Deriving Matrix of Peptide-MHC Interactions in Diabetic Mouse by Genetic Algorithm

Menaka Rajapakse[1,2], Lonce Wyse[1], Bertil Schmidt[2], and Vladimir Brusic[1,3]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{menaka,lonce,vladimir}@i2r.a-star.edu.sg
[2] School of Computer Engineering, Nanyang Technological University,
Block N4, Nanyang Avenue, Singapore 639798
asbschmidt@ntu.edu.sg
[3] School of Land and Food Sciences and the Institute for Molecular Bioscience,
University of Queensland, Brisbane QLD 4072, Australia

Abstract. Finding motifs that can elucidate rules that govern peptide binding to medically important receptors is important for screening targets for drugs and vaccines. This paper focuses on elucidation of peptide binding to I-A^{g7} molecule of the non-obese diabetic (NOD) mouse - an animal model for insulin-dependent diabetes mellitus (IDDM). A number of proposed motifs that describe peptide binding to I-A^{g7} have been proposed. These motifs results from independent experimental studies carried out on small data sets. Testing with multiple data sets showed that each of the motifs at best describes only a subset of the solution space, and these motifs therefore lack generalization ability. This study focuses on seeking a motif with higher generalization ability so that it can predict binders in all A^{g7} data sets with high accuracy. A binding score matrix representing peptide binding motif to A^{g7} was derived using genetic algorithm (GA). The evolved score matrix significantly outperformed previously reported motifs.

1 Introduction

An I-A^{g7} motif shown in Fig. 1 describes commonly observed amino acid residues find in peptides that bind major histocompatibility complex (MHC) molecule of the non-obese diabetic (NOD) mouse (Rammensee I-A^{g7} motif) [1]. These residues, which contribute significantly to peptide binding, are called primary anchor residues and the positions they occur are called anchor positions. Anchor positions may be occupied by so called preferred residues which are tolerated, but alone contribute little to peptide binding strength. I-A^{g7} is critical for the development of insulin-dependent diabetes mellitus (IDDM) in NOD mice [2-10]. To understand the molecular basis of development of IDDM in NOD mice it is important to understand peptide binding properties to I-A^{g7}. I-A^{g7} binds peptides that are 9-30 amino acids long. Peptide binding to I-Ag7 is mediated through a binding core that is 9 amino acids long. For example, a well-known I-A^{g7} binding peptide EEIAQ**V**AT**I**S**A**NG**D**KDIGNI (mouse HSP protein 166-185) binds to I-A^{g7} via residues 171V, 174I, 176A, and 179D [4]. Of these positions, three (171V, 176A, and 179D) correspond to the primary anchors and 174I corresponds to a preferred residue in the Rammensee motif. When associated with appropriate metrics, a binding motif can be used for prediction

M. Gallagher, J. Hogan, and F. Maire (Eds.): IDEAL 2005, LNCS 3578, pp. 440–447, 2005.

of peptides that bind I-A^{g7}. For example, weights of primary anchors can be set to 4 and of preferred residues to 2. The score for mouse HSP peptide 166-185 will be 14 (4+2+4+4 for 171V, 174I, 176A, and 179D). High scoring peptides are thus predicted as I-A^{g7} binders. A widely used extension of the binding motif scoring scheme is a quantitative matrix that contain 9×20 coefficients. Nine rows represent positions and 20 columns represent each of the 20 amino acids, while matrix cells contain weights for each amino acid at a given position. The score for the prediction is calculated by summing or multiplying the coefficients. Examples of binding matrices are given in [8,9]. To our best knowledge, a quantitative matrix for I-A^{g7} has not been reported to date.

	Position								
	1	2	3	**4**	5	**6**	7	8	**9**
Primary anchors	K,H, S,A, V			L		V,A			D,S, E
Preferred residues	R,T			I,V, M		T			

Fig. 1. Peptide binding motif for the I-A^{g7} molecule – see the main text for the description

Some high-affinity binders to I-A^{g7} such as mouse GAD (247-261) peptide NMY**A**ML**I**ERYKMEPE [7] do not correspond well to the Rammensee motif – the best 9-mer window in this peptide has one primary anchor (250A) and one preferred residue (253I). This indicates that any one binding motif is likely to be an imperfect approximation of rules that describe peptide binding to I-A^{g7}. Indeed, we found seven different I-A^{g7} motifs derived from largely unrelated experimental data sets. These include reported motifs Reizis [4], Harrison [5], Gregori [7], Latek [6], Rammensee [1], Reich [2], and Amor [3]. These seven motifs are mutually inconsistent and some are completely different. Each motif describes amino acids at primary and secondary anchor positions, as well as "forbidden" amino acids at specific positions. We interpreted these as well-tolerated, weakly-tolerated, and non-tolerated amino acids. We adopted the following metrics: well-tolerated residues have weight 4, weakly-tolerated 2, and non-tolerated amino acids -4. Anchor positions were assigned weights – primary anchor positions have weight 4 and secondary anchor positions weight 2. The primary and secondary anchor positions were defined according to the motif descriptions by the authors. The binding motifs and the scoring scheme can be accessed at <research.i2r.a-star.edu.sg/Ag7motifs>. In this work we seek to: a) compare the predictive ability of the seven reported motifs, b) combine existing data and develop a method for the derivation of a unified motif that describes well all available data, and c) compare several data-driven methods for the identification of the unified motif. The I-A^{g7} 7-related data were extracted from multiple data sets shown in Table 1. We adopted a quantitative matrix as a model for the unified I-A^{g7} motif. Three well-known methods were employed in the search for the best I-A^{g7} quantitative matrix: Multiple EM for Motif Elicitation (MEME) [14], Gibbs sampling [see 9], and genetic algorithm [see 15]. Here we report the unified motif for I-A^{g7}, and the comparative analysis of the motifs used in this study.

Table 1. I-A^{g7} related peptide data sets

Data set	Non binders	Binders	Reference
Reizis	21	33	[4]
Harrison	19	157	[5]
Gregori	31	109	[7]
Latek	8	37	[6]
Corper	35	13	[10]
MHCPEP	-	176	[11]
Yu	16	10	[12]
Stratmann	3	118	[13]
Brusic	37	-	[unpublished]

2 Characterization of Motif Using a Binding Score Matrix

In this section we give a formal definition of the target model as a quantitative matrix. A *k*-mer motif in an amino acid sequence is usually characterized by a binding score matrix **Q** = $\{q_{\mathrm{ia}}\}_{\mathrm{kx20}}$ where q_{ia} denotes the *binding affinity* of the site *i* of the motif, when it occupies by the amino-acid $a \in \sum$; $\sum$ denotes the set of 20 amino-acid residues. The cumulative binding score for a *k*-mer not only indicates the likelihood of the presence of a particular motif but also determines the likelihood that a sequence containing the motif binds to another sequence. Therefore, the binding score matrix can be viewed as a quantification of a real biological functioning or binding of the motif to other peptides as described in [13]. Given a binding score matrix **Q** of size *k*×20 we define the *binding score, s* for a *k*–mer motif in a sequence of length *n* as:

$$s = \max_{j \in \{1,\ldots,n-k+1\}} s_j \tag{1}$$

$$s_j = \sum_{i=0}^{k-1} \sum_{a \in \Sigma} q_{ia} \cdot \delta_{ij} \qquad with \qquad \delta_{ij} = \begin{cases} 1 & \text{if } x_{j+i} = a \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

We define *m** as the *k*-mer of sequence *x* at position *j*, i.e. $m^* = (x_j, \ldots, x_{j+k-1})$, where

$$j = \arg\max_{j \in \{1,\ldots,n-k+1\}} s_j \tag{3}$$

3 Description of the Method

Let the number of training data sets extracted be *d,* and the number of motifs inferred from different experiments be Ψ. We can then express the available prior information as **D** = $\{(D_i, m_l): i=1,2,\ldots.\mathrm{d}, l=1,2,.., \Psi\}$ where m_l is the consensus motifs found in the experiments. Let $D_i = \{(\mathbf{x}_{\mathrm{ij}}, b_{\mathrm{ij}}): j = 1, 2,\ldots, n_i\}$ where $\mathbf{x}_{ij}$ is the j^{th} sequence in the i^{th} dataset and $b_{ij} \in \{0, 1\}$ indicates whether the sequence $\mathbf{x}_{ij}$ is a binder (when equal to one), or a non-binder (when equal to zero). The collated dataset is then given by $\Gamma = \{\mathbf{x}_{ij}: i=1,2,\ldots..d; j=1,2,\ldots n_i\}$ where n_i is the number of sequences in i^{th} dataset. With these information extracted from the experimentally validated motifs we seek a motif **m*** that best describes the consensus segment in all the sequences in Γ.

3.1 Training and Test Data Sets

The training and test data sets in the experiments are given in Table 1. These data sets consist of short peptides ranging from 9-30 amino acids per sequence. Except for the Stratmann data set, all other data sets were used in the training. The Stratmann test set contains only 118 binders and three non binders. Because of the small number of experimentally determined non-binders, we extended the number of non-binders in this set to 1000 by generating random peptides. The generation of random peptides involved adding correct proportions of amino acids to each peptide so that the generated peptide mimics real protein peptides [16]. Of 1000 random peptides generated, at most five percent are presumed to be binders. This percentage was estimated based on the analysis of I-A^{g7} binding data given in [10].

3.2 Multiple EM for Motif Elicitation (MEME)

MEME is a tool for discovering motifs in protein or DNA sequences in an unsupervised manner [14]. All I-A^{g7} binders were converted to *fasta* format and submitted to the public domain MEME analysis tool [17] and three motifs were requested. The position scoring matrices retrieved were assessed for predictive accuracy.

3.3 Gibbs Sampling

Another tool, Gibbs sampling is also used in the analysis. Gibbs sampling is less susceptible to becoming trapped in a local minima. Details about the Gibbs Motif Sampler can be found in [18, 19]. Input data for the Gibbs sampling are the same as for MEME. A single motif was retrieved. A scoring scheme was formulated based on the mutual information contain in each position and assessed for predictive accuracy.

3.4 Genetic Algorithm

Genetic algorithms work with a fixed number of individuals as its population each representing a particular solution. Let the population at time or iteration t of evolution be Q(t). During a single iteration, each chromosome is evaluated against the goodness of the solution by using a fitness function, f.

Binary String Representation: The binding strengths of elements of binding score matrix, q_{ia}, $\{i=1,\ldots.k, \text{a} \in \Sigma\}$ for each a sequence have been empirically determined and quantitatively expressed in the data sets. Each individual (binding score matrix) in the population is represented by a binary string. A binding score matrix of size k x n, where k represents motif length and n represents number of residues has kn elements.

Fitness Computation: The definition of the fitness function is crucial. The fitness function, in our case, is expected to yield a unified consensus motif for the training set. The dataset of each experiment in the literature gives the information whether the particular sequence is a binder or non-binder. Using this information, the numbers of true positives (TP) and true negatives (TN) determined by solutions in the population could be computed. A highly probable candidate solution must produce lower binding score when tested on a non-binding peptide than on a binding peptide. By incorporating the TPs and TNs resulting from the evaluation and taking into account binding scores for binders and non-binders, we defined a fitness function f on a putative motif, m, representing a binding score matrix Q is defined as:

$$f(Q) = \frac{\sum_{i=1}^{d} \sum_{j=1}^{n_i} s(x_{ij} : m)(1-b_{ij})}{TP + \eta * TN} \tag{4}$$

The GA finds a score matrix that minimizes the above fitness value. An empirically obtained constant, η is used to minimize the number of possible false positives that can arise from the solution matrix with respect to the non binders.

Construction of Template Score Matrices (Seeds) for Initial Population: In our analysis, seven template scoring matrices were constructed as seeds for the initial population. These template scores are based on the knowledge inferred from the literature. A scoring scheme was enforced on the template formation of scoring matrices by assigning a score of 0 that are non tolerant at a specific site. The highest score of 80 was assigned if the site is in a critical position and the amino acid at that position is categorized as a well tolerant. A base score of 10 was assigned for all the other positions that have no significant contribution. The seeding for the rest of the population was carried out with a super-uniform random generator which yields a population representing all schemata up to a certain defining length (limited by the population size) with large global correlations [see 20].

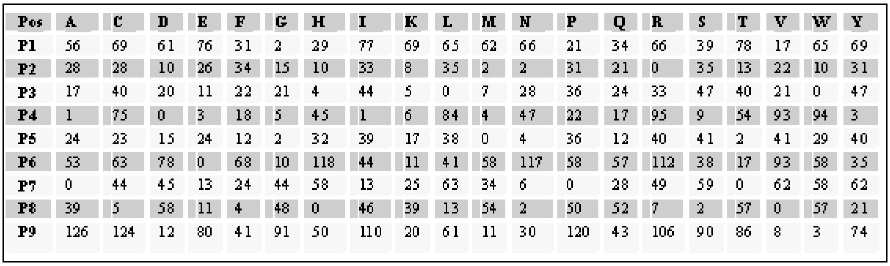
4 Experimental Results

The motifs generated from MEME, Gibbs sampling and the best GA-derived scoring matrix for the cumulative data set are shown below. Using these motifs and scoring matrices we measured the predictive performance on the Stratmann data set combined with randomly generated non-binders (Tables 3 and 4). The performance was measured by the area under the receiver operating characteristics (AROC) curve and estimates of cut-off points between sensitivity and specificity plots (SE=SP). These metrics indicate the generalization ability of each method across different data sets.

MEME Motifs: Motif1: MKRHGLDNY *Motif2:* AE(Y)Y(Q)LI(K)N(T)VMD *Motif3*:CAKKIVSDG. Multi-level motif derived from the Gibbs Sampling: *Gibbs Motif*: N(MP)K(V)A(RI)T(H)G(A)E(FL)D(Q)N(YL)K(YV). The amino-acid inside the bracket indicates a possible substitution for the amino-acid to its immediate left.

The predictive performance was measured by A_{ROC} values and generalization ability by estimating the cut-off points between sensitivity and specificity plots for previously unseen data (Stratman A^{g7} binders, plus randomly generated non-binders). The GA based scoring matrix outperformed the next best method by some 10% on Stratmann data set. Most of the motifs showed marginal predictive accuracy ($0.8>A_{ROC}>0.7$). As expected, majority of the motifs performed well on the data sets they were derived from, but less so on the independent data sets. While Gibbs sampling produced marginal results, the performance of MEME on these data sets was unremarkable.

Table 2. The final scoring matrix derived by GA



Pos	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	56	69	61	76	31	2	29	77	69	65	62	66	21	34	66	39	78	17	65	69
P2	28	28	10	26	34	15	10	33	8	35	2	2	31	21	0	35	13	22	10	31
P3	17	40	20	11	22	21	4	44	5	0	7	28	36	24	33	47	40	21	0	47
P4	1	75	0	3	18	5	45	1	6	84	4	47	22	17	95	9	54	93	94	3
P5	24	23	15	24	12	2	32	39	17	38	0	4	36	12	40	41	2	41	29	40
P6	53	63	78	0	68	10	118	44	11	41	58	117	58	57	112	38	17	93	58	35
P7	0	44	45	13	24	44	58	13	25	63	34	6	0	28	49	59	0	62	58	62
P8	39	5	58	11	4	48	0	46	39	13	54	2	50	52	7	2	57	0	57	21
P9	126	124	12	80	41	91	50	110	20	61	11	30	120	43	106	90	86	8	3	74

Table 3. Table 3. The A_{ROC} values from predictions using each motif across all the data sets. A_{ROC}>0.9 correspond to excellent, 0.8<A_{ROC}<0.9 to good, 0.7<A_{ROC}<0.8 to marginal prediction accuracy. A_{ROC}=0.5 corresponds to random guessing, and 0.5<A_{ROC}<0.7 to poor predictions

	A_{ROC} values							
Motif for predictions	**INDIVIDUAL DATA SETS**							
	Reizis	**Harrison**	**Gregori**	**Latek**	**Corper**	**MHCPEP**	**Yu**	**Stratman**
Reizis	**0.95**	0.68	0.74	0.95	0.50	0.59	0.48	0.67
Harrison	0.75	**0.88**	0.69	0.64	0.53	0.72	0.33	0.79
Gregori	0.64	0.68	0.71	0.73	0.40	0.64	0.61	0.79
Latek	0.66	0.72	**0.80**	**0.95**	0.64	0.52	0.75	0.75
Rammense	0.49	0.64	0.76	0.82	0.60	0.48	0.43	0.77
Reich	0.55	0.64	0.69	0.58	0.56	0.47	0.50	0.73
Amor	0.69	0.54	0.66	0.70	0.56	0.66	0.40	0.78
MEME1	0.61	0.58	0.49	0.60	0.43	0.55	0.36	0.49
Gibbs	0.33	0.79	0.77	0.81	0.39	0.64	0.58	0.82
GA	0.76	0.86	0.76	**0.96**	**0.79**	**0.83**	**0.94**	**0.88**

Table 4. Cutoff points (SE=SP) for predictions using each motif across all the data sets

	Cut-off points SE=SP							
Motif for predictions	**INDIVIDUAL DATA SETS**							
	Reizis	**Harrison**	**Gregori**	**Latek**	**Corper**	**MHCPEP**	**Yu**	**Stratman**
Reizis	**0.87**	0.79	0.69	0.89	0.50	0.57	0.56	0.74
Harrison	0.64	**0.84**	0.65	0.58	0.50	0.68	0.40	0.73
Gregori	0.58	0.62	0.62	0.65	0.42	0.63	0.58	0.74
Latek	0.66	0.68	**0.73**	**0.92**	0.60	0.50	0.65	0.72
Rammense	0.52	0.58	0.70	0.77	0.52	0.46	0.51	0.70
Reich	0.54	0.60	0.62	0.57	0.56	0.48	0.50	0.67
Amor	0.62	0.54	0.62	0.66	0.55	0.60	0.42	0.71
MEME1	0.48	0.55	0.68	0.47	0.63	0.50	0.47	0.47
Gibbs	0.32	0.68	0.71	0.66	0.37	0.58	0.56	0.71
GA	0.72	0.80	0.72	0.92	**0.64**	**0.75**	**0.90**	**0.83**

5 Discussion and Conclusions

We have devised a scoring matrix representing a consensus motif with higher generalization ability than other proposed motifs derived for I-A^{g7} data sets found in the literature. Motifs described in the literature for I-A^{g7} data were tested on an independent data set (Stratmann data set together with 1000 randomly generated non-binders) for the estimation of the prediction accuracy of the evolved matrix. Random non-binders were generated using approximated amino acid compositions. The GA matrix performed well across all data sets, and showed higher generalization ability than the other proposed motifs. The ability of the GA to search a larger solution space in a context independent manner may have eliminated biases in the data sets such as fewer training data, an unequal number of binders and non-binders in the data sets, thereby providing a better solution in finding a consensus motif for difficult and unbalanced data sets.

References

1. Rammensee, H. *et al.*: SYFPEITHI:database for MHC ligands and peptide motifs. Immunogenetics 50 (1999) 213-219
2. Reich,E.P. *et al.*: Self peptides isolated from MHC glycoproteins of non-obese diabetic mice. J. Immunology 152 (1994) 2279-2288
3. Amor,S. *et al.*: Encephalitogenic epitopes of myelin basic protein, proteolipid proteing, and myelin oligodendrocyte glycoprotein for experimental allergic en-cephalomyelitis induction in Biozzi AB/H(H-2A^{g7}) mice share an amino acid motif. J. Immunology 156 (1996) 3000-3008
4. Reizis,B., *et al.*: Molecular characterization of the diabetes mouse MHC class-II protein, I-A^{g7}. Int. Immunology 9 (1997) 43-51
5. Harrison,L.C. *et al.*: A peptide binding motif for I- A^{g7}, the class II major jistocompatibility complex (MHC) molecule of NOD and Biozzi AB/H mice. J. Exp. Med. 185 (1997) 1013-1021
6. Latek,R.R. *et al.*: Structural basis of peptide binding and presentation by the type I diabetes-associated MHC class II molecule of NOD mice. Immunity 12 (2000) 699-710
7. Gregori,S. *et al.*: The motif for peptide binding to the insulin-dependent diabetes mellitus-associated class II MHC molecule I-A^{g7} validated by phage display library. Int. Immunology 12(4) (2000) 493-503
8. Brusic,V. *et al.*: Application of genetic search in derivation of matrix models of peptide binding to MHC molecules. Proc. Int. Conf. Intell. Syst. Mol. Biol. 5 (1997) 75-83
9. Nielsen,M, *et al.* Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. Bioinformatics 20 (2004) 1388-1397.
10. Corper,A.L. *et al.*: A structural framework for deciphering the link between I-A^{g7} and autoimmune diabetes. Science 288 (2000) 505-511
11. Brusic,V., Rudy,G., Harrison,L.C. MHCPEP, a database of MHC-binding peptides: update 1997. Nucleic Acids Res. 26 (1998) 368-371
12. Yu,B. *et al.*: Binding of conserved islet peptides by human and murine MHC class II molecules associated with susceptibility to type I diabetes. J. Immunology 30(9) 2497-506
13. Stratman,T. *et al.*: The I-A^{g7} MHC class II molecule linked to murine diabetes in a promiscuous peptide binder. J. Immunology 165 (2000) 3214-3225
14. Bailey,T.L., Elkan,C. The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol. 3 (1995) 21-29
15. Brusic V, Schonbach C, Takiguchi M, Ciesielski V, Harrison LC. Application of genetic search in derivation of matrix models of peptide binding to MHC molecules. Proc Int Conf Intell Syst Mol Biol. 5 (1997) 75-83
16. Pe'er I. *et al.* Proteomic Signatures:Amino Acid and Oligopeptide Compositions Differentiate Among Phyla. Proteins 54 (2004) 20-40

17. http://meme.scdc.edu/meme/website/meme.html
18. Neuwald, A. F. *et al.*: Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Science 4 (1995) 1618-32
19. Lawrence,C.E *et al.*: (1993) Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. Science 262 (1993) 208-214
20. Schraudolph,N., Grefenstette J.: A User's Guide to GAucsd 1.4, Technical Report, University of California, San Diego, (1992) CS 92-249