Predicting Binding Peptides with Simultaneous Optimization of Entropy and Evolutionary Distance

Menaka Rajapakse[1,2] and Lin Feng[2]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
[2] School of Computer Engineering, Nanyang Technological University, Block N4, Nanyang Avenue, Singapore 639798
menaka@i2r.a-star.edu.sg, asflin@ntu.edu.sg

Abstract. Identifying antigenic peptides that bind to Major Histocompatibility Complex (MHC) molecules plays a central role in determining T-cell epitopes suitable as vaccine targets. Prediction of the binding ability of antigenic peptides to MHC class II molecules is more complex that for class I. Class II molecules bind to peptides of different lengths and the core region that interacts with the binding site on the class II MHC molecule is located anywhere within the peptide. Obtaining an alignment for these binding sites is an important first step in determining the binding motif of MHC class II alleles. In this paper, we exploit entropy and evolutionary distance of the key binding positions (anchor positions) of an alignment in determining the best possible alignment for a given set of peptide data. Once an optimal alignment is found, a weight matrix representing the binding motif is estimated. The weight matrix designed is subsequently applied to predict MHC binding peptides.

1 Introduction

T cells play a key role as the mediators of immune response against diseases. These cells recognize viral antigens (short peptides) bound to Major Histocompatibility Complex (MHC) molecules through T cell receptors (TCR). Predicting such binding peptides assists in selecting epitopes for use in vaccine design. Prediction of MHC class II peptide binding is more difficult than that of class I [1]. This is due to the open-ended nature of MHC class II peptide binding groove which allows binding to a broader range of peptide lengths (approximately 11 to 22aa) [1,2]. While MHC class I binds to peptides of a narrow range (usually 8-10 aa), a core of nine aa within a peptide is sufficient to bind to MHC molecules of both classes [3]. However, often, the exact location of the binding core (motif) within a peptide longer than nine aa is unknown. Therefore, given a set of experimentally validated MHC class II binders of different length distribution, an accurate alignment of the binding cores must be first obtained before a motif can be determined. According to previous studies carried out on the structural features of MHC class II molecules indicate five binding sites, also known as *anchor positions* at positions 1,4,6,7 and 9 within a 9-mer peptide [4-6].

A peptide binding motif is represented either by a consensus sequence or as a quantitative matrix [7]. A widely used representation of a motif is the quantitative

J.C. Rajapakse, B. Schmidt, and G. Volkert (Eds.): PRIB 2007, LNBI 4774, pp. 349–355, 2007.

matrix. Each element in the matrix depicts a weight corresponding to the interaction between an amino acid and a position in the motif. Derivation of quantitative matrices based on experimentally derived position specific binding profiles is costly and time consuming. Hence, such matrices can not be easily updated as with machine-learning techniques when new data become available [8]. Other popular computational tools available for finding motifs in protein sequences are: MEME [9][23], Gibbs motif sampler [10] and Rankpep [11].

In this study, our aim is to obtain an optimal alignment of the binding cores for MHC class II, I-A^b molecule peptide sequence dataset. This is carried out with the help of an evolutionary algorithm [12] by simultaneously optimizing the relative entropy and the evolutionary distance of possible alignments. The obtained best alignment is then used to derive the quantitative matrix which will subsequently be used to predict binding peptides. Relative entropy, a measurement of uncertainty is often used to analyze sequence features and alignments, to measure sequence conservation. The evolutionary distance is measured using the BLOSUM62 substitution matrix, a matrix suitable for modeling evolutionary problems [13]. As anchor positions are known to influence peptide-MHC binding, a higher weightage is given to such positions during the estimation of evolutionary distance. In order to reduce the sequence redundancy in an alignment, we employed sequence clustering followed by sequence weighting.

2 Materials and Methods

2.1 Peptide Sequence Dataset

Peptide sequences and their binding affinities were obtained from SYFPEITHI [4], MHCPEP [17], AntiJen [20] and EPIMHC [21] databases. An independent test dataset was used to evaluate the predictive ability of the I-A^b mouse model. The extracted dataset consists of 251 unique binders with a length distribution ranging from 9 to 24 amino acid residues and 58 non-binders. Binder set was divided into two sets, training and test set so that there is no overlap between the two datasets. While training set consists of 167 binders, the testing dataset consists of 84 binders and 58 non-binders.

2.2 Peptide Sequence Clustering and Weighting

Sequence clustering and weighting is carried out according to [14]. A set of sequences with sequence identity greater or equal to 62% forms a cluster. Cluster assignment is followed by the sequences weighting. Sequence weighting reduces over-representation of sequences in an alignment. A peptide s of length k-mer in cluster, c is assigned a weight, $w_s = 1/n_c$, when n is the number of sequences in cluster c.

2.3 Pseudo-count Correction

Pseudo-count correction is carried out as given in [15], which uses the prior knowledge of amino acid relationships represented by substitution matrices. For a given column, pseudo-count frequencies, g_{al}, for amino acid a at position l of the alignment are calculated according to the following equation where f_{bl}, q_b, and q_{ab} represent observed frequency of amino acid b in position l, background frequency of amino

acid *b*, and the target frequency implicit in the substitution matrix (the frequency by which amino acid *a* is aligned to amino acid *b*), respectively.

$$g_{al} = \sum_b \frac{f_{bl}}{q_b} q_{ab} = \sum_b f_{bl} q_{a|b} \tag{1}$$

where $q_{a|b}$ is the conditional probability derived from the BLOSUM62 substitution matrix. The effective amino acid frequencies were then determined according to [15] by applying weight on pseudo-count correction as below:

$$g'_{al} = \frac{\alpha . f_{bl} + \beta . g_{al}}{\alpha + \beta} \tag{2}$$

Where α and β represent the effective sequence number and an arbitrary weight on the pseudo-count correction, respectively. Let the number of peptide clusters generated be C, and the value of α = C – 1. An empirically determined suitable setting for β = 10 [15].

2.4 Identification of Binding Core of Peptides

The first step in designing a weight matrix is to obtain an accurate alignment of th binding cores that are distributed within experimentally determined binding peptides of varying length. Therefore, our goal here is to identify the starting position of the binding core in each peptide. Let *S* be a set of *N* peptide sequences, $S = \{s_1, s_2, \ldots s_{i,\ldots} s_N\}$. For a given alignment, let s_{il} denotes the i^{th} peptide whose binding core starts at the l^{th} position within the peptide. Let $\kappa = (k_1 k_2 \ldots k_9)$ represents the selected best nine aa length binding core in a peptide. Once all the starting positions are identified, an alignment is obtained for the binding peptides so that the weight matrix can be derived.

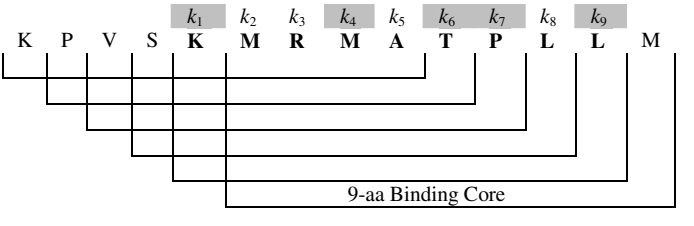

Fig. 1. An illustration of putative 9 aa binding cores within a peptide sequence, s_i. Highlighted positions indicate anchor positions within a putative binding core.

2.5 Generating an Optimal Alignment

We use the evolutionary approach described in [12] to optimize two objective functions associated with relative entropy (*E*) and evolutionary distance (*D*) of all alignments. Each individual in the evolving population represents possible starting positions of binding cores within each experimentally determined binding peptide in the training dataset. An individual is represented by a concatenated string of starting

positions of peptides in an alignment. The bit size for representing each starting position is determined as below. Given a peptide of length r, the number of 9-mer peptides that can be derived from r is p=r-9+1. Hence, the starting positions are located between 0 and p-1 where each peptide is overlapped by a single amino-acid. The bit size, θ, is chosen such that $p < min\ (2^{\theta})$ whereby all 9-mer peptide positions in the peptide are taken care of.

Based on the starting positions embedded in an individual, an alignment is generated for the N peptides. The alignment is then use to estimate D and E for anchor positions as given by the Eq. (3) and Eq. (4) below. The evolutionary distance between two peptide sequences s_m and s_n at anchor positions $\kappa' = (k_1\ k_4\ k_6\ k_7\ k_9)$ of the κ binding core in the alignment is calculated as below, where $B(.)$ is the score estimates from the BLOSUM62 substitution matrix for s_m and s_n. Then D is estimated as:

$$D = \sum_{\substack{m=1 \\ n=m+1}}^{N} \sum_{j=k} W_j * B(s_{m,j}, s_{n,j}) \tag{3}$$

Where W is a weighting factor; W=w for j= κ' and W=1.0 otherwise.

And E is estimated as:

$$E = -\sum_{j=k'} \sum_{a} g_{aj} \log \frac{g'_{aj}}{q_a} \tag{4}$$

where g_{aj} is the frequency of amino acid a occupying at position j in the alignment, g'_{aj} is the frequency of pseudo-count and sequence weight corrected amino acid a at position j, and q_a is the background frequency of amino acid a. A number of different approaches are available for estimating background frequencies, also known as background model or null model: amino acid distribution in the SWISS-PROT database [16], a flat distribution where all amino acid frequencies are equal to 1/20, or an amino acid distribution estimated from a non-binder dataset.

Fitness of an alignment is scored according to Eq.(3) and Eq(4). Best population comprises of individuals that maximize Eq.(3) and minimize Eq.(4) simultaneously. The alignment, which scored the highest D and lowest E is then used to build the weight matrix, M, and subsequently for predicting binders in the testing datasets. Each position of the weight matrix, m_{aj} is calculated according to the equation given below.

$$m_{aj} = g_{aj} \log \frac{g'_{aj}}{q_a} \tag{5}$$

3 Experiments and Results

The experiment in determining the weight matrix representation of I-A^b binding motif is carried out as follows. The values for the parameters β in Eq.(2) and w in Eq.(3) are chosen as 10.0 and 2.0, respectively.

During a single iteration of the evolutionary process, the values of the objective functions, D and E are estimated for the resulting alignments embodied in the individuals. A population of 1000 was evolved for 500 generations with the empirically determined values 0.9 and 0.0004 as the crossover and mutation probability. By using Eq. (5), the weight matrix, M is built with the best alignment, and subsequently used to test the peptides in the testing dataset. A peptide in the testing set is evaluated by scoring all possible 9 aa length binders within the peptide against the weight matrix. Of all the scores, the highest value obtained is assigned as the binding score of the tested peptide. Binding and non-binding status of peptides were determined using a threshold. The performance was measured by estimating Area under Receiver Operating Characteristics (AROC). Let the score estimated for the binding core κ in the peptide s_i be e_i. The binding status, binder (b) or non-binder (nb) is determined according to a threshold, t, as follows:

$$v_i = \begin{Bmatrix} b & if \;\; e_i \geq \mathrm{t} \\ nb & if \;\; e_i < \mathrm{t} \end{Bmatrix}$$

We obtained an ROC curve by evaluating sensitivity and specificity values for various thresholds as illustrated in Figure 2. The final AROC value estimated for the testing dataset is 0.79, a value considered as good prediction accuracy according to [22]. We also compared our results with MEME [23]. The same training dataset was submitted to the on-line web server http://meme.sdsc.edu/meme/meme.html, and the resulting log-odds matrix was used to measure the prediction accuracy. The AROC value estimated for the testing dataset is 0.71.

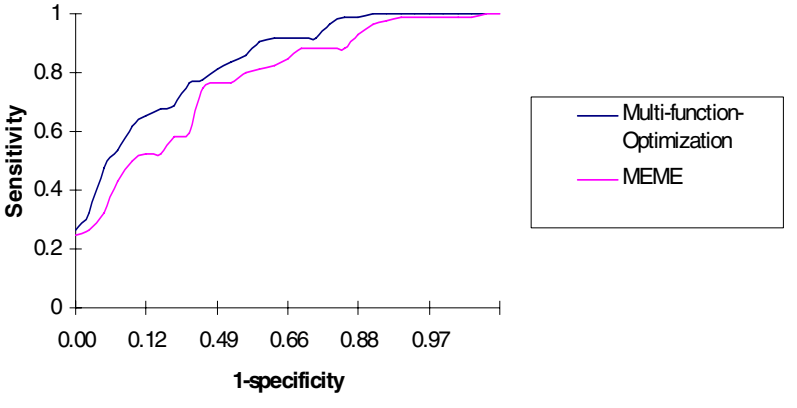

Fig. 2. The ROC plot illustrating the specificity and sensitivity values at different thresholds

4 Discussion and Future Directions

A weight matrix representing motif for MHC class II, I-A^b was derived by simultaneously optimizing entropy and evolutionary distance. The anchor positions of a putative binding core were given higher weightage during the calculation of evolutionary

distance. In order to reduce the sequence redundancy in an alignment, we employed sequence clustering and weighting. The weight matrix developed was subsequently applied to discriminate binders from non-binders. The initial results are promising. Better predictive accuracy can be envisaged by incorporating structural properties as an additional objective function. Currently we are extending our investigations towards evaluating different background models, predictive accuracy of the proposed method on multiple alleles of HLA class I and class II molecules, and determining the applicability of multiple substitution matrices.

References

1. Reche, P.A., Glutting, J.P., Reinherz, E.L.: Prediction of MHC class I binding peptides using profile motifs. Hum. Immunology 63(9), 701–709 (2002)
2. Hammer, J., et al.: Precise prediction of major histocompatibility complex class II – peptide interaction based on peptide side chain scanning. J. Exp. Medicine 180(6), 2353–2358 (1994)
3. Rammensee, H., et al.: MHC ligands and peptide motifs: first listing. Immunogenetics 41(4), 178–228 (1995)
4. Rammensee, H., et al.: SYFPEITHI:database for MHC ligands and peptide motifs. Immunogenetics 50, 213–219 (1999)
5. Stern, L.J., et al.: Crystal structure of the human class II MHC protein HLA-DR1 complexed with an inluenza virus peptide. Nature 368, 215–221 (1994)
6. Dessen, A., et al.: X-ray crystal structure of HLA-DR4 (DRA*0101, DRB1*0401) complexed with a peptide from human collagen II. Immunity 7, 473–481 (1997)
7. Mamitsuka, H.: Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. Proteins 33(4), 460–474 (1998)
8. Nielsen, M., et al.: Improved Prediction of MHC class I and class II epitopes using a novel Gibbs Sampling Approach. Bioinformatics 20(9), 1388–1397 (2004)
9. http://meme.scdc.edu/meme/website/meme.html
10. Neuwald, A.F., et al.: Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Science 4, 1618–1632 (1995)
11. Reche, P.A., et al.: Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. Immunogenetics 56, 405–419 (2004)
12. Deb, K., et al.: A Fast and Elitist Multiobjective Genetic Algorithm. IEEE Trans. on Evolutionary Computation 6, 182–197 (2002)
13. Heinkoff, S., Heinkoff, J.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad.Sci. USA 89, 10915–10919 (1992)
14. Hobohm, U., et al.: Selection of representative protein datasets. Protein Sci. 1, 409–417 (1992)
15. Altschul, S.F., et al.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389–3402 (1997)
16. Bairoch, A., Apweiler, R.: The SWISS-PROT protein sequence database and its supplement TrEMBL. Nucleic Acids Research 28, 45–48 (2000)
17. Brusic, V., Rudy, G., Harrison, L.C.: MHCPEP, a database of MHC-binding peptides: update 1997. Nucleic Acids Res. 26, 368–371 (1998)
18. Bhasin, M., Singh, H., Raghava, G.P.S.: MHCBN, a comprehensive database of MHC binding and non-binding peptides. Bioinformatics 19, 665–666 (2003)

19. Nielsen, M., Lundegaard, C., Worning, P., Hvid, C.S., Lamberth, K., Buus, S., Brunak, S., Lund, O.: Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. Bioinformatics 20(9), 1388–1397 (2004)
20. Blythe, M.J., Doytchinova, I.A., Flower, D.R.: JenPep: a database of quantitative functional peptide data for immunology. Bioinformatics 18, 434–439 (2002)
21. Reche, P.A., Zhang, H., Glutting, J.P., Reinherz, E.L.: EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics 21, 2140–2141 (2005)
22. Swets, J.A.: Measuring the accuracy of diagnostic systems. Science 240, 1285–1293 (1988)
23. Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28–36. AAAI Press, Menlo Park, California (1994)