Pacific Symposium on Biocomputing 4:182-189 (1999)

Use of BONSAI decision trees for the identification of potential MHC Class I peptide epitope motifs.

C.J. SAVOIE, N. KAMIKAWAJI, T. SASAZUKI
Dept. of Genetics, Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-Ku, Fukuoka 812 Japan

S. KUHARA
Graduate School of Genetic Resources Technology Kyushu University, Hakozaki, Fukuoka 812 Japan

Recognition of short peptides of 8 to 10 mer bound to MHC class I molecules by cytotoxic T lymphocytes forms the basis of cellular immunity. While the sequence motifs necessary for binding of intracellular peptides to MHC have been well studied, little is known about sequence motifs that may cause preferential affinity to the T cell receptor and/or preferential recognition and response by T cells. Here we demonstrate that computational learning systems can be useful to elucidate sequence motifs that affect T cell activation. Knowledge of T cell activation motifs could be useful for targeted vaccine design or immunotherapy. With the BONSAI computational learning algorithm, using a database of previously reported MHC bound peptides that had positive or negative T cell responses, we were able to identify sequence motif rules that explain 70% of positive T cell responses and 84% of negative T cell responses.

1 Introduction

1.1 MHC Class I Peptide Motifs

MHC class I molecules bind short peptides of 8 to 10 mer that are primarily derived from endogenous proteins. MHC class I molecules possess peptide binding preferences at certain amino acid positions that are referred to as binding anchor residues. The bound peptides are recognized by the T cell receptors of CTL and are the primary antigenic determinants of the cellular immune response. The affect of certain amino acid residues of MHC class I bound peptides on binding to MHC has been well characterized by studies of naturally bound, eluted peptides and binding affinity studies of synthesized peptides[1,2]. This has facilitated the prediction of which peptides might bind to MHC molecules with high affinity. Such knowledge has been useful in reducing the amount of synthesized peptides that must be produced and screened in the search as peptide epitope candidates.

However, while affinity to MHC is necessary for recognition by the TCRs of cytotoxic T cells, TCR affinity alone is insufficient to cause activation and immune response by T cells. Even among of peptides that bind to MHC with equally high affinities, there can be great variations in T cell responsiveness. The affect of single

amino acid substitutions on responses by particular CTL clones have been well reported, but generalized sequence motif rules unrelated to binding affinity that influence or predict T cells have yet to be reported, although T cell responses to large synthesized peptide libraries indicate the existence of donor-independent activation motifs[3]. Identification of such motifs would be extremely useful for targeted vaccine development and epitope prediction in the investigation of immune responses in viral infection, cancer and autoimmune disease. The idea of mining potential MHC peptide epitopes based on the notion that T cell epitopes in proteins motifs tend to be concentrated into epitope-rich regions has been suggested [4,5]. However, this approach does not take account of the different binding characteristics of different MHC molecules or allelic variants nor does it discriminate between binding affinity for MHC and T cell activation. To elucidate motifs that accurately predict the likelihood of a particular bound peptide to elucidate a response, which would be useful for mining of candidate epitopes from biological databases, independent, MHC allele-specific knowledge of both the sequence motifs responsible for MHC affinity, and those motifs which affect T cell activation, is critical.

1.2 Identification of Peptide Motifs with the BONSAI Program

Knowledge acquisition from amino acid sequences by learning algorithms has been useful in the prediction of functional motifs in proteins, such as transmembrane domains [6,10,11]. The BONSAI program is based on computational learning and is used to elucidate sequence motif rules that explain the distinguishing sequence properties between two groups of amino acid sequences. In this paper, we used the BONSAI program to investigate the motif rules that predict T cell activation from a data set of peptides with reported high binding affinity to the same MHC class I molecule (HLA-A*0201).

2. Data and Methods

2.1 MHC Class I Peptide Sequence Data Set

The data samples used for this work were obtained from the MHCPEP database[8]. This database is comprised of over 13,000 MHC-bound peptides that have been previously described in the literature. The database entries include fields for the MHC corresponding molecule, binding affinity and the presence or lack of T cell

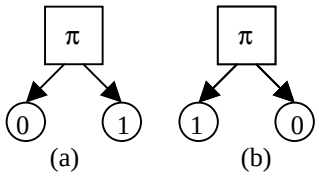
activity in response to a given peptide. Of the peptides in the database, 243 peptides were reported to exhibit high binding affinity to HLA-A*0201 and have either a positive or negative T cell response. Peptides for which the T cell response characteristics were unknown were excluded from our study. Of the high-binders 174 were reported to elicit positive cytotoxic T cell responses and 69 were reported to be negative for T cell activation. These 174 peptides comprised the study data set.

2.2 The BONSAI Program

The BONSAI algorithm is based on a elementary formal system algorithm which uses dynamic indexing to determine the appropriate regular expression clusters that produce optimum rules to explain differences between positive and negative amino acid sequence samples. Decision tree algorithms have been shown to be an effective approach for the identification of motifs from protein sequences such as the transmembrane regions in proteins. Unlike ID3[7,9], in BONSAI, the index attributes of variable regular expressions are not predefined. Rather, the index expressions are defined by optimization of discrimination resolution between positive and negative examples at runtime. Only the allowable maximum of index expressions is pre-defined outside the system. An allowable maximum for variables is necessary because it is known that the class of regular pattern languages is not polynomial-time learnable unless NP ≠ RP. The proofs for the algorithms upon which the BONSAI program are described in detail elsewhere[10,11]. Fig. 1 describes the algorithm in brief. For a decision tree *T* over regular patterns, let nodes (*T*) be the number of nodes in *T*, and τ(T) be the set of trees constructed by replacing a leaf *v* of *T* by the tree of Figure 1 (a) or Figure 1 (b) for some pattern π. The score function *Score(T,P,N)* balances the information gains in classification and is defined as

$$\text{Score}(T, P, N) = \frac{|P \cap L(T)|}{|P|} \cdot \frac{|N \cap \overline{L(T)}|}{|N|}.$$

This algorithm *DT*(*P*,*N*,*MaxNode*) checks all leaves at each phase of node generation. This algorithm is noise-tolerant in that it allows conflicts between positive and negative training examples.




```
function DT (P, N: sets of strings,
             MaxNode: int    ): tree;
begin
   if N = 0 then
   return ( CREATE ("1", null, null) )
   else if P = 0 then
   return ( CREATE("0", null, null) )
   else begin
   T ← CREATE("1", null, null);
   while (nodes(T) < MaxNode
      and Score(T,P,N) < 1 ) do
      begin
        find T_max ∈ τ(T)
             that maximizes Score (T_max, P, N);
        T ← T_max
      end
      return (T)
   end
```

Figure 1. BONSAI decision tree algorithm.

3. Results

3.1 Indexing Clusters

Figure 2 shows the decision tree generated by BONSAI for the HLA-A*0201 peptide binding peptide sequences. The decision tree for the panels of T Cell

immunoreactivity negative (TCI-) and positive (TCI+) peptides resulted in four index groups of amino acids; [0: A, K, M,], [1: C, D, T, V], [2: E, F, H, I, L, N, P, Q] and [3: G, R, S, W, Y]. These were chosen by dynamic optimization of the maximum score on the BONSAI algorithm. That is, they are free from artificial bias outside the algorithm as occurs in other implementations of ID3-based elementary formal systems, in which the index clusters are predetermined by a limited set of attributes. In the case of amino acids, these attributes could include features such as hydrophobicity or structural similarities. This cluster list, chosen based on the optimization of rules generated to explain differences between positive and negative examples of HLA-A*0201 peptides epitopes, is interesting in that it does not reflect the well-characterized features of HLA binding preferences. It is well established, for example, that the binding anchor motif strongly prefers V or L residues at P2 and at P9 of HLA-A associated peptides. However, according to the index produced by this data set, L and V are in separate index clusters. Therefore, the grouping must reflect features other than differences in the ability to bind appropriately to MHC molecules by the peptides in the negative example set, whereas more subtle differences in affinity to MHC and/or the TCR may be involved in the differentiation between negative and positive sequences, thus generating this grouping. Alternatively, features unrelated to binding affinity may be driving this selection, such as structure or differences in the size of the pool of T cells which recognize certain structural motifs of the peptide.

Indexing:

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	1	1	2	2	3	2	2	0	2	0	2	2	2	3	3	1	1	3	3

Decision Tree ----------
x33y
[YES]*** NEG *** (27, 29)
[NO]
|x100y
| [YES]*** NEG *** (24, 24)
| [NO]
| |x103y
| | [YES]*** NEG *** (2, 5)
| | [NO]*** POS *** (121, 11)
Score ----------
POS : 121 / 174 = 69.540 %
NEG : 58 / 69 = 84.058 %
Max Score ... 0.58454

Figure 1. Decision tree for T cell reactivity motif. A motif explaining 84.058% of negative cases was identified using the BONSAI program.

3.2 Identification of Adverse Motifs Affecting T cell Recognition

29 of 69 sequences negative for T cell reactivity did not contain the amino acid sequence [3,3]. Of the remaining sequences, 24 did not contain the motif [1,0,0]. This means that cumulatively, 76.8% of the negative peptides had the sequences [3,3] or [1,0,0]. Conversely, only 29.3% of the positive sequences contained either of these sequence motifs, indicating that the presence of these sequence combinations is inversely correlated with the potential for T cell reactivity. Furthermore, 121/174 (84%) of sequences positive for T cell reactivity did not possess the sequences [3,3] nor [1,0,3]. A peptide sequence that contains these "adverse motif" amino acid combinations is therefore much less likely to fall in the group of potential T cell epitopes than peptides without these sequences. Conversely, sequences that do not contain these motifs are more likely to elicit a T cell response.

4. Discussion

While the present data set that the motif rules were derived from represent a limited number of peptides compared to the complete set of peptides that could theoretically associate with MHC class I molecules, the present results do indicate the existence of sequence characteristics that affect the probability of a given bound peptide being an epitope. Only extensive experimental data can validate these rules or elucidate the mechanisms underlying such T cell recognition preferences. Such validation could potentially be aided by examining T cell responses to large combinatorial peptide libraries. Nevertheless, rules which could limit the number of peptides that should be screened for potential immunogenicity would greatly aid the work of immunologists and who presently must either synthesize all potential target epitopes (which is not always feasible), or to make educated "guesses" as to which proteins might be likely targets for investigation and limit their investigation to "suspicious" proteins, such as oncogenes in the case of cancer or envelope proteins, in the case of viral immunity. These rules will also assist in the prioritization of screening even in cases where exhaustive screening is necessitated. Further work in our laboratory will focus on the validation of the preliminary findings with experimental data, the investigation of T cell preference motifs for other MHC molecules (including class II molecules) and the generation of larger and more inclusive MHC peptide reactivity data using combinatorial chemistry to generate large peptide libraries.

Acknowledgments

This work has been supported in part by grants in aid from the Ministry of Education, Science, Sports and Culture, Japan.

References

1. T. Sudo, N. Kamikawaji, A. Kimura, Y. Date, C.J. Savoie, H. Nakashima, E. Furuichi, S. Kuhara, and T. Sasazuki, "Differences in MHC Class I self peptide repertoires among HLA-A2 subtypes." *J. Immunol.:* **155***:* 4749-4756, (1995)
2. H.G. Rammensee, T. Friede, S. Stevanovic, "MHC ligands and peptide motifs: first listing." *Immunogenetics* **41***:* 178-228 (1995)
3. T. Tana, N. Kamikawaji, C.J.Savoie, T. Sudo, Y. Kinoshita, T. Sasazuki, "A "HLA binding motif-aided peptide epitope library: A novel library design for the screening of HLA-DR4-restricted antigenic peptides recognized by CD4+ T cells." *J. Human Genet*., **43**:14-21 (1998)
4. G.E. Meister, C.G.P. Roberts, J.A. Berzofsky, A.S. De Groot, "Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from Mycobacterium tuberculosis and HIV protein sequences" *Vaccine*, **13**:581-591, (1995)
5. G.E. Meister, C.G.P. Roberts, B.T. Edelson, J.A. Berzofsky, A.S. De Groot in *Vaccines '95*, 219 – 226, "Two novel MHC-binding motif-based T cell epitope prediction algorithms; prediction of epitopes for six Mycobacterium tuberculosis protein antigens" (Cold Spring Harbor Laboratory Press, Plainview (1995)
6. S. Arikawa, S. Kuhara, S. Miyano, A. Shinohara, T. Shinohara, "A learning algorithm for elementary formal systems and its experiments on identification of transmembrane domains" *Proc. Twenty-fifth Hawaii International Conference on System Sciences*, 675-684 (1992)
7. J.R. Quinlan, "Induction of decision trees", *Machine Learning*, **1**:81-106 (1986)
8. V. Brusic, G. Rudy, L.C. Harrison, "MHCPEP, a database of MHC-binding peptides: update 1997" *Nucleic Acids Research*, **26**, Issue 01: January 1 (1998)
9. P.E. Utgoff, "Incremental induction of decision tree" *Machine Learning* **4**:161-186 (1989)
10. S. Arikawa, S. Kuhara, S. Miyano, Y. Mukouchi, A. Shinohara, T. Shinohara, "A machine discovery from amino acid sequences by decision trees over regular patterns" *Proc. of the International Conference on Fifth Generation Computer Systems*, 618-625 (1992)

11. S. Shimozono, A. Shinohara, T. Shinohara, S. Miyano, S. Kuhara, S. Arikawa, “Knowledge acquisition from amino acids sequences by machine learning system BONSAI” *Trans. of Information Processing Society of Japan*, Vol. **35**:2009-2018 (1994)