A neural network modeling of the immunogenic activity of tumor derived peptides

Ordorica Vargas Miguel Angel[1], Velázquez Monroy María de la Luz[1], Sánchez Barbosa Sandra[2]

[1]Instituto Politécnico Nacional. Escuela Superior de Medicina. Departamento de Formación Básica Disciplinaria. Plan de San Luis y Díaz Mirón. Col. Santo Tomás. México, D.F., CP 11340.

[2]Instituto Politécnico Nacional. Escuela Nacional de Ciencias Biológicas. Sección de Estudios de Posgrado e Investigación. Plan de Ayala y Manuel Carpio. Col. Santo Tomás. México, D.F. CP 11340.

mordoricav@ipn.mx

Fecha de aceptación: 12 de agosto de 2015 Fecha de publicación: 23 de septiembre de 2015

RESUMEN

Los péptidos derivados de proteínas específicas de tumores pueden activar el sistema inmune contra las células tumorales. Para promover la respuesta inmune, los péptidos deben primero unirse a las moléculas del complejo principal de histocompatibilidad. Aunque algunas posiciones de la secuencia del péptido inmunogénico necesitan aminoácidos "ancla" específicos, el resto de los residuos también influye en la afinidad de la unión. Modelamos el efecto de la sustitución de aminoácidos, en la afinidad de unión de los péptidos utilizando una red neural artificial. La red predice correctamente la especificidad de las posiciones de anclaje, pero también sugiere que los aminoácidos con propiedades similares pueden ser incluidos en la secuencia de los péptidos con alta probabilidad de tener alta afinidad de unión.

Palabras clave: redes neuronales artificiales, HLA-A.

ABSTRACT

Peptides derived from tumor-specific proteins can activate the immune system against tumor cells. To promote the immune response, peptides must first bind to the major histocompatibility complex molecules. Although some positions of the immunogenic peptide sequence need specific “anchor” amino acids, the rest of the residues also influence the binding affinity. We modeled the effect of amino acids substitution on the binding affinity of peptides using an artificial neural network. The network correctly predicts the specificity of the anchor positions but also suggests that amino acids with similar properties can be included in the sequence of the peptides with high probability of having high binding affinity.

Key words: artificial neural networks, HLA-A.

INTRODUCTION

Cancer is a major health problem worldwide. One of the most promising approaches to cancer treatment is the arousal of the patients' immune response (van der Bruggen *et al.*, 1994; Marchand *et al.*, 1995). With this strategy, it is possible to specifically eliminate cancer cells that already exist, and prevent the formation of new ones. To be able to attack cancer cells the immune system must be activated against tumor-specific antigens. The activation depends on a group of molecules known as the Major Histocompatibility Complex (MHC). The function of this set of molecules is to bind peptides derived from degraded proteins and to "present" them to the different kinds of cells that constitute the immune system. The interaction with the MHC-peptide complexes activates the immune cells. Experimentally, it has been found that the best binders are peptides 9 or 10 residues long (Parker *et al.*, 1992; Kast *et al.*, 1994) that have some positions where only specific amino acids are allowed, these are the so called "anchor" residues. However, the presence of these amino acids is not always enough to guarantee binding or high affinity; this is an important problem because higher affinity means higher probability of inducing the immune response. Several researchers have developed different strategies, including artificial neural networks, to predict binding with varying degrees of success (Hagmann, 2000). All these methods depend on the detection of anchor residues in the sequence of a target protein, but not all the methods can predict the binding affinity.

As explained before, the presence of anchor residues is a necessary, but not sufficient condition to achieve high affinity (Kast *et al*., 1994a), and sometimes not even binding. This means that the other residues can modify the binding affinity and this is an aspect that has not been explored. We have trained an Artificial Neural Network to perform a quantitative sequence activity model of the effect of changes in the amino acid sequence, on the binding affinity of the immunogenic peptides.

METHODS

Sequence and Biological Activity of Peptides

We used a sample of 82 peptides sequences (Table 1) 9 amino acids long, found in the literature (Salazar-Onfray *et al.*, 1997; Sette *et al.*, 1994; Kawakami *et al.*, 1995; Rivoltini *et al.*, 1995).

The analyzed activity (Table 1) is the binding affinity to the Human Lymphocytes Antigens A2.1 (HLA-A2.1), one of the MHC class-I molecules most frequently expressed, measured as molar concentration of a test peptide, needed to inhibit 50% of the binding of a standard peptide (IC50%), transformed to:

$$\mathbf{pIC_{50} = \text{-}\textit{log}\ IC50\%} \qquad \textbf{(i)}$$

Description of Peptide Sequence

The quantitative description of amino acids is crucial in any quantitative structure activity relationships analysis, due to the information content of the description. In our study we used the principal component scales, or Z scales, extracted by Wold (Hellberg *et al.*, 1987) from a matrix of 29 physicochemical properties of the 20 coded amino acids. These calculated descriptors are interpreted as related to hydrophilicity (Z_1), size and shape (Z_2) and electronic properties (Z_3). Each peptide was described by a vector of 9 elements, corresponding to the Z value of the residue in each position starting at the amino end. The complete set of Z values is presented in Table 2.

Network Training

We used a program that simulates a fully connected, feed forward neural network, trained with the back propagation algorithm (Rumelhart y McClelland, 1986). The network was trained with the descriptor vectors of one of the Z scales as input, and the transformed activity as target. The three scales were used individually, one at a time, and the number of hidden units was varied, to find the combination, that best predicted the affinity.

***Table 1**. Sequence and Activity of Peptides*

Sequence	pIC50	Sequence	pIC50	Sequence	pIC50	Sequence	pIC50
LLAQFTSAI	9.284	GLSRYVARL	7.377	GLKGAVTLT	6.623	MLARACQHA	5.851
YLVSFGVWI	8.721	ILSPFMPLL	7.347	IISCTCPTV	6.58	SLHVGTQCA	5.842
GILTVILGV	8.347	VLLDYQGML	7.328	WILRGTSFV	6.556	GLAANQTGA	5.785
GLLGWSPQA	8.237	YMLDLQPET	7.31	LLLEAGALV	6.544	AMFQDPQER	5.74
WLSLLVPFV	8.161	ITDQVPFSV	7.076	KLPQLCTEL	6.484	NLQSLTNLL	5.699
FLLTRILTI	8.149	YLEPGPVTA	7.022	MLLAVLYCL	6.478	VLETAVGLL	5.681
LLMGTLGIV	8.097	ILCLIFLLV	6.996	RLLGSLNST	6.447	TILLGIFFL	5.623
FLLSLGIHL	8.0	QLFHLCLII	6.886	ILTVILGVL	6.419	LTVILGVLL	5.58
KTWGQYWQV	7.959	FAFRDLCIV	6.886	NLSWLSLDV	6.415	HLLVGSSGL	5.556
MMWYWGPSL	7.921	AAGIGILTV	6.886	SLCFLGAIA	6.368	RLHKRQRPV	5.532
SLYADSPSV	7.854	ILLLCLIFL	6.845	ALVARAAVL	6.342	LQTTIHDII	5.501
KLHLYSHPI	7.77	TLIDVCPI	6.815	MLDLQPETT	6.335	TTAEEAAGI	5.38
ALMDKSLHV	7.77	PLLPIFFCL	6.796	SLNSTPTAI	6.274	ALIICNAII	5.322
LLVSGSNVL	7.62	VILGVLLLI	6.785	GLVSLVENA	6.272	CLALSDLLV	5.301
LLLCLIFLL	7.585	ILLLEAGAL	6.777	ILLGIFFLC	6.199	LLCLVVFFL	5.255
FLCKQYLNL	7.538	TLHEYMLDL	6.726	LVLMAVLYV	6.17	FLHLTLIVL	5.0
GLYSSTVPV	7.481	GTLIDVCPI	6.714	TVILGVLLL	6.072	FLALIICNA	4.954
AIIDPLIYA	7.432	SLVENALVV	6.686	FLCWGPFFL	6.049	TLPRARRRV	4.602
HLYSHPIIL	7.42	LLWFHISCL	6.682	GIGILTVIL	6.0	GLFLSLGLV	4.301
FLSLGLVSL	7.409	VLQAGFFLL	6.623	LLSSNLSWL	5.964		
YMNGTMSQV	7.398	FLGGTPVCL	6.623	FLAMLVLMA	5.917		

Activity Modeling

After the training, we selected the system that made the best prediction and presented it with a set of model peptides, made by substituting each coded amino acid for one of peptides' residues. With 9 residues and 19 possible substitutions at each position, there are 171 peptides.

***Table 2**. Principal Component Scales*

Amino acids	Symbols	Z_1	Z_2	Z_3
Alanine	Ala, (A)	0.07	-1.73	0.09
Arginine	Arg, (R)	2.88	2.52	-3.44
Asparagine	Asn, (N)	3.22	1.45	0.84
Aspartate	Asp, (D)	3.64	1.13	2.36
Cysteine	Cys, (C)	0.71	-0.97	4.13
Phenylalanine	Phe, (F)	-4.92	1.30	0.45
Glycine	Gly, (G)	2.23	-5.36	0.30
Glutamate	Glu, (E)	3.08	0.39	-0.07
Glutamine	Gln, (Q)	2.18	0.53	-1.14
Histidine	His, (H)	2.41	1.74	1.11
Isoleucine	Ile, (I)	-4.44	-1.68	-1.03
Leucine	Leu, (L)	-4.19	-1.03	-0.98
Lysine	Lys, (K)	2.84	1.41	-3.14
Methionine	Met, (M)	-2.49	-0.27	-0.41
Proline	Pro, (P)	-1.22	0.88	2.23
Serine	Ser, (S)	1.96	-1.63	0.57
Tyrosine	Tyr, (Y)	-1.39	2.32	0.01
Threonine	Thr (T)	0.92	-2.09	-1.40
Tryptophan	Trp, (W)	-4.75	3.65	0.85
Valine	Val, (V)	-2.69	-2.53	-1.29

To explore as much of the sample space as possible, we randomly selected four peptides, one from the best binders, a moderately high binder, a moderately low and one of the low binders.

The contribution of each amino acid was calculated subtracting the activity of the sample peptide from the calculated activities of its derivatives. The final contribution is the average of the four calculated values.

RESULTS

Network Training

The optimum network has 9 neurons in the input layer, 5 in the hidden layer and 1 in the output layer. We did not use a bigger network because it has been shown (Andrea & Kalayeh, 1991) that better predictions are obtained when the number of patterns is from 1.8 to 2.2 times bigger than the number of bonds in the network, and with this configuration our network is within these limits.

The best results were obtained with the Z_1 scale, for which the average difference between the observed and calculated activities was 4%. For 57 of the peptides the activity was calculated with less than 5% difference, and only 10 had a difference bigger than 10%. Because of this, we decided to use Z_1 to model the activity. A summary of results for the three scales are shown in Table 3.

***Table 3**. Training Results with the Z scales*

Z scale	Average Difference %	Difference < 5%	Difference > 10 %
Z_1	4	56	11
Z_2	4.8	54	15
Z_3	5.2	51	13

The average contribution for coded amino acids in each position is presented in Table 4

***Table 4**. Average contribution of amino acids at each position, the highest contribution at each position is underlined*

Aa	1	2	3	4	5	6	7	8	9
F	1.519	0.763	1.291	0.472	0.589	0.129	0.514	0.240	0.845
W	1.482	0.749	1.288	0.488	0.605	0.138	0.518	0.312	0.817
I	1.415	0.750	1.269	0.518	0.648	0.151	0.530	0.484	0.752
L	1.362	0.760	1.244	0.545	0.683	0.162	0.542	0.630	0.689
V	1.056	0.593	1.042	0.718	0.675	0.242	0.628	0.870	0.410
M	1.012	0.549	1.009	0.740	0.655	0.261	0.627	0.853	0.391
Y	0.650	0.335	0.844	0.843	0.524	0.411	0.503	0.746	0.246
P	0.565	0.312	0.829	0.856	0.503	0.435	0.490	0.730	0.213
A	-0.045	0.218	0.747	0.949	0.368	0.693	0.614	0.607	0.028
C	-0.332	0.202	0.708	0.992	0.319	0.869	0.734	0.549	-0.001
T	-0.450	0.199	0.696	1.006	0.306	0.921	0.776	0.531	-0.008

S	-1.141	0.190	0.631	1.075	0.258	1.090	0.996	0.447	-0.077
Q	-1.278	0.189	0.617	1.089	0.251	1.113	1.043	0.431	-0.110
G	-1.308	0.189	0.614	1.092	0.250	1.118	1.054	0.427	-0.118
H	-1.411	0.189	0.603	1.104	0.245	1.136	1.093	0.414	-0.152
K	-1.616	0.188	0.576	1.132	0.236	1.18	1.185	0.387	-0.254
R	-1.631	0.188	0.573	1.134	0.235	1.185	1.194	0.384	-0.264
E	-1.699	0.187	0.560	1.147	0.232	1.209	1.236	0.372	-0.314
N	-1.737	0.187	0.551	1.156	0.230	1.229	1.266	0.364	-0.348
D	-1.816	0.187	0.525	1.183	0.226	1.297	1.351	0.341	-0.429

These values are not absolute, they represent a relative change when the one amino acid is substituted for another.

All the positions are predicted to have some kind of selectivity. With the exception of positions 5 and 8, the highest contributions are predicted to be on the extremes of the scale. At positions 1, 2, 3, 5, 8 and 9 the network predicts a better contribution from hydrophobic residues; while at 4, 6 and 7 the hydrophilic amino acids have better contributions. Four amino acids are predicted to have high contributions at position 2: Leucine, Isoleucine, Phenylalanine and Tryptophan. L and I are known to be anchor residues, F and W are not found in the sample, but have Z_1 values similar to L and I. Methionine and Threonine are also present at position 2 of some peptides, M is predicted to have medium contribution, but T has a very low one. Valine is predicted to have a medium contribution but is not present in the sample. Predictions for position 9 are very similar. Contributions at positions 2 and 9 are presented in Figures 1 and 2.

At positions 5 and 8 the network predicts better contributions for a small group of amino acids I, L, V and M at 5, and V, M, Y and P at 8, these amino acids are all hydrophobic.

CONCLUSIONS

Although all the scales have a good performance, the best results in the training were obtained with the first principal component scale, that is related to hydrophilicity, this in coherent with composition of the peptides in the training sample. All of them are designed from experimental data which shows a better binding and affinity for peptides with hydrophobic residues at anchor positions. This design limits the available information to the net, and maybe one of the reasons of some mistaken predictions. On the other hand, the best description of the amino acids would be by using the three component scales at the same time; in this way, all the properties of the molecules are properly described. We had to work them out individually because of the small size of our sample. It would take a sample three times as big as the one we have to use the complete set of values.

Others have used different methods to predict the binding and affinity of peptides, with average success between 75% and 100% (Hagmann, 2000), but their sample sizes are much bigger than ours. Gulukota *et al*. (Gulukota *et al.*, 1997) trained a network using over 400 peptides and a frequency description, and successfully predicted binding in over 80% of the cases. Sette (Sette, 2000), used a sequence frequency analysis and a database of over 50 000 peptides to predict binding to the MHC with almost 100% of success. Parker (Parker *et al.*, 1992) also uses a sequence analysis methodology, based on several hundreds of peptides, to predict the half-life of binding to the MHC molecules with an accuracy of over 80%. Our results have not yet reached this refinement, but our data base is very small and with more data the results can be improved. In spite of the small size, the results are almost as good as the published referred data. We got almost 70% of correct predictions during the training. We believe that the information content of the principal component scales is big enough to partially compensate the small number of peptides. The simultaneous use of the three scales to describe the peptides should increase the accuracy.

Recent advances have produced bioinformatics systems to predict de binding of peptides to several classes of MHC molecules, some of them available on line, but their predictions are not significantly better than those previously mentioned (Reche, P.A. & Reinherz, E.L. 2005; Rammensee *et al.*, 1995)

As for the predicted contributions, L is the main anchor residue at position 2, and I, M and T are tolerated (Parker *et al.*, 1992; Kubo *et al.*,1994). The network predicts the contribution of L in this position as the second biggest, followed very closely by I. M is predicted to have only a medium contribution. Both, I and M, are in peptides with a wide range of affinity, but M has a more positive value of Z_1, and probably that is why its calculated contribution is smaller. Probably the very small contribution of T can be explained in the same way. Valine as M, has an intermediate calculated contribution and is found in the sample in peptides of intermediate affinity; therefore, the prediction is essentially correct.

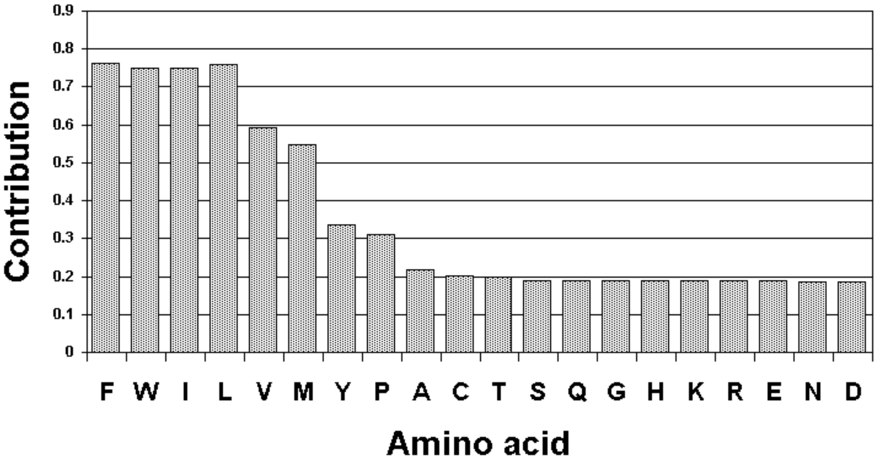

***Figure 1**. Effect of substitution at position 2. For the meaning of the symbols see Table 2.*

The predicted contributions for F and W are high, probably because they have Z_1 values more negative than L. Since neither F nor W are present at position 2 in the training sample, this prediction cannot be contrasted with experimental data. Describing the peptides with the three Z scales, should change this situation.

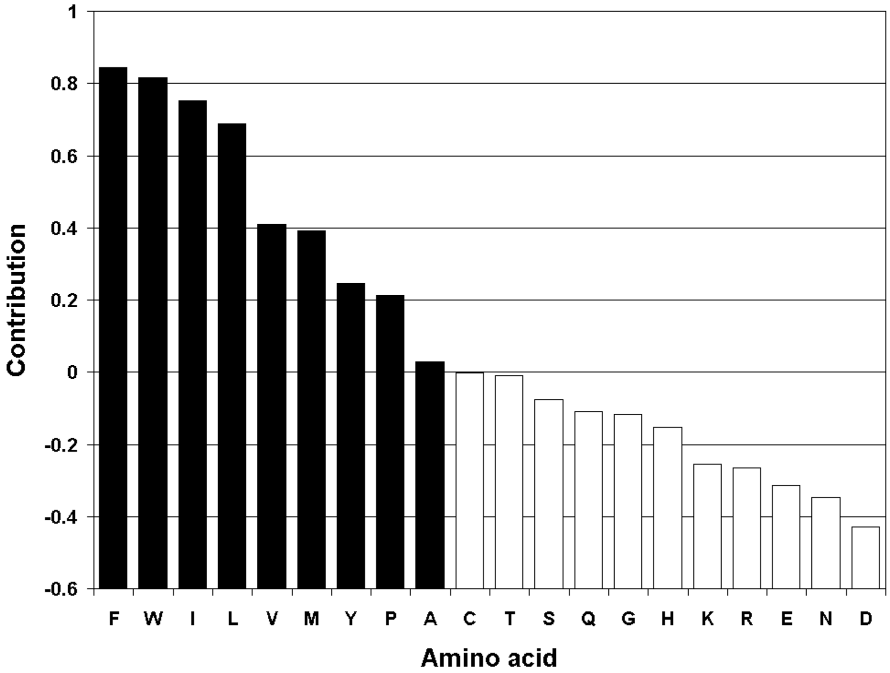

***Figure 2**. Effect of substitution at position 9. Filled bars, positive contributions; empty bars, negative contribution. For the meaning of the symbols see Table 2.*

At position 9 the effect seems to be the same, the amino acids with positive contribution are mainly hydrophobic, probably because almost all the peptides in the sample have hydrophobic residues in this position. As in position 2, L, I and V, the anchor residues (Parker *et al.*, 1992; Kubo *et al.*,1994) are predicted to have positive contribution; this is probably the best prediction of the network since they are present in 70 of the peptides, covering all the range of affinity. Again, F and W have a high predicted contribution although they are not present at this position in the sample, the situation can be explained as in position 2. Other amino acids not present in the sample, but with positive calculated contribution, are M, Y and P. They do not have a contribution as high as F and W but their Z_1 values are also less negative. T, which is tolerated at this position, is predicted to have a very small negative contribution, paralleled by the change in Z_1.

The situation at other positions is not clear, but it is known that they are not as restrictive of binding as 2 and 9 (Kubo *et al.*,1994). It has been mentioned that secondary anchor residues may affect binding and affinity, but the information about them is scarce (Parker *et al.*, 1992). In this situation the predictions of the network are more valuable. At positions 1, 3, 5 and 8, the network predicts better contribution for hydrophobic amino acids, while at positions 4, 6 and 7 the hydrophilic ones have better contributions. Since these positions are not restricted in the sample, the amino acid variation is large. In fact, there are 19 or 20 different amino acids at these positions, but number 8, and, therefore the network, has more information available. With this variability, the predictions are to be considered of value, but again the use of only one aspect of the information on amino acids, limits its application.

Up to this moment, the work has interesting results. We have shown that the principal component scales have enough information to describe the peptide sequence, and calculate the binding affinity, even when used individually. With enough peptides in our sample, to use the complete description, our results will probably improve.

The network predictions are affected by the partial description of the peptides, and the residue variability at each position. These factors limit the applicability of the results, but the calculated relationship is a demonstration that with this approach it is possible to propose solutions to the problem.

REFERENCES

Andrea T.A. & Kalayeh H. (1991). Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. *J. Med. Chem.*, 34:2824-2836.

Gulukota K, Sidney J, Sette A, DeLisi C. (1997). Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.*, 5:1258-1267.

Hagmann, M. (2000). Computers Aid Vaccine Design. *Science*, 290:80-82.

Hellberg S, Sjostrom M, Skagerberg B, Wold S. (1987). Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem.*, 30:1126-1135.

Kast W.M., Brandt R.M., Sidney J., Drijfhout J.W., Kubo R.T., Grey H.M., Melief C.J., Sette A. (1994). Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins. *J. Immunol*., 152:3904-3912.

Kast W.M., Brandt R.M., Sidney J., Drijfhout J.W., Kubo R.T., Grey H.M., Melief C.J., Sette A. (1994a). Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins. *J. Immunol*., 152:3904-3912.

Kawakami Y., Eliyahu S., Jennings C., Sakaguchi K., Kang X., Southwood S., Robbins P.F., Sette A., Appella E., Rosenberg S.A. (1995). Recognition of multiple epitopes in the human melanoma antigen gp100 by tumor-infiltrating T lymphocytes associated with in vivo tumor regression. *J. Immunol*, 154:3961-3968.

Kubo, R.T., Sette A., Grey H.M., Appella E., Sakaguchi K., Zhu N., Arnott D., Sherman N., Shabanowitz J., Michel H., Bodnar W.M. Davis T.A. & Hunt, D.F. (1994). Definition of specific peptide motifs for four major HLA-Aalleles. *J. Immunol.,* 152:3913-3924.

Marchand M., Weynants P., Rankin E., Arienti F., Belli F., Parmiani G., Cascinelli N., Bourlond A., Vanwijck R., Humblet Y. (1995). Tumor regression responses in melanoma patients treated with a peptide encoded by gene MAGE-3. *Int. J. Cáncer*, 63: 883-885.

Parker K.C., Bednarek M.A., Hull L.K., Utz U., Cunningham B., Zweerink H.J., Biddison W.E., Coligan J.E. (1992). Sequence motifs important for peptide binding to the human MHC class I molecule, HLA-A2. *J. Immunol*., 149: 3580-3587.

Rammensee H.G., Friede T., Stevanovic S. (1995). MHC ligands and peptide motifs: 1st listing. *Immunogenetics*, 41, 178-228. recovered from: http://www.syfpeithi.de/bin/MHCServer.dll/Info.htm. On May 25, 2012.

Reche, P.A. & Reinherz, E.L. (2005). PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands. *Nucleic Acids Research*, 33: 138-142. England. recovered on June 13, 2010 from: http://nar.oxfordjournals.org/content/33/suppl_2/W138.full.pdf+html.

Rivoltini L., Kawakami Y., Sakaguchi K., Southwood S., Sette A., Robbins P.F., Marincola F.M., Salgaller M.L., Yannelli J.R., Appella E., Rosenberg S.A. (1995). Induction of tumor-reactive CTL from peripheral blood and tumor-infiltrating lymphocytes of melanoma patients by in vitro stimulation with an immunodominant peptide of the human melanoma antigen MART-1. *J. Immunol*., 154:2257-2265.

Rumelhart, D.E. y McClelland, J.L. (1986), Parallel Distributed Processing Exploration in the Microstructure of Cognition Vol 1: Foundations. MIT Press, Cambridge, USA.

Salazar-Onfray F., Nakazawa T., Chhajlani V., Petersson M., Karre K., Masucci G., Celis E., Sette A., Southwood S., Appella E., Kiessling R. (1997). Synthetic peptides derived from the melanocyte-stimulating hormone receptor MC1R can stimulate HLA-A2-restricted cytotoxic T lymphocytes that recognize naturally processed peptides on human melanoma cells. *Cáncer Res.*, 57:4348-4355.

Sette A. (2000). Tools of the Trade in Vaccine Design. *Science*, 290:2074-2075.

Sette A., Vitiello A., Reherman B., Fowler P., Nayersina R., Kast W.M., Melief C.J., Oseroff C., Yuan L., Ruppert J., Sidney, J., del Guercio M.F., Southwood S., Kubo R.T., Chesnut R.W., Grey H.M., Chisari F.V. (1994). The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol*., 153:5586-5592.

van der Bruggen P., Bastini J., Gajewski T., Coulie P.G., Boel P., De Smet C., Traversari C., Townsend A., Boon T. (1994). A peptide encoded by human gene MAGE-3 and presented by HLA-A2 induces cytolytic T lymphocytes that recognize tumor cells expressing MAGE-3. *Eur J Immunol.*, 24: 3038-3043.