frequently many experimental data are given as having only an indefinite value, as described above, the $D(M)$ becomes a piecewise quadratic function of the elements of M, whose numerical minimization can be more troublesome.

It is very important to quantitatively evaluate an epitope prediction. To this end a number of indicators can be used: among them, for example, the AUC of the ROC (the Area Under the Receiver Operating Characteristic Curve that is a plot of the sensitivity vs. 1-specificity) [4], is widely used recently, and will be considered later on.

Published results about the methods based on IBS exhibit basically the same, moderately good, predictions; some minor improvements in epitope prediction have been recently obtained by adding to the total square deviation a "regularisation function", i.e. a small perturbation, producing a greater stability of the minimizers (e.g. with respect to abnormal experimental data). Such regularizazion is usually called after the russian mathematician Tikhonov [5], but sometimes other terms are used [6] .

The IBS assumption accounts for the simplicity of such methods, but also for their limitations; and in order to improve them it is only natural to resort to less restrictive assumptions.

OUR MODEL

The model we propose in the present paper is based on a generalization, inspired by simple thermodynamics considerations, of the model based on the IBS assumption,

which is here completely abandoned. More specifically our model, in analogy to the allosteric model of Monod, Wyman and Changeux, is based on the assumption that there exist a number N of alternative peptide-MHC binding modes, or, in other words, N different docking possibilities. This amounts to consider N quantitative matrices M_k, each one relative to the binding mode k, so that Eq. (1) becomes:

$$S_{a,k}\left(M_k\right) = \sum_i M_{k,i,a(i)} \tag{8}$$

where $M_{k,i,a(i)}$ is relative to the binding of aa $a(i)$ and the pocket i for binding mode k.

We note that such quantitative matrices (as well as those used under the IBS assumptions) are examples of the "position specific scoring matrices", used also in other bioinformatics fields, such as those used, for example, to represent the specificity of transcription factors.

The total binding strength, according to thermodynamics, is related to the binding strength of the different binding modes, by the equation

$$\tilde{S}_a\left(M_1,\cdots,M_N\right) = -\ln\left(\sum_{k=1,N}\exp\left(-S_{a,k}\left(M_k\right)\right)\right) \tag{9}$$

An intuitive interpretation of this equation can be simply obtained as follows. If we consider the time fraction during which the epitope and the MHC molecule remain bound, we can say that: for the binding in any single mode k the fraction is proportional to $\exp\left(-S_{a,k}\right)$, while for the overall binding for all modes the fraction is both equal to the sum of all time fractions relative to all modes, and proportional to $\exp\left(-\tilde{S}_a\right)$, whence the result. Obviously the special case of only one binding mode ($N=1$) corresponds to the IBS assumption, and Eq. (9) reduces to Eq. (1).

In practice a very good approximation to Eq. (9) (that however is not used in our calculations) can be given by $\tilde{S}_a\left(M_1,\cdots,M_N\right) \cong \min_{k=1,N}\left(S_{a,k}\left(M_k\right)\right)$.

We note that the equations of an allosteric model deal with different states quite independently from their geometrical difference; in more detail we hypothesize that different docking modes for a given peptide-allele pair may exist, also when, as in our case, the relative positions are very similar, for example if only small rigid displacements (translations or rotations) are allowed by the steric constraints of the epitope in the groove.

Our model (9) enables to perform a quantitative epitope prediction relative to a given MHC allele, using the corresponding matrices M_k; if nothing is known (or guessed) about them, but a number of experimental data about the allele-peptide interaction are available, an obvious step is to estimate such matrices by suitably minimizing a total square deviation, as was done in the above one-matrix models using Eq. (1).

The deviation (2), when the measured value has a definite value ($A_a^+ = A_a$), becomes $\tilde{d}_a\left(M_1,\cdots,M_N\right) = \tilde{S}_a\left(M_1,\cdots,M_N\right) - A_a$ otherwise we use the obviously adapted version of Eq. (7).

The total square deviation (3) becomes

$$\tilde{D}\left(M_1,\cdots,M_N\right) = \sum_a \tilde{d}_a{}^2\left(M_1,\cdots,M_N\right) \tag{10}$$

We choose to minimize the regularised deviation (following Tikhonov)

$$\hat{D}\left(M_1,\cdots,M_N\right) = \tilde{D}\left(M_1,\cdots,M_N\right) + \lambda\cdot\sum_{k,i,j}\left(M_{k,i,j}-\bar{M}\right)^2 \tag{11}$$

where $\overline{M}$ is the arithmetic average of all the $M_{k,i,j}$, and λ is a suitable constant.

The global minimizer $\left\{M_1^{opt},\cdots,M_N^{opt}\right\}$ of $\hat{D}$, implicitly defined by

$$\hat{D}\left(M_1^{opt},\cdots,M_N^{opt}\right) = \min_{M_1,\cdots,M_N}\left(\hat{D}\left(M_1,\cdots,M_N\right)\right) \tag{12}$$

is of course defined up to permutations, since $\tilde{D}$ (and therefore also $\hat{D}$) is a symmetrical function of $M_1,\cdots,M_N$.

RESULTS

In order to perform a check of our model, we have chosen the large database of experimental values of IC_{50} relative to peptide binding to MHC class I molecules that is freely available at http://mhcbindingpredictions.immune-epitope.org. Such database has been proposed [7] as a common benchmarking resource for algorithm testing, together with suggestions for obtaining homogeneous comparisons of predictions: the data are already grouped in five sets, thus allowing a uniform use of the cross-validation procedure, as described below.

We have not used all the data, but we have "filtered" them as follows. We have first considered only human MHC class I alleles and peptides of length 9. In order to avoid making prediction based on clearly insufficient data, we have considered only alleles with at least 1000 experimental values of IC_{50}. As a last filtering we have discarded the alleles with less than 800 definite values, so that we end up with a final choice of 10 alleles.

Our results have been obtained using the "five-fold cross-validation" procedure, advocated by [7].

In more detail, for each one of the selected ten MHC alleles, we use the data, already partitioned in five sets, taking in turn as test set one of the five sets, and as training set the union of the remaining four sets, and we proceed as follows:

- we obtain, with a suitable local minimization algorithm, based on conjugate gradients, the minimizer matrices (12) of the deviation $\hat{D}$, using as the source of experimental values the training set.

- we obtain the predicted value $\tilde{S}_a$ for each peptide a of the test set, using Eq. (9) with the minimizer matrices.

This procedure is repeated by selecting each time a different test set, so that we have finally, collecting the results, a prediction $\tilde{S}_a$ for each peptide.

Finally, the quality of the prediction for a given MHC allele is evaluated, by means of the value of the AUC of ROC using as IC_{50} threshold value (i.e. the value under which a peptide is considered an epitope) the concentration of 500 nM, as in [7].

We note that if one considers only one binding mode, and therefore only one quantitative matrix (just as in any IBS-based procedure), and all the experimental data have definite values, the required minimizations boil down to simple quadratic expression minimizations.

If instead some data are given as having an indefinite value belonging to an "out of range" incertitude zone, the function to be minimized becomes a convex piecewise quadratic function, for which the classical quasi-Newton techniques are well suited.

Things become markedly different when we consider several binding modes, i.e. several quantitative matrices, since in this case many local minima of the deviation arise; therefore, in order to perform the above procedure, we have developed a global minimization algorithm inspired by simulated annealing.

We report in Table **1** the results obtained with the above procedure for each one of the ten selected alleles, in the simplest case of only two matrices (with a value of $\lambda = 1$), together, for the sake of comparison, with some other results, as described in the legenda.

As for the choice of λ, some preliminary test had indicated that the best results were obtained, with minimal changes in the quality of the prediction, only when λ remained around 1; we therefore simply selected $\lambda = 1$.

From a simple inspection of Table **1** a number of conclusions can be drawn.

We recall that, as explained above, the idea of introducing several quantitative matrices (instead of only one) was suggested by the hope of better modelling the case of several alternative peptide binding modes (as in allosteric models), and therefore the aim of our model was to assess the possible amount of improvement, if any, that could be obtained when using a several-matrix model with respect to using a simple one-matrix model.

We first note that, according to our expectations, our results in column $N = 1$ are well compatible with those in column SMM (small differences in the results are unavoidable when equivalent procedures differ in many details, such as regularisations and optimisation techniques).

On the other hand the results for the two-matrix case are truly disappointing: contrary to our expectations, not only the results are not better, but they are clearly worse than those in the one-matrix case, the only barely better case being in the first line (A_0201).

DISCUSSION

It is only natural to attempt an explanation for such unexpected result. We put forward a few simple ideas.

Table 1. The Table Reports (in Col. 2, 3, 4) the ROC-AUC Values Obtained in Different Conditions for a Number of MHC Alleles, Following the "Five-Fold Cross-Validation" Procedure

MHC Allele	*N* = 1	*N* = 2	SMM	Total Data	Definite-Valued Data
A_0201	0.952	0.953	0.952	3089	1998
A_0203	0.913	0.900	0.916	1443	1293
A_0202	0.894	0.871	0.899	1447	1279
A_1101	0.947	0.933	0.948	1985	1272
A_0206	0.913	0.886	0.914	1437	1237
A_0301	0.938	0.921	0.940	2094	1179
A_6802	0.897	0.871	0.898	1434	1177
A_3101	0.930	0.906	0.930	1869	1063
A_6801	0.879	0.840	0.885	1141	988
A_3301	0.922	0.894	0.925	1140	822
A_0201_{half}	0.944	0.928		1545	999

For each allele (specified in the first column), we report our results, obtained with one (N = 1) and two (N = 2) matrices, while in the SMM column we report, for comparison, the results obtained by Peters et. al (2006) with the SMM algorithm (which is IBS-based, and therefore, up to programming details, should be equivalent to our column N = 1). The last two columns report the number of data used to obtain the above results: the fifth column contains the total number of data (i.e. both the experimental data having a definite value and those given as having an indefinite value belonging to an "out of range" incertitude zone) while the last column contains only the number of data having a definite value. We note that for N = 1 we used a local optimisation algorithm (based on conjugate gradients), while for N = 3 we used a global optimisation algorithm. The last row (A_0201_{half}) contains our results obtained for the allele A_0201, using only half of the data.

The first naïve hypothesis could be simply that (contrary to our hopes) adding a second matrix is quite useless, and in this case this would be correctly suggested by the cross-validation procedure, that should give worse results when useless parameters are added.

This is analogous to the case of a mean-square regression with few data, where, when adding clearly useless parameters, a simple fitting always gives some (possibly small) improvement, while a correct cross-validation generally produces worse results.

A closer look to the results suggests another hypothesis that of course needs a better validation: the second matrix could be useful, but the data are simply not enough; in more detail, the data are barely sufficient to fit one matrix and are too few to reasonably fit two matrices. In other words, the hypothesised gain in using a second matrix could be overridden by the loss due to the combined effect of the scarcity of data and the cross-validation procedure, thus producing an overall negative balance.

A full confirmation of such hypothesis will of course only be obtainable when much more data, homogeneous and benchmark-ready, will be available. As for now (i.e. with the current data) a first partial confirmation can be obtained by reducing the number of data: if results are nearly equivalent, the hypothesis is clearly disproved, i.e. the data are enough; if instead the results are already worse for the one matrix case, and still worse for the case of two matrices, then we have a good cue for the validity of the hypothesis. To this end we have performed a small test: we have considered the allele A_0201, having roughly twice the number of data available for the others, and for which the two-matrices results are not clearly worse than the one-matrix results, and we have suitably taken one half of the data (A_0201_{half}), so that their number becomes comparable to those of the other alleles. The results for one matrix are worse (0.944 vs. 0.952), and those for two matrices are even worse (0.928 vs. 0.953), which is just what we expected (based on our hypothesis); moreover the results are comparable to those of the other alleles, and since now the size of the data is comparable, it is tempting to consider that such a validation of our hypothesis could be applicable also to all the other alleles.

From a somehow different point of view, a rough confirmation of the above assumption about the scarcity of available data can be provided by a well-known conventional rule of thumb for pattern recognition stating that the number of training samples should be 5–10 times the number of model parameters [8].

According to this rule, in order to satisfactorily estimate the 360 parameters relative to our case of two matrices, we should need 1800–3600 training samples, which in the framework of the adopted five-fold cross-validation procedure, amounts to 2250–4500 required total samples; and it appears that there is a remarkably good agreement between these figures and our results for N=2 in all the considered cases: the data are barely sufficient in the first row, and grossly insufficient in all the others.

For the sake of completeness we may add that, as far as we know, the only other generalization of the model based on the IBS assumption is the model by [6] that aims to represent a possible influence between local bindings.

More specifically [6] considers a number of correction terms M' to be simply added to Eq. (1), each term representing the (pairwise) influence of the interaction between a side chain $a(j)$ and its pocket j on the interaction between another side chain $a(k)$ and its pocket k, so that the score of the peptide-MHC interaction can be written

$$S_a(M) = \sum_{i=1}^{9} M_{i,a(i)} + \sum_{j=1}^{8} \sum_{k=j+1}^{9} M'_{j,a(j),k,a(k)} \qquad (13)$$

In fact in the considered paper for practical reasons only a small number of correction terms in the above general formula are taken as non null.

As for the results obtained with the above model, we note that, in spite of the great increase of the number of adjustable parameters, the author claims only very slight improvements in epitope prevision.

We may end up by noting that, at least with the considered results obtained so far with the available data, there is no evidence of significant improvements obtained by generalizing IBS.

APPENDIX

A Toy Model

In order to provide some insight in the comparison of various assumptions, we consider a simplified toy model representation, enabling a pictorial interpretation. While obviously the involved molecular objects are three-dimensional, and their displacements are roto-translations, a very simple two-dimensional graphical representation of the mutual influence between two aa-pocket interactions, with or without IBS, could be sufficiently illuminating and therefore is presented in what follows.

Both the epitope and the MHC groove are represented as simple strip-shaped horizontal objects, and we model different aa-pocket binding modes by means of small horizontal relative displacements, so that, with respect to the same pocket, it may occur that while a given aa could energetically prefer to bind in a given place within a pocket, another aa could energetically prefer to bind in a slightly different position within the same pocket, thus possibly involving corresponding displacements within other pockets.

We note however that of course the formulations of the mathematical model are completely independent from the pictorial representation.

For example a simplified special case of the binding equilibrium MHC-epitope (showing a four aa section) with the IBS assumption is shown in the toy model in Fig. (**1**). The MHC molecule is designed with only one binding site in each pocket, and the geometry allows simultaneously for each aa an optimal binding, so that there is only one possible binding mode.

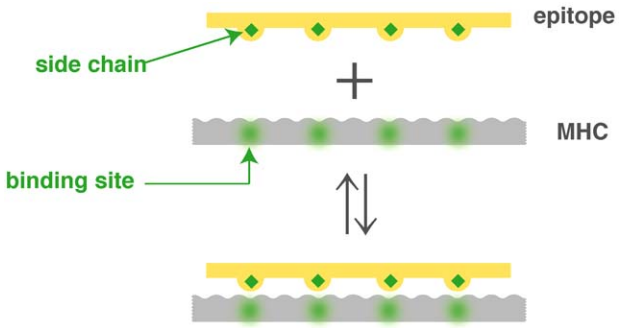

Fig. (1). Illustration of the binding equilibrium MHC-epitope by showing a restricted section (with only four equal aa) of an epitope, and the corresponding restricted section of the MHC groove (with only four pockets). The four side chains are represented by four green diamonds and for each side chain the energetically favoured binding site inside the corresponding pocket is shown as a blurred disk of the same colour. The example represents the standard IBS situation where all amino acids bind simultaneously each one to its optimal site.

We now consider our model. A possible effect of introducing two quantitative matrices, i.e. two binding modes, can be well discussed with a simple toy model, with reference to Fig. (**2**), which represents the binding equilibrium MHC-epitope for the same allele and two different epitopes, both cases with two different binding modes.

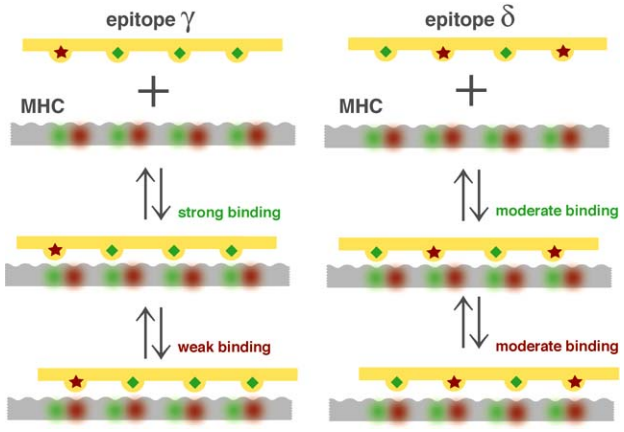

Fig. (2). Two arbitrary illustrative cases of binding equilibrium MHC-epitope relative to two different epitopes binding to two MHC equal grooves (showing only four aa and four pockets); in both cases the equilibrium involves the transitions between three states: epitope and MHC molecule unbound, or bound in two different modes (energetically evaluated by two different matrices M_1 and M_2, and graphically differentiated by a small horizontal shift). For the sake of simplicity the chemical equilibrium arrows between the first and the third state are omitted. The various side chains are identified by different symbols (green diamonds and red stars), and for each side chain the energetically favoured binding site inside the corresponding pocket is shown as a blurred disk of the same colour.

The MHC molecule is designed with two binding sites in each pocket, so that there are two sequences of equidistant binding sites (with green or with red sites). The two possible binding modes are as follows: epitope γ has a "strong" binding (with three aa in a favoured binding site) and a "weak" binding (with only one aa in a favoured binding site), while for epitope δ the two bindings are both "moderate" (both with two aa in a favoured binding site).

As for the model described in Eq. (13), a possible effect of the corrections terms can be well discussed with reference to the toy model in Fig. (**3**). According to such model, the Fig. (**3**) represents the binding equilibrium MHC-epitope for the same MHC allele and two different epitopes. The MHC molecule is designed with two binding sites in the second pocket, so that the sequence of blue-green-blue-blue binding sites is equispaced, perfectly matching the epitope α, while the sequence of blue-red-blue-blue binding sites is non equispaced, so that the matching with the epitope β is painful.

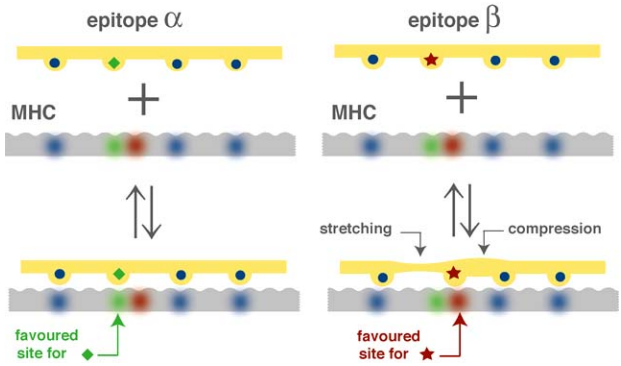

Fig. (3). Two arbitrary illustrative cases of binding equilibrium MHC-epitope relative to two different epitopes binding to two MHC equal grooves (showing only four aa and four pockets). The various side chains are identified by different symbols (green diamonds, blue balls and red stars), and for each side chain the energetically favoured binding site inside the corresponding pocket is shown as a blurred disk of the same colour.

In the first case each aa binds optimally with its preferred pocket so that the energy is given only by the matrix M; in the second case each aa still binds with its preferred pocket, but in a non-optimal way, with an energy penalty which is given by the terms M', and which, in the graphics of the toy model, is represented by the small dislocation of some aa with respect to the optimal position inside the pocket, and by a deformation of the epitope.

REFERENCES

[1] K. C. Parker, M. A. Bednarek, and J. E. Coligan, "Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains", *J. Immunol.*, vol. 152, no. 1, pp. 163-175, January 1994.

[2] A. Kondo, J. Sidney, S. Southwood, M.-F. Del Guercio, E. Appella, H. Sakamoto, H.M. Grey, E. Cells, R.W. Chesnut, and R.T. Kubo, "Two distinct HLA-A*0101-specific submotifs illustrate alternative peptide binding motifs", *Immunogenetics*, vol. 45, no. 4, pp. 249-258, January 1997.

[3] P. Guan, I. A. Doytchinova, C. Zygouri and D. R. Flower, "MHCPred: a server for quantitative prediction of peptide-MHC binding", *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3621-3624, July 2003.

[4] A.P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognit.*, vol. 30, no. 7, pp. 1145-1159, July 1997.

[5] A.N. Tikhonov and V.Y. Arsenin, *Solutions of Ill-Posed Problems*. New York: John Wiley, 1977.

[6] B. Peters, W. Tong, J. Sidney, A. Sette, Z. Weng, "Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules", *Bioinformatics*, vol. 19, no. 14, pp. 1765-1772, September 2003.

[7] B. Peters, H.-H. Bui, S. Frankild, M. Nielson, C. Lundegaard, E. Kostem, D. Basch, K. Lamberth, M. Harndahl, W. Fleri, S.S. Wilson, J. Sidney, O. Lund, S. Buus, and A. Sette, "A community resource benchmarking predictions of peptide binding to MHC-I molecules", *PLoS Comput. Biol*. vol. 2, no. 6, p. e65, June 2006.

[8] L. Kanal, B. Chandrasekaran, "On dimensionality and sample size in statistical pattern recognition", *Pattern Recognit.*, vol. 3, no. 3, pp. 225-234, October 1971.

Received: June 15, 2009 Revised: August 01, 2009 Accepted: August 04, 2009