Rule Induction for Prediction of MHC II-Binding Peptides

An Zeng[1], Dan Pan[2], Yong-quan Yu[1], and Bi Zeng[1]

[1] Faculty of Computer, Guangdong University of Technology, Guangzhou, 510006 Guangdong, China
csanzeng@gmail.com, yyq@gdut.edu.cn, z9215@163.com
[2] China Mobile Group Guangdong Co., Ltd., Guangzhou, 510100 Guangdong, China
pandan@gd.chinamobile.com

Abstract. Prediction of MHC (Major Histocompatibility Complex) binding peptides is prerequisite for understanding the specificity of T-cell mediated immunity. Most prediction methods hardly acquire understandable knowledge. However, comprehensibility is one of the important requirements of reliable prediction systems of MHC binding peptides. Thereupon, SRIA (Sequential Rule Induction Algorithm) based on rough set was proposed to acquire understandable rules. SRIA comprises CARIE (Complete Information-Entropy-based Attribute Reduction algorithm) and ROAVRA (Renovated Orderly Attribute Value Reduction algorithm). In an application example, SRIA, CRIA (Conventional Rule Induction Algorithm) and BPNN (Back Propagation Neural Networks) were applied to predict the peptides that bind to HLA-DR4(B1*0401). The results show the rules generated with SRIA are better than those with CRIA in prediction performance. Meanwhile, SRIA, which is comparable with BPNN in prediction accuracy, is superior to BPNN in understandability.

1 Introduction

T lymphocytes play a key role in the induction and regulation of immune responses and in the execution of immunological effector functions [1]. Binding of peptides to MHC (Major Histocompatibility Complex) molecules conveys critical information about the cellular milieu to immune system T cells. Different MHC molecules bind distinct sets of peptides, and only one in 100 to 200 potential binders actually binds to a certain MHC molecules. And it is difficult to obtain sufficient experimental binding data for each human MHC molecule. Therefore, computational modeling of predicting which peptides can bind to a specific MHC molecule is necessary for understanding the specificity of T-cell mediated immunity and identifying candidates for the design of vaccines.

Recently, many methods have been introduced to predict MHC binding peptides. They could be classified as 4 categories: 1) Prediction method based on motif [2]; 2) Prediction method based on quantitative matrices [3]; 3) Prediction method based on structure [4]; 4) Prediction method based on machine learning [5]. Because the methods in category 4 consider the interactive effect among amino acids in all positions of

J.T. Yao et al. (Eds.): RSKT 2007, LNAI 4481, pp. 300–307, 2007.

the peptide, their prediction performance has been improved a lot. The involved machine learning approaches are mainly from ANNs (artificial neural networks) and HMMs (Hidden Markov Models). Brusic et al. proposed PERUN method, which combines the expert knowledge of primary anchor positions with an EA (evolutionary algorithm) and ANNs, for prediction of peptides binding to HLA-DRB1*0401 [5].

Category 4 has better prediction performance than other categories when much structure information cannot be obtained since category 4 owns the strongest nonlinearity processing capability and generalization ability and self-organization specialty among the four categories. However, category 4 has been mainly focused on the application of ANNs so far. Meanwhile, it is very hard to understand the weights in ANNs and it is very difficult to provide the rules for the experts to review and modify so as to aid them to understand the reasoning processes in another way.

Rough set theory (RS), which was advocated by Pawlak Z. [6] in 1982, gives an approach to automatic rule acquisition, i.e., one might use RS to find the rules describing dependencies of attributes in database-like information systems, such as a decision table. The basic idea of RS used for rule acquisition is to derive the corresponding decision or classification rules through data reduction (attribute reduction and attribution value reduction) in a decision table under the condition of keeping the discernibility unchanged.

The rest of the paper is organized as follows: Section 2 proposes the methodology for prediction of MHC II-binding peptides, which consists of two subparts: peptide pre-processing and the SRIA (Sequential Rule Induction Algorithm) algorithm based on rough set theory. Section 3 describes and discusses the comparable experiment results of various algorithms. Section 4 summarizes the paper.

2 Methodology

The process of prediction of MHC II-binding peptides is composed of two phases: 1) an immunological question is converted into a computational problem with peptide pre-processing, 2) SRIA, which consists of Complete Information-Entropy-based Attribute Reduction sub-algorithm (CARIE) and Renovated Orderly Attribute Value Reduction sub-algorithm (ROAVRA), is advocated to acquire sequential rules from pre-processed peptides.

2.1 Peptide Pre-processing

MHC class II molecules bind peptides with a core region of 13 amino acids containing a primary anchor residue. Analysis of binding motifs suggests that only a core of nine amino acids within a peptide is essential for peptide/MHC binding [7]. It was found that certain peptide residues in anchor positions are highly conserved, and contributed significantly to the binding by their optimal fit to residues in the MHC binding groove [8]. Moreover, evidence further shows that MHC class II-binding peptides contain a single primary anchor, which is necessary for binding, and several secondary anchors that affect binding [5,7]. Thereupon, all peptides with the variable lengths could be reduced to putative binding nonamer cores (core sequences of nine amino acids) or non-binding nonamers.

In terms of domain knowledge about primary anchor positions in reported binding motifs [7], position one (1) in each nonamer corresponds to the primary anchor. Each non-binder is resolved into as many putative non-binder nonamers as its first position is occupied by primary anchor residue. And for binders, after the position one (1) as primary anchor residue is fixed, each binder yields many putative nonamer subsequences. Among these subsequences, the highest scoring nonamer subsequence scored by the optimized alignment matrix is regarded as pre-processed result of the corresponding binding peptide. Here, just like the description in the paper [5], an EA is utilized to obtain the optimized alignment matrix. In this way, the problem of predicting MHC class II-binding peptides is converted into the classification problem. The detailed description of peptide pre-processing is shown in paper [5].

With the pre-processed peptides (nonamers), we can form a decision table where every object represents a nonamer and the numbers of decision attributes and condition attributes respectively are one and 180 (nine positions by 20 possible amino acids at each position, i.e., amino acids are represented as binary strings of length 20, of 19 zeros and a unique position set to one). The values of condition attributes are $\{0, 1\}$ and the values of the decision attribute are $\{0, 1\}$, which corresponds to peptide classes (0: non-binders; 1: binders).

2.2 Sequential Rule Induction Algorithm

Complete Information-Entropy-Based Attribute Reduction Sub-algorithm
Here, we proposed CAR^{IE}. It can acquire an attribute reduct only comprising the essential condition attributes with higher importance measured by the information entropy, while the algorithm in [9] could obtain an attribute subset with redundancy.

Given a decision table $T=(U, A, C, D)$ and a partition of U with classes X_i, $1 \le i \le n$. Here, $C, D \subset A$ be two subsets of attributes A, called condition and decision attributes respectively. We define the entropy of attributes B as in formula (1) [9]:

$$H(B) = -\sum_{i=1}^{n} p(X_i)\log(p(X_i)) \tag{1}$$

where $p(X_i) = |X_i| / |U|$. Here, $|X|$ means the cardinality of set X.

The conditional entropy of D (U/Ind(D)=$\{Y_1, Y_2, \cdots, Y_m\}$) with reference to another attribute set $B \subseteq C$ (U/Ind(B)=$\{X_1, X_2, \cdots, X_n\}$) is defined as in formula (2):

$$H(D \mid B) = -\sum_{i=1}^{n} p(X_i)\sum_{j=1}^{m} p(Y_j \mid X_i)\log(p(Y_j \mid X_i)) \tag{2}$$

where $p(Y_j \mid X_i) = |Y_j \cap X_i| / |X_i|$, $1 \le i \le n$, $1 \le j \le m$.

The relative importance degree of an attribute a for a condition attribute set B ($B \subseteq C$) is defined as in formula (3):

$$Sgf(a, B, D) = H(D \mid B) - H(D \mid B \cup \{a\}) \tag{3}$$

For a given decision table *T=(U, A, C, D)*, the detailed steps of the CAR^{IE} sub-algorithm are as follows:

(1) Calculate the conditional entropy $H(D \mid C)$.

(2) Compute the RDCT, which was detailed in paper [10]. Here, assume the RDCT be *DD*, and let *R* be an attribute reduct of the decision table *T*.

(3) All of the columns of *DD* are summed transversely and the result is *CC*.

(4) Among the rows of *DD*, find out the rows of $\{r_1, r_2, \cdots, r_k\}$ corresponding to the rows of the locations of the minimal elements among *CC*.

(5) If the minimal element among *CC* is one, find out the columns $\{c_{r_1}, c_{r_2}, \cdots, c_{r_k}\}$ where the element *1*s in the rows $\{r_1, r_2, \cdots, r_k\}$ are located and initialize *R* be the attribute set responding to the columns $\{c_{r_1}, c_{r_2}, \cdots, c_{r_k}\}$. Otherwise, initialize *R* be empty.

(6) Compute the conditional entropy $H(D \mid R)$. Here, if *R* is empty, $H(D \mid R) = H(D)$.

(7) Judge if $H(D \mid R)$ is equal to $H(D \mid C)$; if not, repeat the steps i~iii till $H(D \mid R) = H(D \mid C)$.

i. $E = C - R$;

ii. For every attribute a $(a \in E)$, compute the relative importance degree of a for R $SGF(a, R, D)$ according to formula (3).

iii. Find out maximum $SGF(a, R, D)$, let $R = R \bigcup \{a\}$.

(8) Call the RJ algorithm in paper [10], and decide whether *R* is a reduct or not.

(9) If so, the attribute reduct is *R*. **STOP**.

(10) If not, obtain the possibly redundant attributes with RJ algorithm, and delete the condition attributes with the least $SGF(a, R, D)$ among them one by one till *R* is an attribute reduct. **STOP**.

Here, RJ algorithm is the complete algorithm for judgment of attribute reduct in paper [10], which can be used to completely and correctly judge whether an attribute subset is an attribute reduct or not. Paper [10] gives the detailed steps and proof about RJ algorithm.

Renovated Orderly Attribute Value Reduction Sub-algorithm

In order to more efficiently acquire the rule set with the stronger generalization capability, we advocate ROAVRA, which combines OAVRA [10] with domain knowledge of the primary anchor positions.

In ROAVRA, firstly, one object in a decision table is taken out in a certain sequence one at a time, and the current object's attribute value is classified. Secondly, a rule is generated according to the classification result. Finally the objects that are consistent with the current rule in the decision table are deleted. The above steps repeat until the decision table is empty. Thus, compared with OAVRA, ROAVRA need not classify the attribute values of the objects consistent with obtained rules so that it can reduce the scanning costs to a great extent.

The description of ROAVRA is as follows:

(1) Select an object in the decision table in a certain sequence one at a time. Here, we adopted a random sequence.

(2) For the selected object, classify the attribute value as three classes.

(3) According to the classification results for the selected object, judge whether the first-class attribute values are enough to constitute a correct rule. If not, one at a time precedently choose the second-class attribute value correspond-ing to an amino acid on the primary anchor position. If all attribute values of the second class have used to compose a rule and a correct rule can't be formed yet, one at a time precedently choose the attribute value corresponding to an amino acid on the primary anchor position among the third class until a correct rule can be generated. Save the obtained rule in the rule set.

(4) Delete all the objects that are consistent with the current rule in the decision table. The rest of the decision table is saved as the decision table.

(5) Repeat step (1) – (4) until the decision table is empty. STOP.

The obtained rules are sequential and have the priority order, i.e. the rule generated earlier has the higher priority order. When the rule set is used to make a decision for an unseen instance, the rules must be used in the same sequence as they were produced. If a rule is qualified to making a decision, the others with the lower priority order than its need not to be used.

3 Experiment Results and Discussions

The data set is composed of 650 peptides to bind or not bind to *HLA-DR4 (B1*0401)*, which is provided by Dr. Vladimir Brusic. The lengths of peptides are variable from 9 to 27 amino acids. With the help of SYFPEITHI software [3], the primary anchor of peptides binding to *HLA-DR(B1*0401)* can be obtained. The alignment matrix [5] is used to score each nonamer within the initial peptide after fixing the first position into any one among F, Y, W, I, L, V or M. The highest scoring nonamer sequence is seen as pre-processed results of the corresponding peptide.

Here, 915 pre-processed nonamers are obtained. There are some nonamers with unknown affinity and some inconsistence nonamers (i.e. the same nonamers have different binding affinity) among the 915 nonamers. After removing the inconsistent and unknown nonamers from 915 pre-processed peptides, we have 853 nonamers remained to analyze. The decision table is composed of 853 nonamers (553 non-binders, 300 binders). The numbers of condition attributes and decision attributes are 180 and one respectively.

In the experiment, the decision table is divided into two parts by a 4-fold stratified cross-validation sample method. The following experimentation consists of eight 4-fold stratified cross-validations.

CAR^{IE} sub-algorithm is called to compute an attribute reduct. According to the resulting attribution reduct, ROAVRA is used to acquire sequential rules. The rules have been examined and the results are shown in Table 1.

For comparison purposes, two different algorithms are utilized to process the same decision table. The first is CRIA consisting of attribute reduction sub-algorithm [11] and attribute value reduction sub-algorithm [12]. The second is BPNN.

Table 1. Test Results with SRIA

No. of Test	*sensitivity* (%)	*specificity* (%)	*precision* (%)	*accuracy* (%)
1	81.333	88.788	81.879	86.166
2	81.000	90.054	83.219	86.870
3	81.000	88.969	82.373	86.166
4	80.667	89.693	81.757	86.518
5	79.667	87.884	81.293	84.994
6	78.667	88.427	81.661	84.994
7	78.667	90.235	82.517	86.166
8	78.667	90.958	83.688	86.635
Average (%)	79.959	89.376	82.298	86.064

Table 2. Test Results with CRIA

No. of Test	*sensitivity* (%)	*specificity* (%)	*precision* (%)	*accuracy* (%)
1	62.333	76.130	83.111	71.278
2	63.667	76.492	80.591	71.981
3	64.000	79.566	82.759	74.091
4	64.667	76.673	79.835	72.450
5	65.333	76.673	86.344	72.685
6	65.667	74.503	82.083	71.395
7	65.667	75.226	82.427	71.864
8	65.667	74.684	78.800	71.512
Average (%)	64.625	76.243	81.994	72.157

Table 3. Test Results with BPNN

No. of Test	*sensitivity* (%)	*specificity* (%)	*precision* (%)	*accuracy* (%)
1	79.667	91.682	83.860	87.456
2	84.667	91.501	84.385	89.097
3	83.000	91.139	83.557	88.277
4	78.333	90.958	82.456	86.518
5	77.000	92.224	84.307	86.870
6	78.667	89.512	80.272	85.698
7	80.333	92.405	85.159	88.159
8	80.333	92.405	85.159	88.159
Average (%)	80.250	91.478	83.644	87.530

With the help of CRIA, the rules have been acquired with the training part and examined with the test part. The results are shown in Table 2.

The structure of ANNs is 180-4-1 style, i.e., the input layer and hidden layer consist of 180 nodes and 4 nodes respectively, and output layer with a single node. The learning procedure is error back-propagation, with a sigmoid activation function.

Values for learning rate and momentum are 0.2 and 0.9 respectively. The prediction performance of ANNs is shown in Table 3.

From comparisons of the test results listed in table 1, 2 and 3, we can see that the sensitivity, specificity, precision and accuracy with SRIA are much higher than those with CRIA, and very close to those with BPNN. This suggests that SRIA is much better than CRIA in the generalization capability of induced rules though the both algorithms can obtain the plain and understandable rules. In addition, compared with BPNN, SRIA can provide the comprehensible rules that can help experts to understand the basis of immunity.

Table 4 shows a part of rules generated from SRIA in the experimentation.

Table 4. A part of rules generated with SRIA

Rule No.	Antecedents	Consequent
1	1L(1)&2A(0)&2L(0)&2R(0)&2T(0)&3Q(1)&4M(0)&4Q(0)&4V(0)&5A(0)&5L(0)&6A(0)&6S(0)&6T(0)&7L(0)&7P(0)&8L(0)&8R(0)&8S(0)&9A(0)&9G(0)&9S(0)&9V(0)&9W(0)	0
2	1F(1)&2A(0)&2R(0)&4M(0)&4Q(0)&4V(0)&5A(0)&5L(0)&6A(0)&6T(0)&7L(1)&8R(0)&9A(0)&9G(0)&9S(0)&9V(0)	0
3	2R(0)&6A(1)&8R(0)&9V(1)	1

Here, we can write the third rule in table 4 as "2R(0)&6A(1)&8R(0)&9V(1) 1", i.e., if amino acid code "R" does not appear in the second position of a nonamer and "A" appears in the sixth position and "R" does not appear in the eighth position and "V" appears in the ninth position, the nonamer is classified into "binders".

4 Conclusions

In order to minimize the number of peptides required to be synthesized and assayed and to advance the understanding for the immune response, people have presented many computational models mainly based on ANNs to predict which peptides can bind to a specific MHC molecule. Although the models work well in prediction performance, knowledge existing in the models is very hard to understand because of the inherent "black-box" nature of ANNs and the difficulty of extraction for the symbolic rules from trained ANNs. In fact, comprehensibility is one of the very important requirements of reliable prediction systems of MHC binding peptides.

Thus, SRIA based on RS theory is proposed to acquire the plain and understandable rules. The CAR^{IE} algorithm, which is adopted as a sub-algorithm of SRIA, could compute an attribute reduct only comprising essential and relatively important condition attributes in a decision table composed of 180 condition attributes. The ROAVRA in SRIA is used to extract sequential rules from the reduced decision table based on the attribute reduct. Experimental results suggest SRIA is comparable to the conventional computational model based on BPNN and is obviously superior to CRIA

in prediction performance. Moreover, the SRIA algorithm can extract plain rules that help experts to understand the basis of immunity while BPNN cannot.

Acknowledgements

The authors gratefully acknowledge Dr. Vladimir Brusic for providing the data set used in this paper and Prof. Dr. Hans-Georg Rammensee for providing his paper. This work is supported in part by the NSF of Guangdong under Grant 6300252, and the Doctor Foundation of GDUT under Grant 063001.

References

1. Peter, W.: T-cell Epitope Determination. Current Opinion in Immunology. 8 (1996) 68–74
2. Falk, K., Rötzschke, O., Stevanovié, S., Jung, G., Rammensee, H.G.: Allele-specific Motifs Revealed by Sequencing of Self-peptides Eluted from MHC Molecules. Nature. 351 (1991) 290–296
3. Rammensee, H.G., Bachmann, J., Emmerich, N.P., Bachor, O.A., Stevanovic, S.: SYFPEITHI: Database for MHC Ligands and Peptide Motifs. Immunogenetics. 50 (1999) 213–219
4. Jun, Z., Herbert, R.T., George, B.R.: Prediction Sequences and Structures of MHC-binding Peptides: A Computational Combinatorial Approach. Journal of Computer-Aided Molecular Design. 15 (2001) 573–586
5. Brusic, V., George, R., Margo, H., Jürgen, H., Leonard, H.: Prediction of MHC Class II-binding Peptides Using An Evolutionary Algorithm and Artificial Neural Network. Bioinformatics. 14 (1998) 121–130
6. Pawlak, Z: Rough Sets. International Journal of Information and Computer Sciences. 11 (1982) 341–356
7. Rammensee, H.G., Friede, T., Stevanovic, S.: MHC Ligands and Peptide Motifs: First Listing. Immunogenetics. 41 (1995) 178–228
8. Madden, D.R.: The Three-dimensional Structure of Peptide MHC Complexes. Annu. Rev. Immunol.. 13 (1995) 587–622
9. Guo-Yin, W.: Rough Reduction in Algebra View and Information View. International Journal of Intelligent Systems. 18 (2003) 679–688
10. Dan, P., Qi-Lun, Z., An., Z, Jing-Song, H.: A Novel Self-optimizing Approach for Knowledge Acquisition. IEEE Transactions on Systems, Man, And Cybernetics- Part A: Systems And Humans. 32 (2002) 505–514
11. Fu-Bao, W., Qi, L., Wen-Zhong, S.: Inductive Learning Approach to Knowledge Representation System Based on Rough Set Theory" (in Chinese). Control & Decision. 14 (1999) 206–211
12. Pawlak, Z., Slowinski, R.: Rough Set Approach to Multi-attribute Decision Analysis. European Journal of Operational Reaserch. 72 (1994) 443–459