Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Predicting Peptide Binders of Flexible Lengths with Genetic Annealing Algorithm

Menaka Rajapakse[1,2]
[1]Institute for Infocomm Research
21, Heng Mui Keng Terrace
Singapore 119613
menaka@i2r.a-star.edu.sg

Lin Feng[2]
[2]School of Computer Engineering
Nanyang Technological University, BLK N4
Nanyang Ave. Singapore 639798
asflin@ntu.edu.sg

***Abstract*—Prediction of peptides that bind to Major Histocompatibility Complex class II (MHC-II) molecules is vital for drug discovery and vaccine development. Prediction of peptides binding to MHC-II molecules is complicated because of the broad range of their lengths. Peptides bind to the molecules at an ungapped motif present at the binding site. Obtaining an alignment of binding sites of binding proteins facilitates determining of the binding motif. However, multiple sequence alignment often fails on peptides. In this paper, we propose a Genetic Annealing Algorithm (GAA) to identify an alignment for binding peptides that can subsequently be used to predict binding peptides. Our approach is demonstrated with a dataset having difficulty in finding a consensus motif through experimental means and using existing motif detection methods. GAA based approach outperformed Gibbs motif sampler and RANKPEP approaches in predicting peptides binding to MHC II molecules.**

Keywords- Genetic algorithm; MHC molecules; motif; peptide binding

I. INTRODUCTION

Major histocompatibility complex (MHC) molecules play a key role in initiating an immune response. They bind to and expose an antigen (or short peptides) so that they are recognized by T cell receptors (TCR) which then identify the foreign peptide and trigger an immune response against the infected cell or foreign agent. MHC molecules make multiple contacts with the side-chains of a binding peptide, which determines the specificity of binding and define the binding motifs [1]. Prediction of MHC class II peptide binding is more difficult than that of class I [2]. This is due to fewer restrictions being imposed on the type of side chains by MHC class II molecules and the ability of MHC class II molecules to bind to peptides longer than 9 amino acids (aa) (approximately 11 to 22aa) [1, 3]. A core of 9 aa within a peptide is sufficient to bind to a MHC class II molecule [4]. However, the exact location of the binding core (motif) within a peptide is unknown.

A peptide binding motif is represented either by a consensus sequence or as a weight matrix [5]. The presence of a motif binding to a particular peptide can be determined experimentally from a large pool of known binding peptides [4, 6]. However, such experimental methods are costly, time consuming, and cumbersome. Amino acids at specific positions that contribute significantly to the binding are referred to as *primary anchor residues* and the corresponding sites as *anchor positions*. Anchor positions are occupied by *preferred residues* that are tolerated with varying strengths at binding sites but alone contribute little to the binding of the peptide to the molecule. Earlier studies, using more comprehensive information, found complex matrix models that elaborate the exact nature of the binding strength [7, 8]. These matrix models offering position specific binding strength of each residue within the binding core are known as Position Specific Scoring Matrices (PSSM).

Advanced classifiers such as artificial neural networks (ANN) [9-14], hidden Markov models (HMM) [5, 15], support vector machines (SVM) [16-18] and their hybrids [19] have been used to discriminate binding peptides (binders) and non-binding peptides (nonbinders). However, these classifiers require the input training peptides be of equal lengths. Given a set of peptides of different lengths with known binding affinities, the location of the binding core within each peptide must be first identified and then extracted before classification. Classical multiple sequence alignment techniques often fail to detect the binding cores due to weak instances of binding motifs.

Recently, iterative learning methods [20-23], stochastic approaches such as multiple EM elicitation (MEME) [24, 25], Gibbs motif sampler [26-29], profile motifs (RANKPEP) [2, 30], etc., and evolutionary algorithms (EA) [31] have been used to try and uncover motifs in datasets of peptides with varying length. An iterative step-wise discriminant analysis (SDA) has been used to derive a quantitative matrix for MHC class II peptide data of variable length [20, 21]. Given two mutually exclusive sets, SDA is able to build a Bayesian discriminant function that is used to implement a binary classifier by generating binders based on a predefined anchor motif. The results are refined according to a score calculated based on the presence of anchor positions specified in the motif. This approach is more suitable when binding and non-binding sequences are significantly distinct. A linear programming model has been utilized as the learning model for the binary classification of binders and nonbinders in [22]. This supervised model generates a predictor while learning features of the negative samples (nonbinders) and iteratively filtering them out of an unlabeled dataset consisting of possible binders and nonbinders. The reported results are comparable or better

over the Gibbs approach on different datasets. An ant colony system (ACS) has been used to search for an optimal local alignment for a set of peptides of variable length [23]. The performance of the ACS strategy has rendered comparable or better results than the Gibbs sampler for a number of different datasets. A set of profile motifs has been used in RANKPEP to predict peptide binding to a number of MHC class I and class II molecules [2, 30]. MEME [25] and Gibbs sampler [26, 28] are two widely used statistical approaches for motif detection in unaligned peptide sequences. Gibbs sampler performs a random walk through the space of multiple alignments and is less prone to get trapped in a local minimum compared to greedy algorithms such as MEME. The main drawbacks associated with Gibbs sampler include different results at each run, frequent false positives, and attraction to local maxima.

To date, there is no one optimal model or algorithm for predicting the peptides that bind to all MHC class I or class II molecules. Therefore, different algorithms that perform well on previously unseen data are needed. We propose the use of EA to align a set of experimentally determined binding peptides at their binding cores and subsequently derive the binding motif. The accuracy of an EA-based technique mainly depends on the fitness function defining the proximity to the optimal solution. We explore Genetic Annealing Algorithm (GAA) for predicting MHC-II peptide binding. The GAA explores the solution space to identify a motif that can best explain the peptide binding for a given dataset.

We demonstrate our method on experimental datasets of peptide binding to I-A^{g7} molecule obtained from literature (Table 1). I-A^{g7} is the MHC class II molecule of the NOD mouse, critical for the development of insulin-dependent diabetes mellitus (IDDM) and other autoimmune disorders [32-37] . The knowledge of peptide binding to I-A^{g7} is important in understanding the molecular basis of the development of IDDM in NOD mice. Finding motifs in peptide binding to I-A^{g7} is a non-trivial problem [38, 39]. Despite numerous attempts, no consensus has been reached in defining motifs that describe the binding rules to A^{g7} molecule [32-42]. Experiments have demonstrated that I-A^{g7} binding peptides are 9-30 aa long [41]. However, computational analyses on multiple datasets show that each experimentally determined motif only explains a subset of the rules describing the optimal motif.

II. Materials And Methods

A. Genetic Algorithms

Genetic algorithms (GA) are based on the principles of biological evolution and have often been successful in solving complex search and optimization problems. GAs find a wide spectrum of applications in bioinformatics. The majority GA applications has been concerned with motif discovery, an example of which is TFBS detection [43-47]. A few researchers have used GAs for peptide binding predictions from protein sequences [31]. For more details on EAs, GA, and their applications in bioinformatics, readers are referred to [48-51].

The basic steps of an EA implementation are: (1) the representation of input variables as individuals or chromosomes (binary or real valued) in a population; (2) the formulation of fitness (objective function) to evaluate individuals; (3) the formation of a new population by genetic operations (reproduction, crossover, and mutation) on the present population; and (4) determine whether the population has achieved the optimal fitness. The algorithm starts with an initial population of individuals and evolves in an iterative manner. In a single iteration, each individual is evaluated by estimating its fitness. New populations (offspring) are produced from highly fit individuals (parents), chosen according to a selection criterion, which then undergo genetic operations. Each offspring is thereafter paired and compared with its parents .The highly fit individuals are retained while the less fit individuals are discarded.

B. Genetic Annealing Algorithm

The GAA incorporates simulated annealing [51] [52, 53] into the crossover process of the GA [54], thereby combining the advantages of the both algorithms [55, 56]. The strategy behind simulated annealing is to allow moves resulting in solutions worse than the current solution in order to avoid local minima. In GAA, offspring produced in crossover between two parents are evaluated for their fitness. Highly fit offspring replace the parents in the next round of crossovers. Less fit offspring than their parents only survive with a selection probability characterized by Boltzmann distribution:

$$P(f) = \frac{1}{Z(T)} \exp\left(-\frac{|\Delta f|}{T}\right) \quad (1)$$

where Δf is the difference of the fitness between an offspring and a parent in the population, T is the temperature of the current population and $Z(T)$ is the normalizing function. After a new population is formed the temperature is lowered by a small fraction (γ) and the process continues until the termination criterion is met. The pseudo code for the GAA is as follows [56]:

An initial temperature, T_0, is defined.
$T = T_0$
Begin: GAA
  recruit
  Repeat
    select
    crossover with annealing
    $T = \gamma \cdot T$
  Until { good solutions found }
End

In the first step, a predefined number of individuals are recruited. Next, the selection process is carried out based on the fitness of the each individual in the population. For selection, the binary tournament selection scheme is used [57]. During which, all the individuals are paired up and their fitness values are compared; the fittest individual of a pair is retained and the other is discarded. The selection process is followed by the

annealing crossover operation which, in turn is followed by the mutation operation. During the annealing crossover process, a highly fit individual (say parent-1) is selected and is allowed to mate with a partner (say parent-2) selected randomly from the population to produce two offspring. The fitness of the two new offspring (say offspring-1 and offspring-2) is evaluated. If the fitness of offspring-1 is better than the fitness of parent-1, then parent-1 is replaced by offspring-1. Otherwise, Boltzmann probability is computed and compared with a normalized random number. The less fit offspring is accepted only if the random number is less than the calculated probability. This process repeats for a predefined number of times to ensure that the best possible solution for the starting pair of parents (parent1 and parent2) is achieved. This is analogous to obtaining thermal equilibrium in simulated annealing at a given temperature [56]. After the entire process is completed the fittest individual in the population is selected, the temperature is reduced, and the entire crossover process is repeated with another partner selected at random.

C. Predicting Peptide Binding to MHC-II I-A^{g7}

Here, we attempt to find an optimal motif describing peptide – MHC-II (I-A^{g7}) molecular binding from experimental binding data that is already available. There are several factors that impede the derivation of such a consensus motif. The first is the strong resemblance among the peptides isolated in a single experiment and the second is the diversity among different datasets. A motif derived from a dataset which lacks diversity indicates a bias towards the dataset used to derive the motif. Such motifs are difficult to generalize for previously unseen datasets. Our aim is to find a consensus motif for I-A^{g7} binding data by using an evolutionary approach that can alleviate the influences that arise from biased datasets.

D. Datasets

Seven I-A^{g7} datasets were extracted from literature [34-37, 39, 58-60] and from Brusic, V.(unpublished data). The numbers of binders and nonbinders in each dataset are given in Table 1. The datasets consist of short peptides ranging from 9-30 amino acids in length. Their binding affinities have been experimentally determined by independent studies and classified as binders or nonbinders based on an inhibitory concentration (IC_{50}) according to the following scheme [35]: good binder (IC_{50}=100nM); weak binder (IC_{50}=2000nM); nonbinder (IC_{50}=50000nM). The datasets in [34-37, 58-60] were combined into a single training dataset and preprocessed by removing duplicates and by discarding: (1) long binder if a binder is a substring of another binder and (2) the substring if a nonbinder is substring of another nonbinder. Let the preprocessed and combined dataset be here onwards referred to as *training* dataset and denoted by $D = \{(x_i, v_i): i = 1, 2,\ldots,d\}$ where d is the number of total peptides and x_i is the i^{th} peptide sequence with the label $v_i \in \{b, nb\}$ indicating whether the sequence x_i is a binder (b), or a nonbinder (nb). The number of peptides in the training set d = 438. Out of which, 304 were binders and 134 were nonbinders.

An independent dataset, Stratmann [39], consists of a diverse set of I-A^{g7} binding peptides with their binding affinities was used as the *testing* dataset. The number of binders and nonbinders in this dataset are 112 and 3, respectively. Due to the fewer number of nonbinders in the testing dataset, we augmented the number of nonbinders to 1000 with randomly generated nonbinders. The generation of random nonbinders involves adding correct proportions of amino acids to each peptide so that the generated peptides mimic real protein peptides [61]. Of randomly generated peptides, approximately five percent is presumed to be binders [58]. The error arising from the five percent of possible binders in the randomly generated nonbinder set was taken into account when calculating the prediction accuracy.

E. Binding Score Matrix

A k-mer motif of amino acids is characterized by a positional binding score matrix (BSM), $Q = \{ q_{ia} \}_{kx20}$ where q_{ia} denotes the binding strength of the site i when it is occupied by amino acid a. The binding score of a motif is computed by adding the binding scores assigned for each amino acid at the respective positions. The binding score indicates the likelihood of the motif binding to the molecule. The binding score s_i of the sequence x_i is given by the maximum value of binding scores calculated for all the k–mer subsequences in x_i:

$$s_i = \max_{l \in \{1,\cdots,n-k+1\}} s_{il} \tag{3}$$

where s_{il} denotes the binding score of the subsequence beginning at location l of the sequence:

$$s_{il} = \sum_{l'=0,1,\cdots,k-1} q_{(l+l')x_{i(l+l')}} \tag{4}$$

and assuming one motif instance per sequence, the location of the motif is given by

$$l^* = \arg\max_{l \in \{1,2,\cdots,n-k+1\}} \{s_{il}\} \tag{5}$$

That is, if x_{il} denotes the k-mer subsequence starting at site l of the sequence x_i, then the most likely motif instance, say m_i, is given by $m_i = x_{il^*}$.

F. Obtaining an Alignment of Binding Cores with GAA

This experiment is aimed at identifying an alignment of binding cores when the training dataset consists of peptides of varying lengths. The alignment is then used to discover the consensus motif in the form of a BSM. The positions of the binding cores within the peptides are unknown. The elements of the BSM, say Q, are represented with linear binary strings, by rearranging as a 20k-tuple (q_{ia}, : i=1,….k; $a \in \Omega$) where a is an amino acid in the amino acid alphabet, Ω. Each element in the k-tuple is converted to a binary representation with a binary word of size θ so that $q_{ia} \in [0, 2^{\theta}-1]$. The k-mer motif is therefore represented by a 20$k\theta$ long binary string. Let the

binary representation of Q at the t^{th} iteration of a GA evolution is denoted by $q(t) = \{q_1(t), q_2(t), \ldots\ldots q_N(t)\}$ where N is the size of the population or the number of individuals.

The fitness function is designed to arrive at an optimal consensus of the motifs, using the information of binding peptides, provided in the training dataset. A solution is evaluated on its ability to maximize the accuracies in identifying true binders (TP) and true nonbinders (TN) as well as to widen the gap between scores for binders and nonbinders. This is achieved by a fitness function that minimizes a linear combination of the sum of false positives (FP) and false negatives (FN) as well as the ratio between the average cumulative scores of nonbinders and binders.

The fitness function f is given by

$$f = \mathrm{FN} + \kappa_1 \mathrm{FP} + \kappa_2 \frac{\mathrm{N_b} \sum_{i=1}^{d} s(m_i)\delta(v_i = nb)}{\mathrm{N_{nb}} \sum_{i=1}^{d} s(m_i)\delta(v_i = b)} \qquad (6)$$

where $s(m_i)$ denotes the score computed for the most likely motif instance of sequence x_i of the training dataset and the Kronecker δ is equal to one when the arguments are satisfied, otherwise it becomes zero; N_b and N_{nb} are the total counts of binders and nonbinders in the dataset. The constant κ_1 ($>N_b/N_{nb}$ for $N_b > N_{nb}$) was empirically determined to minimize the number of false positives with respect to the nonbinders. The constant κ_2 acts as a normalizing parameter between the sum of FN and FP and the ratio between the average cumulative scores.

For this experiment, k=9, θ=7. The GAA was run with a population size of N=1000 and the number of generations set to 40. In each generation, the temperature was adjusted and the population was subjected to 20 iterations. One point crossover operation was performed with mutation probability, p_m=0.035. The temperature was initially set to T_0=0.2 and at each generation was reduced by a proportion γ = 0.9. The parameters were set as follows, κ_1 = 5.5 and κ_2 = 1.0.

III. Results And Discussion

We demonstrate the application of GAA determine a consensus motif. Seven experimental datasets of binding peptides to I-A^{g7} molecules obtained from literature were used for training [34-37, 58-60] and an independent dataset [39] was used as testing dataset to compute prediction accuracy. Receiver operating characteristics (ROC) is used as the measure of prediction accuracies and the overall quality of the prediction is measured using the Area under Receiver Operating Characteristics (AUC) [62]. With GAA, the consensus motif, with the best fitness, was determined by the highest AUC on the training dataset. Of all the attempts, the solution with the highest AUC was chosen as the consensus motif.

Table 1 shows the datasets extracted from literature, which were used in the training. An independent dataset comprises of a diverse set of peptides, the Stratmann dataset, was used as the testing dataset. Let the score corresponding to the motif in the peptide x_i be s_i. Then whether the peptide is a binder or a nonbinder is determined according to a threshold, t, as follows:

$$\hat{v}_i = \begin{Bmatrix} b & if\ \ s_i \geq \mathrm{t} \\ nb & if\ \ s_i < \mathrm{t} \end{Bmatrix} \qquad (7)$$

We obtained ROC curve by evaluating sensitivity and specificity values at various thresholds. The performance of GAA on training and testing datasets are given in Table 2. The performances are compared with the earlier motif prediction approaches RANKPEP [30] and the Gibbs sampler [26]. As seen, Gibbs Sampler and RANKPEP exhibit poorer performance than GAA on the training dataset. This may be due to the fact that, unlike GAA, Gibbs approach searches for a motif by using only positive data (binders) and therefore do not learn the characteristics of nonbinders. AUC plots are shown in Fig. 1.

TABLE I. I-A^{G7} Peptide Datasets

Dataset	Nonbinders	Binders	Reference
Reizis	21	33	[34]
Harrison	19	157	[35]
Gregori	31	109	[37]
Latek	8	37	[36]
Corper	35	13	[58]
MHCPEP	-	176	[59]
Yu	16	10	[60]
Brusic	37	-	[unpublished]

TABLE II. Performance of I-Ag7 motif derived by GAA, Gibbs Sampler, and RANKPEP

Motif	Performance measured by AUC	
	Training set	Testing set
GAA	0.82	0.85
Gibbs Sampler	0.60	0.82
RANKPEP	0.59	0.75

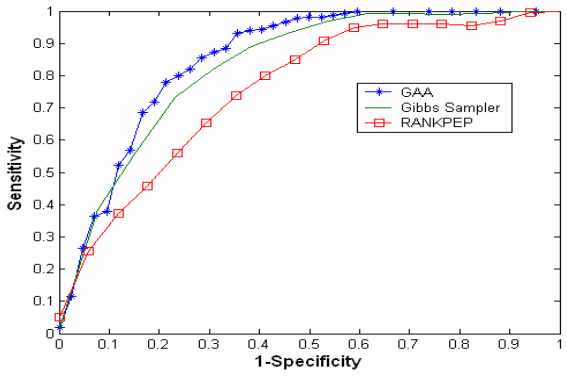

Figure 1. Comparison of performance of GAA with Gibbs Sampler and RANKPEP on the independent dataset for determining I-A^{g7} best motif.

IV. CONCLUSIONS

We proposed a GAA-based approach to identify an alignment of binding cores, which subsequently renders a motif for predicting peptides that bind to MHC class II molecules. Our approach facilitates self discovering a motif, that is, when no information of motifs is available. The GAA approach outperformed earlier approaches to motif detection.

GAA-derived motif outperformed existing motif finding algorithms such as Gibbs sampler and RANKPEP. EAs have the advantages over EM-based algorithms in generating biologically meaningful results by performing a global search [64]. Though a global search by an EA does not guarantee an optimal solution, the likelihood of finding an optimal solution is higher than a local or greedy search. Moreover, the EAs have the advantage of learning the characteristics of both binders and non- binders from the training data while EM or Gibbs algorithms use only the binder dataset in the training dataset. This was reflected in the AUC values calculated for the training dataset. It is important to note that the performance of the EM based Gibbs sampler on the test dataset was comparable with the proposed EA. Though Gibbs sampler is faster, EA gave better performance reaching the global optimum. However, there are number of parameters that need to be tuned in order to obtain the optimal performance with the Gibbs sampler. In the case of GAA, a few parameters must be empirically determined and tuned for optimal performance. Basic rules for selecting parameters for GAA were given. Our future investigations are aimed at defining a set of suitable ranges for these input parameters.

Computational predictions of peptides that bind to MHC class II molecules of the immune system are vital for designing vaccines and discovering drugs for diseases including cancer, infectious diseases, and autoimmunity. Though computationally predicted binders do subsequently need to be validated by wet lab experiments, high costs involved in the initial screening process and clinical testing can be significantly reduced by incorporating computational predictions as a preliminary step.

ACKNOWLEDGMENT

The authors would like to thank Dr Liu Ning, Dr Bertil Schmidt and Dr Vladimir Brusic for their helpful discussions and assistance in earlier part of this work.

REFERENCES

[1] L. J. Stern and D. C. Wiley, "Antigenic peptide binding by class I and class II histocompatibility proteins," *Behring Inst Mitt*, pp. 1-10, 1994.

[2] P. A. Reche, J. P. Glutting, and E. L. Reinherz, "Prediction of MHC class I binding peptides using profile motifs," *Hum Immunol*, vol. 63, pp. 701-9, 2002.

[3] J. Hammer, E. Bono, F. Gallazzi, C. Belunis, Z. Nagy, and F. Sinigaglia, "Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning," *J Exp Med*, vol. 180, pp. 2353-8, 1994.

[4] [4] H. G. Rammensee, T. Friede, and S. Stevanoviic, "MHC ligands and peptide motifs: first listing," *Immunogenetics*, vol. 41, pp. 178-228, 1995.

[5] H. Mamitsuka, "Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models," *Proteins*, vol. 33, pp. 460-74, 1998.

[6] K. Falk, O. Rotzschke, S. Stevanovic, G. Jung, and H. G. Rammensee, "Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules," *Nature*, vol. 351, pp. 290-6, 1991.

[7] J. Ruppert, J. Sidney, E. Celis, R. T. Kubo, H. M. Grey, and A. Sette, "Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules," *Cell*, vol. 74, pp. 929-37, 1993.

[8] M. Bouvier and D. C. Wiley, "Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules," *Science*, vol. 265, pp. 398-402, 1994.

[9] L. Bisset, Fierz, W, "Using a neural network to identify potential HLA-DR1 binding sites within proteins," *J. Mol. Recognition*, vol. 6, pp. 41-48, 1994.

[10] V. Brusic, Rudy, G, Harrison, LC, "Prediction of MHC binding peptides using artificial neural networks," in *In Complex Systems: Mechanism of Adaptation*, R. Stonier, Yu, XS, Ed. Amsterdam: IOS Press, 1994, pp. 253-260.

[11] H. P. Adams and J. A. Koziol, "Prediction of binding to MHC class I molecules," *J Immunol Methods*, vol. 185, pp. 181-90, 1995.

[12] K. Gulukota, Sidney, J., Sette, A., DeLisi, C. , "Two complementary methods for predicting peptide binding major histocompatibility complex molecules," *J. Mol. Biol*, vol. 267, pp. 1258-1267, 1997.

[13] F. D. R. Doytchinova I.A, "Towards the insilico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction," *Bioinformatics*, vol. 19, pp. 2263-70, 2003.

[14] F. R. Burden and D. A. Winkler, "Predictive Bayesian neural network models of MHC class II peptide binding," *J Mol Graph Model*, vol. 23, pp. 481-9, 2005.

[15] H. Noguchi, R. Kato, T. Hanai, Y. Matsubara, H. Honda, V. Brusic, and T. Kobayashi, "Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules," *J Biosci Bioeng*, vol. 94, pp. 264-70, 2002.

[16] P. Donnes and A. Elofsson, "Prediction of MHC class I binding peptides, using SVMHC," *BMC Bioinformatics*, vol. 3, pp. 25, 2002.

[17] Y. Zhao, C. Pinilla, D. Valmori, R. Martin, and R. Simon, "Application of support vector machines for T-cell epitopes prediction," *Bioinformatics*, vol. 19, pp. 1978-84, 2003.

[18] M. Bhasin and G. P. Raghava, "SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence," *Bioinformatics*, vol. 20, pp. 421-3, 2004.

[19] H. H. Takahashi H., "Prediction of peptide binding to major histocompatibility complex class II molecules through use of bossted fuzzy classifier with SWEEP operator method," *Bioscience and Bioengineering*, vol. 101, pp. 137-41, 2006.

[20] R. R. Mallios, "Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm," *Bioinformatics*, vol. 15, pp. 432-9, 1999.

[21] R. R. Mallios, "Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm," *Bioinformatics*, vol. 17, pp. 942-8, 2001.

[22] N. Murugan and Y. Dai, "Prediction of MHC class II binding peptides based on an iterative learning model," *Immunome Res*, vol. 1, pp. 10, 2005.

[23] O. Karpenko, J. Shi, and Y. Dai, "Prediction of MHC class II binders using the ant colony search strategy," *Artif Intell Medicine*, vol. 35, pp. 47-56, 2005 Sep-Oct.

[24] T. L. Bailey, Elkan, C., "Unsupervised learning of multiple motifs in biopolymers using expectation maximization," *Machine Learning*, vol. 21, pp. 51-80, 1995.

[25] T. L. Bailey and E. Charles, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," presented at Second International Conference on Intelligent Systems for Molecular Biology, 1994.

[26] A. F. Neuwald, J. S. Liu, and C. E. Lawrence, "Gibbs motif sampling: detection of bacterial outer membrane protein repeats," *Protein Sci*, vol. 4, pp. 1618-32, 1995.

[27] G. Thijs, K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau, "A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes," *J Comput Biol*, vol. 9, pp. 447-64, 2002.

[28] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment," *Science*, vol. 262, pp. 208-14, 1993.

[29] M. Nielsen, C. Lundegaard, P. Worning, C. S. Hvid, K. Lamberth, S. Buus, S. Brunak, and O. Lund, "Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach," *Bioinformatics*, vol. 20, pp. 1388-97, 2004.

[30] P. A. Reche, J. P. Glutting, H. Zhang, and E. L. Reinherz, "Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles," *Immunogenetics*, vol. 56, pp. 405-19, 2004.

[31] V. Brusic, C. Schonbach, M. Takiguchi, V. Ciesielski, and L. C. Harrison, "Application of genetic search in derivation of matrix models of peptide binding to MHC molecules," *Proc Int Conf Intell Syst Mol Biol*, vol. 5, pp. 75-83, 1997.

[32] E. P. Reich, H. von Grafenstein, A. Barlow, K. E. Swenson, K. Williams, and C. A. Janeway, Jr., "Self peptides isolated from MHC glycoproteins of non-obese diabetic mice," *J Immunol*, vol. 152, pp. 2279-88, 1994.

[33] S. Amor, J. K. O'Neill, M. M. Morris, R. M. Smith, D. C. Wraith, N. Groome, P. J. Travers, and D. Baker, "Encephalitogenic epitopes of myelin basic protein, proteolipid protein, myelin oligodendrocyte glycoprotein for experimental allergic encephalomyelitis induction in Biozzi ABH (H-2Ag7) mice share an amino acid motif," *J Immunol*, vol. 156, pp. 3000-8, 1996.

[34] B. Reizis, M. Eisenstein, J. Bockova, S. Konen-Waisman, F. Mor, D. Elias, and I. R. Cohen, "Molecular characterization of the diabetes-associated mouse MHC class II protein, I-Ag7," *Int Immunol*, vol. 9, pp. 43-51, 1997.

[35] L. C. Harrison, M. C. Honeyman, S. Trembleau, S. Gregori, F. Gallazzi, P. Augstein, V. Brusic, J. Hammer, and L. Adorini, "A peptide-binding motif for I-A(g7), the class II major histocompatibility complex (MHC) molecule of NOD and Biozzi AB/H mice," *J Exp Med*, vol. 185, pp. 1013-21, 1997.

[36] R. R. Latek, A. Suri, S. J. Petzold, C. A. Nelson, O. Kanagawa, E. R. Unanue, and D. H. Fremont, "Structural basis of peptide binding and presentation by the type I diabetes-associated MHC class II molecule of NOD mice," *Immunity*, vol. 12, pp. 699-710, 2000.

[37] S. Gregori, E. Bono, F. Gallazzi, J. Hammer, L. C. Harrison, and L. Adorini, "The motif for peptide binding to the insulin-dependent diabetes mellitus-associated class II MHC molecule I-Ag7 validated by phage display library," *Int Immunol*, vol. 12, pp. 493-503, 2000.

[38] E. Carrasco-Marin, O. Kanagawa, and E. R. Unanue, "The lack of consensus for I-A(g7)-peptide binding motifs: is there a requirement for anchor amino acid side chains?," *Proc Natl Acad Sci U S A*, vol. 96, pp. 8621-6, 1999.

[39] T. Stratmann, Apostolopoulos,V, Mallet-Designe V, Corper AL, Scott CA, Wilson IA, Kang AS, Teyton L, "The I-Ag7 MHC class II molecule linked to murine diabetes is a promiscuous peptide binder," *J. Immunology*, vol. 165, pp. 3214-25, 2000.

[40] E. Carrasco-Marin, J. Shimizu, O. Kanagawa, and E. R. Unanue, "The class II MHC I-Ag7 molecules from non-obese diabetic mice are poor peptide binders," *J Immunol*, vol. 156, pp. 450-8, 1996.

[41] H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanovic, "SYFPEITHI: database for MHC ligands and peptide motifs," *Immunogenetics*, vol. 50, pp. 213-9, 1999.

[42] A. Suri, I. Vidavsky, K. van der Drift, O. Kanagawa, M. L. Gross, and E. R. Unanue, "In APCs, the autologous peptides selected by the diabetogenic I-Ag7 molecule are unique and determined by the amino acid changes in the P9 pocket," *J Immunol*, vol. 168, pp. 1235-43, 2002.

[43] R. G. Beiko and R. L. Charlebois, "GANN: genetic algorithm neural networks for the detection of conserved combinations of features in DNA," *BMC Bioinformatics*, vol. 6, pp. 36, 2005.

[44] G. B. Fogel, D. G. Weekes, G. Varga, E. R. Dow, H. B. Harlow, J. E. Onyia, and C. Su, "Discovery of sequence motifs related to coexpression of genes using evolutionary computation," *Nucleic Acids Res*, vol. 32, pp. 3826-35, 2004.

[45] F. Liu, Tsai, J., Chen R., Chen, S., Shih, S., "FMGA: Finding Motifs by Genetic Algorithm," presented at IEEE BIBE, 2004.

[46] N. Lo, Changchien, S., Chang, Y., Lu, T, "Human promoter prediction based on sorted consensus sequence patterns by genetic algorithms," presented at Intl congr. on Biological and Medical Engineering, 2002.

[47] D. Corne, Meade, A., Sibly, R, "Evolving Core Promoter Signal Motifs," presented at IEEE Congress on Evolutionary Computation, 2001.

[48] J. A. Foster, "Evolutionary computation," *Nat Rev Genet*, vol. 2, pp. 428-36, 2001.

[49] M. Mitchell, *An Introduction to Genetic Algorithms*: MIT press, 1999.

[50] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*: Wiley publishers, 2001.

[51] G. Fogel, Corne, D., *Evolutionary Computation in Bioinformatics*: Morgan Kaufman publishers, 2003.

[52] S. Kirkpatrick, Gelatt, C. D., and Vechi, M. P, "Optimization by Simulated Annealing," *Science*, vol. 220, pp. 671-680, May 1983.

[53] G. B. Sorkin, "Efficient Simulated Annealing on Fractial Energy Landscapes," *Algorithmica*, vol. 6, pp. 367-418, 1991.

[54] J. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor,MI: University of Michigan Press, 1975.

[55] S. Lim, Tan, B, "Performance of the Genetic Annealing Algorithm in DFM Synthesis of Dynamic Musical Sound Samples," *J. Audio Eng. Soc*, vol. 57(5), 1999.

[56] B. Tan, Lim, S., "Automated Parameter Optimization for Double Frequency Modulation Synthesis Using the Genetic Annealing Algorithm," *J. Audio Eng. Soc*, vol. 44, pp. 3-15, 1996.

[57] D. E. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms," in *Foundations of Genetic Algorithms,* G. Rawlins, Ed.: Morgan Kaufmann, 1991.

[58] A. Corper, Stratmann,T,Apostolopoulos,V,Scott, CA, Garcia,KC, Kang,AS,Wilson,IA,Teyton,L, "A Structural Framework for Deciphering the Link Between I-Ag7 and Autoimmune Diabetes " *Science 21*, vol. 288, pp. 505 - 511, April 2000.

[59] V. Brusic, G. Rudy, and L. C. Harrison, "MHCPEP, a database of MHC-binding peptides: update 1997," *Nucleic Acids Res*, vol. 26, pp. 368-71, 1998.

[60] B. Yu, L. Gauthier, D. H. Hausmann, and K. W. Wucherpfennig, "Binding of conserved islet peptides by human and murine MHC class II molecules associated with susceptibility to type I diabetes," *Eur J Immunol*, vol. 30, pp. 2497-506, 2000.

[61] I. Pe'er, C. E. Felder, O. Man, I. Silman, J. L. Sussman, and J. S. Beckmann, "Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla," *Proteins*, vol. 54, pp. 20-40, 2004.

[62] A. Webb, *Statistical Pattern Recognition*, 2nd ed: John Wiley & Sons, 2002.

[63] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, pp. 1285-93, 1988.

[64] M. A. Lones, Tyrrell, A.M., "The Evolutionary Computation Approach to Motif Discovery in Biological Sequences," presented at Genetic and Evolutionary Computation, 2005.