Evolutionary approaches for predicting MHC peptide binding

Rajapakse, Menaka Priyadharasani

2009

Rajapakse, M. P. (2009). Evolutionary approaches for predicting MHC peptide binding.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/14955>

<https://doi.org/10.32657/10356/14955>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**EVOLUTIONARY APPROACHES FOR
PREDICTING MHC PEPTIDE BINDING**

**MENAKA PRIYADHARSANI RAJAPAKSE
SCHOOL OF COMPUTER ENGINEERING**

2009

EVOLUTIONARY APPROACHES FOR PREDICTING MHC PEPTIDE BINDING

MENAKA PRIYADHARSANI RAJAPAKSE

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2009

Table of Contents

Table of Contents	iii
List of Tables	vi
List of Figures	x
Abstract	xv
Author's Publications	xvii
1 Introduction	1
1.1 Motifs and Peptide Binding	2
1.2 Motif Discovery and MHC-Peptide Binding	4
1.3 Motivation	8
1.4 Main Contributions	11
1.5 Organization of the Thesis	13
2 Background	15
2.1 Immunology	16
2.1.1 Immune System	17
2.1.2 Protein Antigen	18
2.1.3 MHC Molecules	19
2.1.4 Validation of Binding Peptides	22
2.1.5 MHC-Peptide Binding Motif	24
2.2 Motif finding Methods	24
2.2.1 Profile Analysis	25
2.2.2 Gibbs Sampling	28
2.2.3 Expectation Maximization Algorithm	32
2.3 Methods of MHC-Peptide Binding	35

2.3.1	Sequence Based Methods	36
2.3.2	Structure Based Methods	38
2.3.3	Modeling Issues	41
2.3.4	Performance Measures	42
2.4	Evolutionary Approach	45
2.4.1	Evolutionary Algorithms (EA)	47
2.4.2	Multi-Objective Optimization	48
3	Self-Discovery of Motifs for Predicting MHC Peptide Binding	53
3.1	Previous Work	54
3.1.1	Prediction of MHC class II peptides using EA and ANN (PERUN)	55
3.1.2	Motif Discovery using a Genetic Algorithm (MDGA)	55
3.1.3	Ant Colony Search Strategy	56
3.2	Motivation	57
3.3	Problem Representation	59
3.3.1	Motif Representation and Peptide Scoring	59
3.3.2	Representing PSSM in the EA Population	60
3.4	GAA based Motif Discovery	60
3.4.1	Fitness Function	62
3.5	MOEA based Motif Discovery	64
3.5.1	Non-Sorting Genetic Algorithm-II	64
3.5.2	Objective Functions and Constraints	73
3.6	MHC-Peptide Binding Data	74
3.6.1	I-A ^{g7} Dataset	74
3.6.2	HLA-DRB1*0401 Benchmark Dataset	76
3.6.3	IEDB Benchmark Dataset	77
3.7	Experiments	78
3.8	Results	80
3.8.1	Performance Comparison	80
3.8.2	EER Plots	87
3.8.3	Run Time Comparison	87
3.8.4	Parameter Settings	88
4	Guided-Discovery of Motifs	90
4.1	Motivation	91
4.2	Guided-Discovery by Experimental Motifs	93
4.2.1	I-A ^{g7} Experimental Motifs	93
4.2.2	Scoring Experimental Motifs	98
4.2.3	Representation of Experimental Motifs	99

4.2.4	Fitness Functions	99
4.3	Guided-Discovery by Computationally Predicted Motifs	101
4.3.1	Representation and Scoring of Computationally Predicted Motifs	101
4.4	Guided-Discovery by Anchor Positions	102
4.4.1	Clustering and Weighting	103
4.4.2	Pseudo-Count Correction	103
4.4.3	Representing an Individual in a Population	104
4.4.4	Fitness Functions	105
4.5	Experiments and Results	106
4.5.1	I-A ^{g7} dataset	106
4.5.2	I-A ^b Positive Modeling dataset	111
4.5.3	Performance of Guided-Discovery by Computationally Predicted Motifs	113
5	Characterization of Motifs	116
5.1	Previous Work	117
5.2	Motivation	118
5.3	Amino Acid Properties	120
5.4	Methods	124
5.4.1	Entropy of a Motif	125
5.4.2	Motif Logos	126
5.4.3	Characterization Based on Binding Sites	127
5.4.4	Characterization by Bootstrapping of Binding Cores	127
5.5	Experiments and Results	128
5.5.1	Characterization of I-A ^{g7} Predicted Binding Sites	128
5.5.2	Characterization of I-A ^{g7} by Random Selection of Binding Cores	130
5.5.3	Analysis of Entropy and Motif Logos	133
5.6	Validation of Physicochemical Profiles	141
6	Conclusion	147
6.1	Summary	147
6.2	Discussion	151
6.3	Future Work	153
	Bibliography	155
	Appendix	173

List of Tables

1.1	Representation of a DNA motif with a consensus motif.	3
1.2	Representation of a DNA motif with position specific weight matrix. P1 to P7 denotes the positions of the motif. The values indicate the likelihood of the presence of a particular nucleotide.	3
2.1	MHC-peptide binding prediction methods. (The list provided is by no means an exhaustive list)	40
2.2	Pseudocode for evolutionary algorithm	48
3.1	Pseudocode demonstrating Genetic Annealing Algorithm	61
3.2	Pseudocode illustrating the fast non-dominated sorting algorithm. . .	67
3.3	Algorithm for estimating crowding distance	69
3.4	Information on I-A ^{g7} related peptide binding datasets. Unavailable information is indicated by ‘-’	76
3.5	Description of the benchmark dataset, BM-SET1.	78
3.6	Description of BM-SET2	79
3.7	Seven-fold cross-validation performance of self-discovered motif on the training dataset	82

3.8	Comparison of performances of the self-discovery methods of motif detection on BM-Set1 (DRB1*0401). Performances are measured in AUC values and the best values are shown in bold. †These values are based on smaller dataset sizes as SVRMHC didn't predict values for some of the peptides. The values from the Gibbs sampler were estimated from the matrix provided by the authors in [1].	84
3.9	Comparison of performances of the self-discovery method of motif detection on BM-Set2. Performance is measured in AUC values. The best values displayed in bold are obtained by comparing all the methods listed in the table except the ARB method as ARB method has been trained on quantitative data obtained from IEDB and therefore a fair conclusion can not be drawn. The values from the Gibbs sampler were estimated from the matrix provided by the authors in [1].	86
3.10	Parameter settings for GAA and MOEA.	89
4.1	Information on I-A ^{g7} related peptide binding datasets and motifs. Unavailable information is indicated by '-'. A motif is denoted by $m()$, for example $m(\text{Reizis}) = \text{motif by Reizis}$	93
4.2	The description of experimentally determined I-A ^{g7} 9-mer peptide binding motif by Reizis: each position accommodates a well-tolerated, weakly-tolerated, or non-tolerated amino acid. The positions P4, P6 and P9 are the primary anchor positions where binding is highly likely to occur.	94
4.3	Harrison motif (Harrison et al., 1997) In Harrison motif, the positions P6 and P9 are primary anchors. Positions P3 and P8 are considered secondary anchor positions.	95
4.4	Gregori motif (Gregori et al., 2000) In Gregori motif, the positions P6 and P9 are primary anchors. Positions P3, P4, P7 and P8 are considered secondary anchor positions.	96

4.5	Latek motif (Latek et al., 2000) The positions P1, P4, P6 and P9 are considered as primary anchors in Latek motif. Position P7 is considered a secondary anchor position	96
4.6	Rammensee motif (Rammensee et al., 1999) In Rammensee motif, the positions P1, P4, P6 and P9 are primary anchors. Position P7 is considered a secondary anchor position.	97
4.7	Reich motif (Reich et al., 1994) The Reich motif given in the paper (Reich et al., 1994) is based on MSA 560-574. This motif has been modified for the present study by taking a 9-mer sequence (565-573) from the consensus sequence. The modified motif is chosen based on the above defined Reizis and Rammensee motifs.	97
4.8	Amor motif (Amor et al., 1996) In Amor motif, four positions have been described as very important. By considering the residues for these four positions and comparing them with the other motifs, positions P6, P7, P8 and P9 were chosen.	98
4.9	Quantitative measures defined for the tolerance level and the binding strength in experimental motifs.	99
4.10	Seven-fold cross-validation performance of guided-discovery motif on the training dataset.	108
4.11	Performance measured by AUC of experimentally determined I-A ^{g7} motifs on their own datasets and other experimental datasets.	110
4.12	Parameter setting for guided-discovery with experimental motifs.	112
4.13	Comparison of performance of guided-discovery using positive modeling of motifs against MEME, Gibbs sampler, MOEA-self discovered motif, and RANKPEP on I-A ^b dataset	112
4.14	Comparison of performance of guided-discovery by using computationally predicted motifs. Same dataset was used to measure the performance of all three methods: MEME, Gibbs sampler, MOEA-guided discovery	115

4.15	Parameter setting for guided-discovery with anchor positions.	115
5.1	Physicochemical properties of Amino Acids: the scales illustrated for hydropathy, polarity, bulkiness are Kyte-Dolittle [2], Grantham [3], and Zimmerman[4], respectively.	123

List of Figures

2.1	Illustration of T cell activation through antigen binding to MHC and MHC-antigen complex presentation to TCR.	16
2.2	The structure of HLA (human MHC) complex in chromosome 6. . . .	20
2.3	Illustration of peptide binding to MHC class I molecules with (a) a ribbon diagram, and (b)a schematic diagram	21
2.4	Illustration of peptide binding to MHC class I molecules with (a)a ribbon Diagram, and (b)a schematic Diagram	22
2.5	Illustration of a profile matrix	26
2.6	Illustration of a constrained optimization problem	47
2.7	Illustration of Pareto optimal front	50
2.8	Illustration of dominance	51
3.1	Illustration of crowding distance estimation. Points marked in circles are solutions of the same non-dominated front.	68
3.2	Illustration of convergence to a Pareto optimal front.	81
3.3	Illustration of the true boundary obtained for I-A ^{g7} training dataset. .	83
3.4	Comparison of performance of self-discovery motifs against MEME, RANKPEP, and experimental motifs on the balanced I-A ^{g7} test datasets (the performance was averaged over 25 test datasets)	85
3.5	Illustration of choosing the best threshold from an EER plot. Threshold which corresponds to the point where FAR and FRR intercept is selected as the threshold that minimizes the error in prediction. . . .	87

4.1	An illustration of putative 9 aa binding cores within a peptide sequence, s_i . Highlighted positions P1,P4,P6,P7,and P9 indicate anchor positions within a putative binding core associated with I-A ^b	103
4.2	The flow chart of the motif discovery process guided by the anchor positions. Anchor positions are given more weightage during the estimation of fitness functions. An individual (solution) in the population individual (solution) represents starting positions of all putative binding cores in the positive sample set. EA operators represent crossover and mutation, respectively.	107
4.3	Illustration of the true boundary obtained for I-A ^{g7} training dataset from guided discovery approach.	109
4.4	Comparison of performance of guided-discovery motifs against MEME, RANKPEP, and experimental motifs on the balanced I-A ^{g7} test datasets (the performance was averaged over 25 test datasets).	111
4.5	Illustration of AUC values obtained during the parameter estimation for α_1 and α_2 . A subset of the I-A ^{g7} training dataset was used for the parameter estimation. Discontinued curves indicate that there were no feasible solutions for given α_1 and α_2 values. Parameter estimation was performed by fixing one parameter (α_1) and evaluating AUC values for a range of values of the other parameter (α_2).	113

4.6	An illustration of MOEA based motif discovery process guided by computationally predicted motifs. As an example, Motifs predicted by MEME and Gibbs approach were used to guide the motif discovery process. Three mutually exclusive binding datasets were used to train MEME, Gibbs, and MOEA respectively. While binding datasets B1 and B2 were used for training MEME and Gibbs, MOEA was trained with B3 binder set combined with a non-binder dataset, NB _x . An independent test dataset comprising binder set B and non-binder dataset NB _y was used for testing where AUC values were estimated for the test dataset from each of the discovered motif $m(.)$	114
5.1	Illustration of the chemical structure of an amino acid.	120
5.2	Venn-diagram indicating physicochemical properties of amino acid residues.	121
5.3	Illustration of relative abundance of hydrophobic residues at the predicted I-A ^{g7} binding site.	129
5.4	Illustration of relative abundance of polarity at binding positions of predicted I-A ^{g7} binding site.	129
5.5	Illustration of relative abundance of volume of residues at the predicted I-A ^{g7} binding site.	130
5.6	Illustration of relative abundance of hydrophobic residues at randomly selected 9-mers of I-A ^{g7} training dataset.	131
5.7	Illustration of relative abundance of polarity at randomly selected 9-mers of I-A ^{g7} training dataset.	131
5.8	Illustration of relative abundance of volume of residues at randomly selected 9-mers of I-A ^{g7} training dataset.	132
5.9	Comparison of variation of relative abundance of physicochemical properties along 9-mer positions of (a) a predicted binder (b) a random selection	132

5.10	Illustration of entropy at each position at the predicted binding site and at the randomly selected 9-mer binding core in the I-A ^{g7} training dataset.	133
5.11	Motif logos derived for I-A ^{g7} mouse allele (a)self-discovery (b)guided-discovery.	134
5.12	Motif logos derived for mouse alleles (a)I-A ^b (b)I-A ^d and (c)I-A ^s . . .	136
5.13	Motif logos derived for HLA alleles in the BM-Set2 (a)DRB1-0101 (b)DRB1-0301 (c)DRB1-0401 and (d)DRB1-0404.	137
5.14	Motif logos derived for HLA alleles in the BM-Set2 (a)DRB1-0405 (b)DRB1-0701 (c)DRB1-0802 and (d)DRB1-0901.	138
5.15	Motif logos derived for HLA alleles in the BM-Set2 (a)DRB1-1101 (b)DRB1-1302 and (c)DRB1-1501.	139
5.16	Motif logos derived for HLA alleles in the BM-Set2 (a)DRB4-0101 and (b) DRB5-0101.	140
5.17	Illustration of relative abundance of residues at predicted binding core (BC) and randomly selected 9-mer core (RBC) in the I-A ^{g7} training set (a) hydrophobic (b) hydrophylic	143
5.18	Comparison of relative abundance of neutral residues between MHC-peptide binding sites and 9-mers randomly selected from the I-A ^{g7} training dataset	144
5.19	Illustration of relative abundance of residues at predicted binding core (BC) and randomly selected 9-mer core (RBC) in the I-A ^{g7} training set (a) polar (b) non-polar	144
5.20	Illustration of relative abundance of residues at predicted binding core (BC) and randomly selected 9-mer core (RBC) in the I-A ^{g7} training set (a) charge (b) uncharge	145
5.21	Illustration of relative abundance of residues at predicted binding core (BC) and randomly selected 9-mer core (RBC) in the I-A ^{g7} training set (a) verylarge (b) large	145

5.22 Comparison of relative abundance of medium residues between MHC-peptide binding sites and 9-mers randomly selected from the I-A ^{g7} training dataset	146
5.23 Illustration of relative abundance of residues at predicted binding core (BC) and randomly selected 9-mer core (RBC) in the I-A ^{g7} training set (a) verysmall (b) small	146

Acknowledgements

Firstly, I would like to express my sincere gratitude to my Ph.D. advisor, Associate Professor Lin Feng, for his valuable suggestions and guidance for this research, and agreeing to supervise this project when my previous advisor left.

I would also like to pay my gratitude towards my previous advisor, Dr. Bertil Schmidt for his support throughout this project. A special thank goes to Dr. Vladimir Brusic, who also co-supervised this research project, for introducing me to the field of immunoinformatics, encouraging me to pursue Ph.D. studies, and for his constant advice.

My sincere thanks go to my ex-colleagues, Dr. Judice Koh, Dr. Guang Lan Zhang, and Mr. Asif Khan, whom I could always count on. Also, thanks go to my colleagues at Institute for Infocomm Research in Singapore, especially the Data Mining Department, for their special friendships and support. I also would like to thank the people who have advised me or share their expertise to help me culminate my thesis.

Thanks to my parents and my husband Jagath, for their constant encouragement and support to pursue my endeavours.

Without all of your encouragement and support, this thesis would not have been possible. Thank you all.

Abstract

T cells of the immune system recognize short linear peptides, known as epitopes, derived from degradation of protein antigens and presented by major histocompatibility complex (MHC) molecules on the surface of antigen processing cells. Antigen presentations in MHC class I and II pathways are critical for the initiation and regulation of immune responses. Recognition of T cell epitopes is a key regulatory step in control of infectious and autoimmunity diseases, and tumors. Accurate computational predictions can dramatically speed up the identification of potential T cell epitopes and thereby minimize the number of wet lab experiments needed for investigation of molecular basis of immunity. Prediction of binding peptides to MHC class II is more difficult than to class I molecules because of their ability to bind to peptides of different lengths. Having variable length binding peptides in a dataset restricts the utilization of classifiers because the number of input features to the classifier is known to be constant.

In this thesis, we address the problem of predicting peptide binding to MHC class II molecules. The varying lengths of binding peptides have become a major bottleneck to earlier motif detection techniques. We define the problem as a multi-objective optimization problem, which in addition to accommodating variable length peptides, allows incorporation of a priori knowledge of motifs and binding residues. Often times, with binding peptides, information of binding motifs (acquired through experimental means or through prediction models) are also available. In order to accommodate such a priori information, we introduce the problem of *guided-discovery of motifs* in this thesis. This thesis proposes two techniques to guide the discovery of

motifs, using consensus sequences and using anchor residues. Self-discovery of motifs uses only the information of binders and non-binders. Self-discovery of motifs is a special case of guided-discovery which uses only the information from binders and non-binders.

The proposed approaches are implemented by using an multi-objective evolutionary algorithm (MOEA) and a genetic annealing algorithm (GAA). The motifs derived from proposed MOEA and GAA approaches demonstrate excellent predictive performances for number of benchmark datasets. These methods also facilitate identifying a motif for a dataset that had difficulty in finding a consensus motif. In addition, MOEA and GAA derived motifs show superior generalization capabilities to those derived with number of computational techniques as well as to the experimentally determined motifs on other datasets. We also study the effects of using positive samples alone in building predictive models.

A characterization based on the abundance of physicochemical properties at the predicted binding sites is presented, where the relative abundance of physicochemical properties at the binding sites corroborates well with the earlier findings, supporting the applicability of the proposed methods to find binding motifs in a wide range of MHC class II alleles. Both self-discovery and guided-discovery techniques proposed in this thesis could lead to efficient way of future discovery of drugs and vaccine targets.

Author's Publications

Journals

1. Menaka Rajapakse, Lin Feng. Prediction and Characterization of MHC-Peptide Binding Sites. *IEEE Engineering in Medicine and Biology Magazine*, (in press).
2. Menaka Rajapakse, Bertil Schmidt, Lin Feng, Vladimir Brusic. Predicting peptides binding to MHC class II molecules using multi-objective evolutionary algorithms. *BMC-Bioinformatics*, Vol. 8, No. 459, November 2007.
3. Menaka Rajapakse, Guanglan Zhang, Srinivasan Kellathurn, Bertil Schmidt, Vladimir Brusic. $PRED^{NOD}$, a prediction server for peptide binding to the H-2^{g7} haplotype of the non-obese diabetic mouse, *Autoimmunity*, Vol. 39, No.4 pp. 645-650, December 2006.

Book Chapters

1. Menaka Rajapakse, Lin Feng. Predicting Binding Peptides with Simultaneous Optimization of Entropy and Evolutionary Distance. In: J. C. Rajapakse et al. (eds). *Pattern Recognition in Bioinformatics*, Second IAPR International Workshop, PRIB 2007, Singapore, *Lecture Notes in Bioinformatics*. Springer-Verlag. 4774: 349-355.
2. Menaka Rajapakse, Bertil Schmidt, Vladimir Brusic. Multi-Objective Evolutionary Algorithm for Discovering Peptide Binding Motifs. In: F. Rothlauf et

- al. (eds). *Applications of Evolutionary Computing, EvoWorkshops2006*, EvoBio 2006, Budapest, Hungary, *Lecture Notes in Computer Science*, Springer-Verlag. 3907: 149-158
3. Menaka Rajapakse, Lonce Wyse, Bertil Schmidt, Vladimir Brusic. Deriving Matrix of Peptide-MHC Interactions in Diabetic Mouse by Genetic Algorithm. In: M. Gallagher et al. (eds). *Intelligent Data Engineering and Automated Learning*, 6th International Conference, IDEAL 2005, Brisbane, Australia, *Lecture Notes in Computer Science*, Springer-Verlag. 3578 :440-447

Conferences

1. Menaka Rajapakse, Lin Feng. Guided-Discovery of Motifs for Peptide Binding Prediction”, *Research in Computational Molecular Biology (RECOMB)*, Singapore, March 2008.
2. Menaka Rajapakse, Lin Feng. Predicting Peptide Binders of Flexible Lengths with Genetic Annealing Algorithm”, *IEEE Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Hawaii, USA, April 2007.
3. Menaka Rajapakse, Bertil Schmidt, Vladimir Brusic. Pareto-Optimal Motif for Predicting Peptide Binders, *The 16th Intl. Conference on Genome Informatics (GIW)*”, Eds. Japanese Society of Bioinformatics, Yokohama, Japan, December 2005.
4. Menaka Rajapakse, Bertil Schmidt, Vladimir Brusic. An Evolved Consensus Motif for Class II MHC Molecule-IA^{g7} dataset, *Asia Pacific Bioinformatics Conference (APBC)*, Singapore, January 2005.

Chapter 1

Introduction

Human immune system acts as the body's defense mechanism against infectious and other diseases, and comprises of a number of organs (bone marrow, spleen, thymus, lymph nodes, etc.), tissues, cell types (lymphocytes, natural killer cells, macrophages, dendritic cells, etc.), and molecules (major histocompatibility complex, Interleukin-2, Interleukin-4, etc.). Infectious diseases are caused by foreign *antigens* found on the surface of *pathogens* (bacteria or virus). The antigens can also generate from the body itself, causing autoimmune disorders such as lupus erythematosus and rheumatoid arthritis.

The *lymphocytes* are the prime movers of the immune system, carrying antibodies throughout the body. Lymphocytes are categorized into two classes: B cells and T cells. While B cells are responsible for immune responses mediated by secreted antibodies, T cells are responsible for cell mediated immune responses. T cells of the immune system are the key mediators of immune responses against cancers, infectious diseases, or allergies. Binding of antigenic peptides to Major Histocompatibility Complex (MHC) molecules is a crucial step for initiation and regulation of immune responses. MHC molecules are categorized into two classes, class I and class II, based

on the mode and outcome of peptide presentation. MHC molecules bind and expose antigenic peptides so that they are recognized by T cell receptors (TCR) on the surface of T cells. Once recognized, an immune response is triggered against the infected cell or foreign agent.

On antigenic peptides, there exist immunogenic regions known as *epitopes* that bind to MHC molecules and facilitate the activation of immune responses. Antigenic peptides involved in MHC binding and trigger T cell recognition are known as T cell epitopes and are important for designing vaccines and immunotherapies. However, not all peptides bind to a given MHC molecule; only one in 100-200 peptides binds to a MHC molecule [5]. Thus, accurate identification of peptides binding to a specific MHC molecule is an important first step in determining potential vaccine candidates. This thesis addresses the problem of detecting binding peptides to MHC class II molecules.

1.1 Motifs and Peptide Binding

Motifs are sequence patterns conserved in DNA and protein sequences as a result of conservation through evolution, and associated with a biological function. Identification of these conserved patterns enables understanding of important biological mechanisms and specific structures underlying their function. Characterization of motifs requires thorough understanding of its biological function. A DNA motif is represented either by a *consensus sequence* or as a *weight matrix* as shown in Tables 1.1 and 1.2.

In protein sequences, conserved patterns sometimes called *fingerprints*. The presence of fingerprints among similar sequences are suggestive of their belonging to a

Table 1.1: Representation of a DNA motif with a consensus motif.

Motif instances	AGTCAAC
	AGTACAA
	AGTCAAC
	AATCAAC
	ACTAAGC
Consensus Motif	AGTCAAC

Table 1.2: Representation of a DNA motif with position specific weight matrix. P1 to P7 denotes the positions of the motif. The values indicate the likelihood of the presence of a particular nucleotide.

	P1	P2	P3	P4	P5	P6	P7
A	1.0	0.0	0.0	0.4	0.8	0.8	0.2
G	0.0	0.6	0.0	0.0	0.0	0.2	0.0
C	0.0	0.2	0.0	0.6	0.2	0.0	0.8
T	0.0	0.0	1.0	0.0	0.0	0.0	0.0

specific protein family. A protein motif is often indicative of a binding site as they tend to be well conserved regions in protein structures. Enzyme active sites, structural domains, cellular localizations tags, and MHC peptide binding sites are examples of protein motifs.

Peptide binding to MHC molecule occurs when a MHC molecule makes multiple contacts with the side-chains of an antigenic peptide. The contact sites define the *binding motif* and determine the specificity of binding [6]. A consensus motif describes amino acids (or nucleotides) at specific sites that influence binding while a matrix model employs position-specific information to describe the likelihood of binding motifs by indicating the strength of binding of nucleotides or amino acids at specific sites of the motif [7, 8].

1.2 Motif Discovery and MHC-Peptide Binding

Motif discovery is the process of finding the instances of a conserved pattern or given *motif* in a single or multiple sequences without the knowledge of the positions and constitutions of the instances. Presence of short sequence motifs may not be apparent from an alignment of the sequences. Therefore, advanced techniques are needed to identify the motifs because of the weak nature of instances, and to subsequently test for their presence in a given sequence.

With the influx of high throughput genomic sequence data from numerous genome sequencing projects, the need for computational analysis methods has escalated over the last decade. Devising computational methods for discovering motifs in genomes and primary sequences has drawn considerable attention from computational biologists in recent years [9, 10, 11, 12, 13, 14]. Using computational methods to isolate

potential targets identified as motifs in sequences can drastically narrow down the search space, thereby saving millions of dollars in wet-lab experiments and clinical testing.

However, though many computational techniques have mushroomed for motif recognition in unaligned sequences, a vast number of algorithms have focused mainly on DNA sequences and only a handful of techniques are available for finding motifs in protein sequences. Widely used general motif discovery tools (not an exhaustive list) in public domain, for both DNA and protein, or protein sequences alone are: Gibbs Sampler [9], MEME [10, 15], Teiresias [11], Block Maker [12], Pratt [13], and Gemoda [14]. Several of these methods have been recently benchmarked for their accuracy, limitations and potentials [16, 17].

Often, these tools are successful on a subclass or a set of subclasses of motif recognition problems because of the numerous challenges in the motif discovery process, and may not perform well across all classes of sequences. For example, a tool which is designed for TFBS detection cannot be directly applied to MHC peptide binding detection even though the underlying problem in both cases is discovering a binding motif. Similarly, due to the characteristics of the binding core of the MHC molecules, a tool designed for MHC class I sequences may not be applicable to detect peptide binding sites in MHC class II sequences. As a result, a large number of motif discovery tools have been developed. Moreover, the need to satisfy other factors such as time complexity, generality of the algorithm, and applicability to both DNA and protein sequences has further lead to a proliferation of motif discovery tools. Nevertheless, the need for new techniques arises when challenges to motif discovery emerge as new

datasets being made available from wet-lab experiments for previously unknown alleles.

Experimental approaches for predicting binding peptides involve cost and labour-intensive procedures involving synthesis of peptides and measuring their binding affinity, and therefore are not feasible for studies involving a large number of peptide sequences. Computational methods can assist in discovering statistically significant motifs compared to those discovered by biologists with pure experimental methods. For instance, when predicting binding peptides to MHC molecules, a motif discovered by a wet-lab experiment may well represent only the peptides similar to the ones used in the experiments thereby hampering the predictions of peptides of other datasets. On the other hand, motifs detected by computational approaches are more flexible in accommodating a wide array of previously unseen data, because of their better generalization capabilities.

Recent advances in bioinformatics have generated a considerable interest in computational tools for predicting MHC-peptide binding interactions. Such computational prediction tools help minimize wet-lab experiments and facilitate efficient scanning for potential T cell epitopes from large sets of protein antigens (e.g., viral genomes) by narrowing down the search space of potential targets [18]. However, these predictions do subsequently need to be verified by experimental methods involving biochemical and functional assays, to determine their *in vivo* relevance.

Computational techniques for prediction of binding peptides to both MHC class I and class II molecules have been extensively studied over the last two decades and success has been achieved with class I prediction models compared to class II molecules [19, 20]. Peptide binding prediction to class II MHC has been more difficult

due to the structural differences between MHC class I and class II binding groove where the binding takes place. The binding groove of the class I molecules is closed and therefore the binding is restricted to fixed length peptides (8, 9, or 10aa). The binding groove of class II MHC molecules on the other hand is open at both ends, enabling binding to peptides of length 9aa and longer (peptide termini will extend outside the binding groove). Therefore, in MHC class II, the exact location of the binding core (or motif) within the peptide is usually unknown in experimentally determined binders. As such, peptide binding prediction is often preceded by a motif discovery process.

The variable nature of lengths of binding peptides to MHC class II molecules introduces many challenges to computational prediction methods of peptide binding. Nevertheless, earlier investigations on analysis of structures on binding motifs have introduced that only a core of 9aa within binding peptide is essential for binding to MHC class II molecules [21][22][23]. Therefore, an approach that searches for the binding core region in binding peptides would give rise to best alignment, which is able to overcome the limitations posed by the variable length peptides.

Often a preprocessing step utilizing existing knowledge such as anchor position specific information is involved in obtaining an alignment prior to building a predictive model [24][25][26]. These methods use the first anchor position to limit the number of possible 9aa alignments. Only alignments bearing a reported amino acid at the first anchor position are considered to be valid for building predictive models.

1.3 Motivation

The prediction of peptide binding to MHC molecules plays an important role in immunology. There is a pressing need for developing immunotherapy, vaccines, and diagnostic tools for infectious diseases such as cancer, autoimmunity, and allergies. One of the hallmarks of peptide binding to MHC molecules is their ability to form stable complexes with peptides while simultaneously accommodating considerable variability of sequences. A given peptide binding groove can bind to hundreds or thousands of different peptides, identical or homologous at only a few side chain positions [21].

Biochemical assays, which empirically test protein-peptide binding affinity, can nowadays provide a high throughput output [27]. However, high costs, and laborious and tedious experimental work involved in T cell epitope screening in conventional methods have made researchers to seek fast and efficient computational prediction tools. Such tools allow researchers to identify allele specific binding peptides, thus reducing the number of peptides to be synthesized and assayed. However, the inherent complexities involved in recognition of pathogens and in activating immune responses pose many challenges to the prediction of epitopes.

T-cell epitope prediction is a grand scientific challenge: it is both difficult, and thus exciting and of true utilitarian value, as a pivotal tool in the drive to develop vaccines and diagnostics.

-Darren Flower;2003

To date, there exist a large number of MHC class II peptide binding prediction models [24, 25, 28, 29, 30, 31, 32, 33, 34, 35]. However, each method has its own constraints such as limiting its prediction capabilities to a fewer number of different MHC class II alleles [24, 25, 30, 35, 36, 37] relying heavily on anchor-position specific information [24], [34], [29], possible biases arising from testing on limited size datasets of high similarity [34] and lack of ability to generalize predictions well for previously unseen data, which is quite common in methods that have been trained with smaller datasets. Due to these limitations, there is a need for a better predictive models and methods to alleviate the shortcomings and realize optimal alignment of peptide binding cores across different types of alleles.

In addition, as mentioned before, there exist sub-optimal motifs resulting from wet-lab experiments because of limited number of used datasets. These sub-optimal motifs represent only dominant characteristics specific to a dataset and not the consensus binding motif of the allele of interest. Nevertheless, fusion of these related sub-optimal motifs can enrich the discovery of motifs while guiding the search for a consensus motif. Thus, we introduce a guided-discovery of motifs that employs motifs found in the experiments into the computational prediction of consensus motif. Theoretically, synergistic fusion of motifs derived from different experiments is known to give better motifs, and an optimal or near-optimal alignment for a consensus motif is envisaged.

Due to degenerated nature of the MHC class II motifs and different lengths of peptides in experimental datasets, it is vital to select a search algorithm that is capable of handling noisy and missing data problems. In a noisy problem, the fitness of candidate solutions approaches the fitness variance caused by the noise and a

good candidate solution might receive worse fitness than a bad candidate solution by chance [38]. As evolutionary algorithms (EA) use nature inspired heuristics for search, they are considered relatively robust to the noise present in the evaluation of the objective function [39]. Therefore, we choose EA based approaches; in addition to their optimization capabilities and robustness to noisy data, they are also effective in handling multi-modal, and discontinuous problems.

Often, EA are modeled as a single objective optimization problems where the set of decision variables is optimized to maximize or minimize a single goal. The use of single objective EA becomes insufficient when there is more than one conflicting objectives or otherwise are present. In solving motif detection problems, in addition to high accuracy high sensitivity and specificity are also expected. To accomodate multiple objectives as such, we model the motif detection problem as a multi-objective optimization problem. Multi-objective EA (MOEA) allows simultaneous optimization of number of objective functions. MOEA yields trade-off solutions that provide a good compromise between the different objectives, for example accuracy, sensitivity, and specificity. A number of recent MOEA approaches are described in [40, 41, 42].

In order to accomplish the above mentioned requirements, we propose two approaches for motif discovery: self-discovery and guided-discovery, allowing discovery of a binding motif in the absence or presence of prior knowledge, respectively. A number of self and guided discovery approaches based on EA are proposed and experimentally analyzed to determine the suitability of each method for predicting binding peptides.

1.4 Main Contributions

Specific contributions of this research are as follows:

- Introduced the problem of guided-discovery of motifs. Often times, with the experimental binding peptides, the information of experimental motifs are available. Previous research has not incorporated the information of experimental motifs in to the motif detection. Therefore, we define and provide solution to guided-discovery of motifs. In addition to the information from both binders and non-binders, the guided-discovery approach used the information of experimentally found motifs.

The motif is represented as a position specific scoring matrix. Two objective functions were proposed; one designed to maximize the sensitivity and specificity of predicting true binders and non-binders, and the other facilitates to incorporate prior knowledge available from experimental motifs.

- Self-discovery of motifs with previous probabilistic approaches uses only the information from binders. The proposed self-discovery approach of motifs uses the information available from both binders and non-binders. The objective functions are designed to maximize the sensitivity and specificity of predicting true binders and true non-binders, and high marginal score in separating two classes.
- To predict peptide binding motifs, we proposed a Genetic Annealing Algorithm (GAA) which employs a fitness function representing the weighted sum of multiple objective functions designed to maximize the sensitivity and specificity of predicting true binders and true non-binders, and to increase the separability

between the two classes. The proposed GA algorithm incorporates annealing into the cross-over operator and implemented as a single objective GA. The approach outperformed previous methods of finding a consensus motif through experimental means and by using existing motif detection methods for MHC class II, I-A^{g7}. This approach facilitates self-discovery of motifs from peptides of different lengths.

- The proposed methods are capable of finding peptides binding to MHC class II molecule, I-A^{g7} better than previous approaches. Finding a motif has been difficult for I-A^{g7} because of its promiscuity. Cross-validation experiments on motifs derived on I-A^{g7} datasets demonstrate better generalization abilities of the motif. Methods were further validated and compared on two publicly available benchmark datasets: (1) an ensemble of qualitative HLA-DRB1*0401 peptide data obtained from five different sources, and (2) quantitative peptide data obtained for sixteen different alleles comprising of three mouse alleles and thirteen HLA alleles. The proposed method outperformed earlier methods on most datasets, indicating that it is well suited for finding peptides binding to a wide range of MHC class II alleles.
- In our methods, we first introduced positive sample modeling with guided-discovery. Most previous methods have only used positive data (binders) to find binding motifs. The fitness functions of the EA approach were modeled to measure relative entropy and the evolutionary distance among amino acids in an alignment, and the EA process is guided by the anchor positions associated with MHC class II.

- Next, we introduced a new guided-discovery approach that uses experimentally validated motifs to drive the motif discovery. The fitness functions of the EA approach were modeled to measure the sensitivity and specificity, and the similarity to the experimentally found motifs.
- Also proposed is a motif search approach guided by motifs predicted by numerous computational predictive methods. The fitness functions adopted were similar to the case where experimental motifs were used to guide the motif discovery process.
- A method to characterize motifs at the predicted binding sites by using physico-chemical properties was proposed. The methods provide a better way to compare numerous motifs. The characterization clearly indicates whether the anchor positions of the binding sites determined are in agreement with the previous studies. Characterization of the binding motif of I-A^{g7} shows that the motif is degenerate and allows it to bind a wide selection of peptides.

1.5 Organization of the Thesis

This report consists of 5 chapters. The next chapter, Chapter 2, provides the background that leads to this research. It gives the relevance of MHC-peptide binding to vaccine discovery and development of immunotherapy, the importance of computer-aided tools for screening vaccine targets, and the areas where advances have been made. It also highlights the effects of the quality of data in motif finding problems. Existing motif prediction models and their advantages and shortcomings are discussed. Introductions to Evolutionary Algorithms (EA), multi-objective EA (MOEA)

and the motivation for using EA for predicting binding peptides are also given.

We introduce self-discovery problem in motif detection in Chapter 3. The application of the MOEA approaches to find a consensus motif for degenerated datasets is presented and discussed. The results are compared with earlier well-known motif detection methods. The method is tested and compared with earlier methods by using number of MHC-class II binding datasets.

Chapter 4 introduces the guided-discovery of motif detection. The application of MOEA based methods to find a consensus motif for degenerated datasets are presented. Finally, results are discussed and a performance comparison of our approach with other well-known motif detection methods is reported.

In Chapter 5 we discuss the use of physico-chemical properties to propose a new method to characterize motifs. Discrepancies among computationally derived and experimentally determined motifs are presented and discussed.

Finally, the last chapter contains the discussion of the advantages and disadvantages of the proposed methods, concluding remarks and a summary of the intended future work.

Chapter 2

Background

Immunoinformatics, a subfield of bioinformatics, is dedicated to understanding immune system with the aid of principles from computer science. The long term goal of immunoinformatics is to devise an *in silico* immune system by modeling the functionalities of components of human immune system. An *in silico* immune system can facilitate understanding a wide range of diseases and clinical conditions. An example of such an initiative is the ImmunoGrid (<http://www.immunogrid.org>) which aims at modelling the human immune system at molecular, cellular, and organ levels. Such technologies will facilitate simulation of processes of human immune system at different scales while providing tools for clinicians and researchers investigating immunological disorders. Furthermore, immunoinformatics research could lead to discovery of new vaccines, drugs, and immunotherapies.

There are two types of lymphocytes: T cells and B cells. While T cells involve in cell-mediated immune responses, B cells involve in the humoral (mediated by the secreted antibodies) immune responses. T cells are the key mediators of immune responses against infectious diseases, cancers, autoimmunity, allergies, and transplant rejections [43]. In addition to lymphocytes, major histocompatibility complex (MHC)

molecules play an important role in the immune system. A crucial step in T cell activation is the presentation of peptides derived from disease related protein antigen to MHC molecules. This process involves intra-cellular fragmentation of disease related protein antigen, followed by binding of derived peptide (known as epitopes) to MHC, which is subsequently presented on cell surface of antigen presenting cells (APC) for the recognition by TCR. The T cell activation process is illustrated in Figure 2.1.

Figure 2.1: Illustration of T cell activation through antigen binding to MHC and MHC-antigen complex presentation to TCR.

2.1 Immunology

Immunology is the study of the way in which the body defends itself against invading organisms (pathogens) or internal invaders (tumors) [44]. Immunoinformatics deals with special problems of immunology targeting at better understanding of the immune

system and related diseases such as allergies, cancer, and infectious diseases. One key goal of immunoinformatics is the development of computer-aided vaccine design (CAVD) for aforementioned diseases. Because of the deluge of biological information available from various resources such as genome databases, vaccinologists are turning to computer-aided tools for efficient and cost-effective screening of vaccine targets. These tools utilize sequence analysis techniques to identify putative protein fragments of infectious agents that can be used as candidates for vaccine design. Areas that have made advances in epitope-based vaccine discovery through immunoinformatics tools are melanoma [45], multiple sclerosis [46], malaria [47] and anti-tumor vaccines [48].

2.1.1 Immune System

The immune system acts as human body's defense mechanism against infectious foreign antigens (bacteria, fungi, parasites, or viruses). It is the second most complex biological system in the human body, which comprises several organs (e.g. bone marrow, spleen, thymus, and lymph nodes), tissues, cell types (lymphocytes, natural killer cells, macrophages, and dendritic cells) and molecules [49]. A species possess two types of immunity: innate (inborn) and adaptive (acquired) immunity. While innate immunity refers to basic resistance to non-specific diseases, the adaptive immunity refers to resistance to antigen-specific threats. An innate immune response is activated by chemical properties of an antigen and through the development of immune memory. The microorganisms causing diseases are fast recognized and destroyed by the innate immunity. An adaptive immune response is activated only when the innate defenses are failed. In order to trigger an immune response against an antigen, the encountered antigen must be first processed and recognized. Once the

antigen is recognized, the adaptive immune system generates a large number of immune cells capable of attacking the antigen and thereby the infectious disease. After the infection is cured, the adaptive immunity develops a memory that enables more rapid and efficient response against the specific antigen when encountered in future.

2.1.2 Protein Antigen

An antigen is a substance that stimulates an immune response, especially the production of antibodies. Antigens are usually proteins or polysaccharides, but could be any type of molecule, including small molecules (haptens) coupled to a carrier-protein. Antigens composed of proteins can trigger a stronger antibody response than other antigens. Based on their origins, antigens are classified into groups such as exogenous antigens, endogenous antigens, and autoantigens. Exogenous antigens are the antigens that have entered the body from external sources, for example by inhalation, ingestion, or injection. These exogenous antigens are taken into the APCs and processed into fragments. APCs then present the fragments to T helper cells (CD4+) by class II major histocompatibility molecules on their surface [44].

The endogenous antigens are antigens that are generated within the cell as a result of normal cell metabolism or because of viral or intracellular bacterial infection. The fragments are then presented on the cell surface in the complex with class I histocompatibility molecules. If activated cytotoxic CD8+ T cells recognize them, the T cells begin to secrete different toxins that cause the destruction of the infected cell.

Autoantigens on the other hand are usually normal proteins or complex of proteins (and sometimes DNA or RNA) that are recognized by the immune system of

patients suffering from a specific autoimmune disease such as diabetes or rheumatoid arthritis [50]. Under normal conditions, these antigens should not be targeted by the immune system, but mainly due to genetic and environmental factors, the normal immunological tolerance for such an antigen is lost in these patients.

2.1.3 MHC Molecules

Major Histocompatibility Complex (MHC), is a set of molecules displayed on cell surfaces that are responsible for the recognition of lymphocyte and presentation of antigen. At anytime, a human cell can display 50,000-100,000 human leukocyte antigen (HLA or human MHC) molecules, an a single cell can display up to 20 different HLA variants. The MHC molecules control the immune response through the recognition of foreign (non-self) and altered self (tumors) antigens. There are two main classes of MHC molecules, class I and class II, and each class can be further divided into variants called *alleles*. Alleles are alternative forms of the same gene.

MHC is encoded by several genes located on human chromosome 6 as shown in Figure 2.2 (source:<http://www.cehs.siu.edu/fix/medmicro/mhc.htm>). Class I molecules are encoded by three different genes on the genome referred to as A, B and C region while class II molecules are encoded by the D region. A genomic region between those class I and II regions on chromosome 6 encodes class III MHC molecules.

Class I genes are encoded by a polymorphic heavy chain which combines with β_2 - microglobulin and is expressed on the surfaces of all cells with a nucleus. The structure of binding groove in the heavy chain enables T cells to recognize bound peptides. Class II genes are composed of two dissimilar polymorphic polypeptide chains, that is α and β chains, both of which contribute to the peptide groove. The

Figure 2.2: The structure of HLA (human MHC) complex in chromosome 6.

overall protein structure of both class I MHC and class II MHC is the same in which a seven-stranded β -sheet represents the floor of the binding groove, and the sides are formed by two long α -helices (or continuous α -helical segments in the $\alpha 2$ or $\beta 1$ helices) that straddle the β -sheet. Polymorphic residues that clustered within and around the binding groove provide the variations in shape and chemical properties which account for specific peptide-binding motifs identified by each MHC allele [51].

Class I MHC Molecules

T lymphocytes recognize disease related foreign peptides presented by MHC class I (also known as class I HLA for human) and subsequently help in destroying infected cells. Class I MHC genes are highly polymorphic and hundreds of different alleles exist for each locus [43]. Therefore, HLA are highly diverse and each allele binds only to a specific set of peptides (approximately 1 in 200 peptides) [5]. The binding groove of MHC class I contains a set of pockets that accommodate side chains of binding peptide. The amino acid sites that secure the peptide to the MHC molecule are called *anchor* residues. Figure 2.3 illustrates a ribbon and schematic diagrams of MHC class I molecule structure. In class I MHC molecules, the peptide-binding site (called the $\alpha 1\alpha 2$ domain) is constructed from the heavy chain only, and an additional light chain subunit, $\beta 2$ -microglobulin ($\beta 2m$), associates with $\alpha 3$ of the heavy chain.

Figure 2.3: Illustration of peptide binding to MHC class I molecules with (a) a ribbon diagram, and (b) a schematic diagram

Class II MHC Molecules

MHC class II samples peptides derived from endocytosed proteins. Peptides presented by MHC class II molecules activate CD4+ helper T lymphocytes and stimulate cellular and humoral immunity against pathogens [43]. Genetic susceptibility to several autoimmune disorders such as insulin-dependent diabetes (IDDM or type I diabetes), multiple sclerosis, and rheumatoid arthritis are associated with the expression of specific MHC class II alleles[50]. MHC class II molecules consist of two chains: α chain and β chain. These polypeptides each have two extracellular domains: membrane proximal domains (α_2 and β_2) and NH2-terminal extracellular domains (α_1 and β_1) [50]. The class II MHC peptide-binding site is assembled from the heavy chains, α_1 and β_1 . This peptide binding site is a groove in which the sides are α helices and the floor is composed of several anti-parallel β -pleated sheets [50]. Figure 2.4 illustrates

Figure 2.4: Illustration of peptide binding to MHC class I molecules with (a) a ribbon Diagram, and (b) a schematic Diagram

a ribbon and schematic diagrams of MHC class II molecule.

2.1.4 Validation of Binding Peptides

Experimental validation of peptides that bind to MHC molecules is performed by T cell based bioassays or by adopting biochemical methods. Bioassays are carried out by measuring either the activation and proliferation of T cells or the ability of the T cells to kill target cells [51]. Experimental methods of analyzing T cells include cytotoxic T lymphocyte (CTL) assays, intracellular cytokine staining (ICS), enzyme-linked immunospot (ELLISPOT), and MHC tetramers. While CTL assays are used to detect T cell recognition of MHC class I bound peptides, proliferation assays are used to detect class II MHC-peptide bound complexes. Biochemical methods include stabilization assays and competitive inhibition assays [52].

Despite the possibilities of experimental validation, there is a little consensus on the best assay systems. And affinity data are often of poor quality due to the differences in measurements of the same peptide, mixtures of different standard peptides used in radioligand assays, and experiments conducted at different temperatures and over different concentration ranges [19]. Moreover, such methods are both cost and labor intensive. Therefore, experimental methods of validation are infeasible for studies involving large numbers of protein sequences. Computational tools of prediction of binding peptides are therefore sought to minimize the number of peptides that are to be synthesized and assayed experimentally.

Both quantitative and qualitative peptide data are available in databases such as IEDB [53] and AntiJen [54]. IEDB being the most recent one contains the largest set of peptide data. The JENPEP database provides quantitative MHC-peptide binding data [55]. Other databases comprising qualitative peptide data are SYFPEITHI [34], MHCPEP [52], MHCBN [56], and EPIMHC [57]. The SYFPEITHI database contains a collection of MHC class I and class II ligands and peptide motifs of humans and other species, such as apes, cattle, chicken, and mouse. Peptides available with SYFPEITHI database are known to be high binders [57]. The entries in the oldest database, MHCPEP, have been compiled from literature as well as from direct submissions of experimental data. The EPIMHC database has been developed to overcome drawbacks in [34][52][56] databases and only contains MHC binding peptides that occur in actual proteins.

2.1.5 MHC-Peptide Binding Motif

The differences between the grooves of class I and class II MHC peptide binding attribute to their specific vital biological functionalities. The binding cleft of Class I MHC molecule allows binding of peptides of fixed lengths (8, 9, or 10aa) [44]. As opposed to MHC class I, the binding groove of Class II MHC is open at both ends allowing binding to peptides of broad length distribution (approximately 9-20aa) [44]. However, existing evidence suggest that a minimum of 9aa core region is generally recognized by the MHC class II molecules [21][22][23]. The non specific length distribution in class II binding peptides have complicated the building computational approaches of prediction models. Moreover, motif signals present in peptides that bind to MHC class II molecules are weak and often degenerated. Obtaining a correct alignment of a binding core region is crucial for identifying the binding motif and subsequently identifying the binding peptides.

2.2 Motif finding Methods

The most fundamental to all computational approaches of motif detection is the profile analysis [58][59]. Two other popular methods are Gibbs motif sampler [9][60] and MEME [10][15]. In general, profile analysis is applicable only when an alignment of a set of preselected sequences is possible. On the contrary, stochastic approaches such as Gibbs motif sampler [9][60] and MEME [10][15] are able to detect motifs in aligned or unaligned set of sequences. In the following subsections, we will describe existing motif detection methods that can handle datasets consisting of variable length sequences. An introduction to profile analysis is also given as it provides an insight into the

definition of a motif. Also the possibility of building profile matrices by various other ways exists such as using the values obtained from biochemical assays performed in wet-lab experiments, an output model of a classifier or a search algorithm.

2.2.1 Profile Analysis

Profile analysis involves building profiles (also known as templates, weight matrices, position specific scoring matrices, etc.) from a set of aligned sequences of protein or DNA sequences. The underlying tenet of profile analysis is that certain positions are more conserved than the others. The conserved regions or *consensus* can therefore be extracted by aligning a set of sequences of a given family. The residues in the consensus sequence correspond to the highest weight values in the columns. A description of building a profile from a given set of sequences is given below.

Given a set of sequences of the same family, sequences are first aligned so that the most similar regions (motif patterns) among the sequences are superimposed. The next step is to create a profile based on this alignment - a profile is a frequency matrix formulated by taking the ratio between the frequency of each residue in every motif position and the frequency of the respective amino acid in the background. Once the profile is built, a new sequence can be tested against the profile to see if the sequence contains the motif belonging to the specific family of the proteins. This is carried out by sliding a window along the test sequence and evaluating a score for each window according to a defined scoring function.

Let Q be the profile matrix formulated using a set of aligned sequences of a protein family where the residue at site i is represented by $a_i \in \Omega$, where the amino acid alphabet is denoted by Ω . A profile of a motif of length k , can be illustrated

q_{1a_1}	q_{2a_1}	q_{3a_1}	$\cdots$	$\cdots$	$\cdots$	$\cdots$	q_{ka_1}
q_{1a_2}	q_{2a_2}	q_{3a_2}	$\cdots$	$\cdots$	$\cdots$	$\cdots$	q_{ka_2}
q_{1a_3}	q_{2a_3}	q_{3a_3}	$\cdots$	$\cdots$	$\cdots$	$\cdots$	q_{ka_3}
$\vdots$	$\vdots$	$\vdots$					$\vdots$
$q_{1a_{ \Omega }}$	$q_{2a_{ \Omega }}$	$q_{3a_{ \Omega }}$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$q_{ka_{ \Omega }}$

Figure 2.5: Illustration of a profile matrix

as shown in the Figure 2.5 where q_{ia_j} denotes the i_{th} position specific score of amino acid $a_j \in \Omega$.

A scoring function for estimating the score s for a window of length k of sequence i , $x_i = (x_{i1}, x_{i2}, \cdots, x_{ik})$ is defined as

$$s = \sum_{i=1}^k \sum_{j=1}^{|\Omega|} \log q_{ia_j} \delta(x_i = a_j) \quad (2.2.1)$$

where the Kronecker $\delta(\cdot) = 1$ when the argument is satisfied, and otherwise 0. A threshold is used to establish the presence of a motif in a given sequence.

Let us consider an example for detecting a possible motif pattern in a test sequence GEEVER, given an alignment of a specific sequence family as shown below:

G	E	V	K
G	D	I	R
G	E	I	K
G	D	V	E

Using the alignment we can build a profile for the given sequence family. Assuming that the amino acids in the background are distributed equally, we can estimate a score for each residue for occupying a specific position in the alignment. For example,

the value for residue G in the position 1 in the profile, $q_{1G} = (4/4)/(1/20) = 4$. To detect motifs in the given test sequence, a window of size $k=4$ is shifted to the left by one position at a time along the test sequence of length l . During each shift, a score is calculated according to the equation 2.2.1 where number of scores corresponds to the number of windows occupied by the test sequence which is equal to $l - k + 1$. A score estimated for a window can be represented as:

$$s(\text{GEEV}) = \text{score1G} + \text{score2E} + \text{score3E} + \text{score4V}$$

$$s(\text{EEVE}) = \text{score1E} + \text{score2E} + \text{score3V} + \text{score4E}$$

$$s(\text{EVER}) = \text{score1E} + \text{score2V} + \text{score3E} + \text{score4R}$$

where $\text{score1G} = \log((4/4)(1/20))$ and $\text{score3V} = \log((2/4)(1/20))$ and so on. The final score is tested against a threshold in determining the presence of a motif in the given test sequence. The profile with the highest score is taken as the most likely motif. While profile analysis is applicable to both protein and DNA data, and also for peptide binding prediction, it can be applied only when an alignment of sequences is possible.

Profile Based Methods

RANKPEP [61] and SYFPEITHI [34] are examples of peptide binding prediction methods based on profile analysis. Profiles are generated by analyzing the amino acid frequency in the binding peptides during pool sequencing [34] by revealing positional amino acid preferences with the use of combinatorial peptide libraries and sequence alignment of binding peptides [61]. RANKPEP provides position specific scoring matrices (PSSM) for predicting peptide binding to both MHC class I and class II molecules. The PSSMs have been derived using motif block alignments of binding

peptides to specific MHC class I and class II molecules. While the motif discovery program MEME [10][15] had been used to obtain the motif block alignments, the PSSMs had been obtained by using BLK2PSSM utility available in the BLIMPS package [62]. SYFPEITHI on the other hand uses PSSM constructed by using domain expert knowledge. Profile analysis had been carried out with the aid of a scoring scheme that assign weights to the amino acids at each position in the alignment. While higher positive weights have been assigned to frequently occurring amino acids in anchor positions and amino acids present in a significant number of ligands, medium weights have been given to amino acids occupying auxiliary anchor positions, rarely occurring amino acids, and preferred amino acids. Amino acids that are considered as unfavorable for binding have been given negative weights. SYFPEITHI contains a number of PSSM based motifs derived for both MHC class I and class II alleles. While both methods are design to predict peptides binding to class I and class II, RANKPEP has the advantage over SYFPEITHI web server as it offers predictions to a large number of MHC class II alleles compared to SYFPEITHI web site, in which the number of MHC class II molecules that are targeted for peptide prediction is very limited.

2.2.2 Gibbs Sampling

Gibbs sampling method is a stochastic approach for optimization, which aims at representing the solutions in a multi-variable joint distribution and finding the optimal solution by iterative stochastic sampling [9, 60]. This is performed by splitting the original joint distribution into a set of conditional distributions from which samples are drawn easily. In stochastic approaches, the solutions are considered as random

variables. Therefore, let the position specific scoring matrix Q be represented by random variable θ_i for clarity.

Given a joint distribution of solutions $(\theta_1, \theta_2, \dots, \theta_n)$, $p(\theta)$ a set of conditional distributions, $p(\theta_i | \theta_{-i})$ are implemented where θ_{-i} represents variables exclusive of θ_i , for drawing samples. The sampling procedure can be repeated until a predefined criterion is met [9]. Gibbs Sampling is a probabilistic approach widely applied for motif detection is sensitive to noise (i.e., presence of sequences that do not contain the motif). For a given model m of motif, let the random variable $\theta = (\theta, \theta_0)$ where $\theta = (\theta_{i1}, \theta_{i2}, \dots, \theta_{i|\Omega|})$ denotes random variables representing the motif and θ_0 represents background distribution.

Motif Site Sampler

Motif site sampler, developed by Lawrence and his colleagues [9] can be applied to find patterns in protein sequences, DNA sequences, or peptide binding predictions alike. The algorithm is initialized by choosing random starting positions of plausible motifs in each sequence in the dataset, followed by two main steps applied iteratively.

Step1: Predictive update step; For a random sequence in the dataset, motif weight matrix model, θ is constructed for a motif m with length k by the motif instances in the rest of the sequences by using Eq. 2.2.2 as below:

$$\theta_{ia} = \frac{q_{ia} + g_i}{N - 1 + G} \quad (2.2.2)$$

where $a \in \Omega$ and θ_{ia} is the frequency of a in position i of the motif instances alignment excluding the instance in x_i . q_{ia} denotes the binding strength of the site i when it is occupied by amino acid a in PSSM Q . Pseudo counts are represented by g_i and G ,

and sequence count is denoted by N .

Step2:Sampling step; The new starting position of the motif, j is chosen according to the weighting scheme. Each segment with length k is calculated according the motif model generated, $\theta_i = (\theta_{i1}, \theta_{i2}, \dots, \theta_{i|\Omega|})$. The frequency of the background distribution by θ_{i0} . Let's assume a sequence $x_i = (x_{i1}, x_{i2}, \dots, x_{in_i})$ and the motif begins at site j . The likelihood of the motif given the motif m is given by

$$P(m|\theta) = \prod_{i'=j}^{j+k-1} \theta_{ia} \quad (2.2.3)$$

The background probabilities are computed by finding the frequencies of amino acids in the background.

$$P(m|\theta_0) = \prod_{i=j+k-1}^{n_i} \theta_{ja} \quad (2.2.4)$$

where n_i is the length of the sequence x_i . A relative weighting score is assigned to each subsequence starting at position j in x_i by the likelihood ratio as below:

$$s_j = \frac{P(m|\theta)}{P(m|\theta_0)}, j = 1, 2, \dots, n_i - k + 1 \quad (2.2.5)$$

The starting position of the new motif in x_i is chosen by sampling from the normalized weight score, $\tilde{s}$ given as:

$$\tilde{s}_j = \frac{s_j}{\sum_{j=1}^{n_i-k+1} s_j} \quad (2.2.6)$$

Step 1 and *step 2* are iteratively applied to the dataset until convergence is achieved or a user specified number of iterations are reached.

Several strategies have been proposed to overcome deficiencies associated with the above algorithm. Given below in details are the improvements that can be utilized

to overcome those drawbacks. The phase shift problem can arise when the algorithm is trapped in a sub-optimal position of a stronger motif (i.e. phase shift from the optimal motif position). This can be resolved by running each iteration multiple times, and then by comparing the current starting position of the motif with the positions shifted left or right by a specific number of positions until the optimum value is determined. For determining the motif length (when not known), the algorithm must run for multiple times for a range of values, and the best length is determined based on the criterion of information content per parameter.

Gibbs Motif Sampler

Gibbs Motif Sampler [60] based on the Bayesian approach is an extension of motif site sampler. This algorithm assumes sequences in the dataset as independent and hence adopts the product multinomial distribution, to model the motifs assuming independence of the motif positions. The background is described by a simple identical and independent distribution (i.i.d) sequence model or low order Markov model. The knowledge of the amino acid pattern of the motif and the location of the motif within the sequences in the dataset are usually unknown. The Bayesian approach takes these locations as missing data and estimates these information as unknown parameters.

The dataset can be modeled using a set of parameters θ where θ_0 represents the unknown parameters of the background and $\theta_i = \{\theta_{ia}\}$ represents the parameters of the motif, $a \in \Omega$. A product Dirichlet distribution $f_{\theta}(\theta)$ is introduced as the conjugate prior for θ , and a Dirichlet prior $f_{\theta_0}(\theta_0)$ for θ_0 . Let Z denote the index for the motif positions in the dataset and x denotes the sequence dataset, and the

predictive product formula can then be given as:

$$\pi(Z|x) \propto \pi(x, Z) = \int \int \pi(x, Z|\theta_0, \theta) f_{\theta_0}(\theta_0) f_{\theta}(\theta) d\theta_0 d\theta \quad (2.2.7)$$

Since $\pi(a_j|Z[-j], x) \propto \pi(Z|x)$, the probability of the element in position a_j in a sequence x_i is calculated as proportional to the likelihood ratio of element being a site to the element being a non-site.

The data structure described in the above models assumes the motif as a contiguous block of amino acids of length k . A fragmentation model is used to model the motif when gaps are allowed within the sites and for the choice of motif width [60]. This model adopts a sampling strategy to select $k < L$ positions, which has better conservation than the rest of the positions within the aligned block of length L . In order to relax the one motif instance per sequence requirement, the algorithm adopts a Bernoulli sampling strategy. There exist several other variations and extensions of Gibbs Sampling approach. Some of these approaches are; Unified Gibbs method [63, 64], Motif Sampler [65, 66].

2.2.3 Expectation Maximization Algorithm

Expectation maximization (EM) is an iterative optimization algorithm for estimating maximum likelihood. Given a sequence dataset $x = \{x_i : i = 1, 2, \dots, N\}$, the Bayes equation can be written as

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \quad (2.2.8)$$

where $P(\theta|x)$ is the posterior distribution of the model characterized by the parameter θ , and $P(x|\theta)$ is the likelihood of dataset x . The terms $P(\theta)$ and $P(x)$ describe the prior of the model of the parameters and the normalizing component which is not

usually involved in the optimization step. With respect to the motif recognition problem, θ acts as the parameter for describing the motif. The EM algorithm tries to find the θ by introducing a set of hidden variables, H that corresponding to the starting location of the motif:

$$\theta^* = \arg \max_{\theta} P(x|\theta, H) \quad (2.2.9)$$

The Eq. 2.2.9 is evaluated by estimating the hidden parameter H and model parameter θ in an iterative manner. The optimization is achieved by applying low bound maximization approach [67] by adopting the following three steps.

1. Expectation step (E-step): initialize θ randomly or heuristically. An optimal lower bound formula that obtained true likelihood function at the current estimation is constructed as the expectation with respect to the hidden variable H and using the value of θ^t in the current step, t .
2. Maximization step (M-step): the lower bound formula derived in E-step is maximized with respect to the parameter θ to get the updated parameter θ^{t+1} .
3. Repeat E and M steps until converged.

The EM algorithm is expected to achieve at a local optimum value of θ .

Multiple EM Elicitation (MEME)

Multiple EM Elicitation is a motif finding algorithm, which extends the EM algorithm for motif detection[10, 15]. MEME is applicable to both DNA and protein sequences, assuming that the subsequences share similar patterns due to the same inherent biological functions. MEME is able to find one or more motifs in a given sequence.

MEME adopts EM to fit a two component model, i.e., background and motif models to the datasets. MEME allows using multiple initialization to avoid solutions to trap in local minima.

The MEME algorithm models the datasets using the parameter structure represented as $(\Lambda, \theta, \theta_0, Z)$, where $Z = \{Z_{i1}, Z_{i2}\}$, $i = 1, 2, \dots, n$, $Z_{ij} = 1$ if sample i belongs to motif ($j = 1$) or background ($j = 2$); $\theta = \{\theta_1, \dots, \theta_N\}$ and θ_0 are the *motif model* and *background model*, respectively. $\Lambda = (\Lambda_1, \Lambda_2)$ is the probability for a certain sample belonging to motif or background. Additional parameters are the erasing factors that are used to find multiple motifs in the same dataset. The likelihood of the dataset, $x = \{x_1, x_2, \dots, x_N\}$, is given as:

$$L(\theta, \Lambda | x, Z) = P(x, Z | \theta, \Lambda) = \prod_{i=1}^N \prod_{j=1}^2 [P(x_i | \theta_j) \Lambda_j]^{Z_{ij}} \quad (2.2.10)$$

The EM algorithm iteratively maximizes the expected log-likelihood over the values of missing data Z , given the estimations of θ and Λ , and maximizes the function relative to θ . The E-step and M-step are given below.

E-step: calculates the expectation formula over the hidden variable Z (the motif start position in each sequence). Let θ^0 and Λ^0 be the current estimates of the parameters. The expectation formula can be given as:

$$E_{Z|x, \theta^0, \Lambda^0} [\log(L(\theta, \Lambda | x, Z))] = \sum_{i=1}^N \sum_{j=1}^2 E(Z_{ij} | x, \theta^0, \Lambda^0) \log(P(x_i | \theta_j) \Lambda_j) \quad (2.2.11)$$

M-step: from Eq. 2.2.11, we can estimate the value of Z^0 , and then by maximizing this equation over the model parameters θ and Λ , we can get the updated values of these model parameters. These two steps are iterated until the convergence

is achieved, i.e., the change of θ falls below a specified threshold or until a user defined number of iterations is reached.

Although MEME algorithm seems to perform well on many motif detection problems, it does possess some limitations. These limitations are attributed to the facts that EM algorithm is a local optimum method, and therefore it is likely to get trapped in a local optimum when heuristic points are chosen, especially for large datasets.

2.3 Methods of MHC-Peptide Binding

Two main avenues, namely sequence based methods and structure based methods have been explored for developing MHC peptide binding prediction tools. Moreover, these methods have been further subdivided into quantitative and qualitative prediction into two main classes based on how prediction results are presented [24]: (a) quantitative predictions and (b) qualitative predictions. The quantitative predictions predict inhibitory concentration (IC₅₀) values, qualitative predictions predict binding status (binder or non-binder, or if a binder is a high, medium, or low binder) of a peptide based on predictive score. In contrast to qualitative prediction methods, only a few quantitative methods have been developed to date ARB [32], SVRMHC[24], MHCpred [35], SMM-align [33].

The ARB approach treats all positions in the peptide equally and relies on automated peptide alignments to derive an overall peptide profile. While this inclusion of weakly interacting positions can be an advantage to predict overall peptide binding, it may lower the accuracy when picking the correct binding core. The ARB method developed at IEDB [68] provides a total of 85 and 13 MHC class I and class II matrices and is trained on quantitative peptide binding data available within the IEDB

database [68]. Both the SVRMHC and MHCpred methods are trained on relatively small sets of quantitative peptide binding data in the AntiJen database [54] and their performance could probably be improved by retraining with more data. The SVRMHC method covers six, and MHCpred three MHC class II alleles, respectively. The SMM-align method is designed to identify a weight matrix that optimally reproduces the measured IC50 values for each peptide in the training set. Results reported in [33] demonstrate better predictive performance when compared with the other methods listed above and is suggestive of the importance of correctly aligned cores in predicting binding peptides. However, a fair comparison of performance among peptide binding prediction methods is difficult to make due to the overlap among peptide databases that are being used in training these methods. A summary of MHC-peptide binding prediction methods (qualitative, quantitative, stochastic, and evolutionary) is given in Table 2.1.

2.3.1 Sequence Based Methods

Sequence based methods use regular-expressions to depict consensus patterns. Peptide binding predictions are performed through sequence motifs [22] derived from alignment of peptides. Sequence motifs describe the presence or composition of a motif and can be experimentally determined from a large pool of putative binding peptides [22][29]. These methods exploit amino acids at specific sites (anchor positions) that contribute significantly to the binding. Amino acids appearing in anchor positions are referred to as *anchor* residues. For instance, an HLA-A2 allele peptide

motif of nine residues has Leu at the second and Val at the ninth position as its anchor residues [5]. However, such simple and partial information extracted from binding peptides is insufficient to describe complex MHC-peptide binding interactions. Short-comings of the motif based methods have led to the modeling of matrix-based methods.

Matrix Models

Matrix models employ position-specific information to describe the nature and strength of binding motifs [7][8]. Each column in the matrix corresponds to a position of a certain length of peptides, and each row corresponds to an amino acid [69][70]. As such, matrix based methods give a quantitative score to each peptide, enabling the ranking of the predicted peptides based on their binding affinity [71]. An advantage of using the matrix-based method compared to the motif-based approach is that it covers a wider range of peptides with binding potential. A major drawback in deriving matrix models is that generation of the matrix for each MHC allele requires a large number of experimentally validated peptides. Although quantitative matrices estimated from experimentally derived position specific matrix models have demonstrated reasonable performance in predicting binding peptides, derivation of such matrices are practically costly. And more importantly, they lack the flexibility of data-driven machine learning methods that allows refinement of solutions in an iterative manner as and when additional data become available [26]. Examples of matrix models based on computational methods are: RANKPEP [61], ARB [32], SMM-align [33], and PERUN [26]. The methods SYFPEITHI [34] and TEPITOPE [72] are examples of matrix models derived based on experimental evidence.

On the contrary to other experimentally derived matrix based methods that measure all possible pocket profiles on a given allele to determine a single allele matrix which requires hundreds of individual peptides and thousands of peptide binding assays, TEPITOPE assumes that pocket profiles are nearly independent of the rest of the binding groove, and once a pocket profile is determined experimentally, it is shared among other alleles as long as their amino acid residues contributing to the pockets are identical. This facilitates the assignment of a small number of pocket profiles to a large number of alleles through sequence comparison and subsequent derivation of virtual matrices. TEPITOPE covers 50 different HLA-DR alleles.

Classifiers

Classifiers such as artificial neural networks (ANN) [26, 73, 69, 74], hidden Markov models (HMM)[5, 31], and support vector machines (SVM) [24, 75] have been employed to discriminate binding peptides from non-binding peptides when peptides of equal length are given. In situations where datasets comprised of variable length peptides, additional preprocessing steps must be performed in order to get an alignment before a classifier is employed [26, 76, 75].

2.3.2 Structure Based Methods

An alternative approach to sequence based prediction methods is the use of structural information for peptides binding prediction [20][77][78][79][80]. Structural approaches rely on the structural conservation observed in crystallographically solved peptide-MHC complexes. An advantage of using structure based methods over sequence based methods is that a large number of experimentally validated peptides

for each allele is not required apriori compared to the sequence-based methods. However, structure-based approaches are applicable only when three-dimensional (3-D) structures of protein-peptide complexes are known or when building of reliable molecular models is feasible. Structural approaches use binding energies of protein-peptide complexes to determine peptide binding [80]. Methods that employ 3-D structures to predict MHC-peptide binding include protein threading, homology modeling and docking [20]. Despite growing interest and promising results, reliability of structural methods fall still behind the strong data-driven sequence based methods.

Protein Threading

Protein threading assesses the compatibility of different peptides to fit into the binding groove of a MHC molecule. By threading the peptide through their known crystal structure templates, an estimate for the binding energy based on the interactions between the peptide in the groove and MHC residues is obtained. A solved structure, or a structural model based on a similar MHC allele, provides the information for generating a list of MHC contact residues for each peptide position [81][80].

Homology Modeling

In homology modeling, the unknown structure of a related amino acid sequence is predicted using the known homologous protein structures. In the context of peptide-MHC prediction, the aim is to model the bound conformation of a peptide sequence with an unknown structure, given the 3D structure of other bound peptides to homologous MHC molecules [20].

Docking

Computational methods for the detection of protein ligand binding or docking have become an area of increasing interest. Docking facilitates investigating intermolecular interactions. In general, the purpose of docking simulation is 2-fold: (1) to find the most probable translational, rotational and conformational juxtaposition of a given ligand-receptor pair, and (2) to evaluate the relative goodness-of-fit or how well a ligand can bind to the receptor [82].

Table 2.1: MHC-peptide binding prediction methods. (The list provided is by no means an exhaustive list)

Category	Method	Reference
Stochastic	MEME	Bailey <i>et al</i> , 1995
	Gibbs Sampler	Liu <i>et al</i> , 1995
Evolutionary Approaches	PERUN	Brusic <i>et al</i> , 1998
	Ant colony search	Karpenko <i>et al</i> , 2005
	MDGA	Che <i>et al</i> , 2005
Other methods	SYFPEITHI	Rammensee <i>et al</i> , 1999
	PROPPRED	Chambers <i>et al</i> , 2001
	MHCPred	Godkin <i>et al</i> , 2001
	HMM	Noguchi <i>et al</i> , 2002
	RANKPEP	Reche <i>et al</i> , 2004
	MHC2PRED	Raghava <i>et al</i> , unknown
	ARB	Bui <i>et al</i> , 2005
	Iterative Learning Model	Murugan <i>et al</i> , 2005
	Predictive Bayesian Neural Network	Burden <i>et al</i> , 2005
	MULTIPRED	Zhang <i>et al</i> , 2005
	Kernel Approach	Salomon <i>et al</i> 2006
	SVRMHC	Wan <i>et al</i> , 2006
	MHC-BPS	Cui <i>et al</i> , 2007
	SMM-align	Nielsen07 <i>et al</i> , 2007

2.3.3 Modeling Issues

The accuracies of prediction models are still far from the level of biologists' acceptance levels. Often, prediction models rely on the quantity and quality of available experimental data. The drawbacks associated with building highly accurate prediction models are due to the lack of sufficient positive samples for training and validating computational methods, flaws and inconsistencies associated with binding data, biases and lack of diversity introduced during the pre-selection of datasets, and the small dataset sizes. Therefore, care must be exercised to eliminate or minimize biases that may arise from imbalance and redundant datasets. Below we investigate the issues related to peptide data which have implications on the performance of prediction model.

The quality of a prediction model usually depends on the quantity of the available peptide binders to specific alleles used for building the model. Motifs derived from experimentally determined small datasets indicate bias towards the derivative datasets and are usually poor performers. Structure-based predictive techniques become the viable choice of model when little data are available [20]. However, application of structure based techniques is constrained by the inherent complexities involved in building models, underfitting to data and computational speed. As the number of known peptide binders increases, sequence-based predictive techniques become more useful predictors.

Previous studies have reported existence of potential errors and inconsistencies of binding data in repositories [19]. This can be attributed to many factors such as inability to verify binding data independently [54] across many databases as well as discrepancies in assays used in studies. Noise and errors in datasets also have

an undesirable effect on building predictive models. Earlier studies carried out to investigate the impact of noise in datasets have revealed that 5% of errors in a dataset will usually double the number of data points required to build a matrix model [70]. However, no significant increase in error has been observed with robust classifiers such as ANN that can handle noise and incomplete data [83].

Another aspect that have a direct impact on the construction of predictive model is the diversity of data. Motif models derive from data lacking diversity show biases and lack generalization ability performing poorly on previously unseen data. Care must also be taken to avoid overfitting which occurs when the predictive model adapts too well to the training data and includes random distributions in the training set as being significant, affecting model performance [20]. The overfitting problem can be overcome by adapting [26, 84].

2.3.4 Performance Measures

In practice, there are number of different measures to evaluate the performance of predictive algorithms. The accuracy of performance depends not only on the correct predictions of positives (binders), but also on the negatives (non-binders) in an independent testing dataset. A set of performance measures can be defined by taking into account outputs predicted by the prediction method and using the experimentally observed as the ground truth. By thresholding the predicted scores, the presence of a motif is determined. If the predicted score of a binder is equal or higher than the threshold, it is classified as a true positive (TP), or as a false negative (FN) otherwise. Similarly, a true negative (TN) is a non-binder acquired a score lower than the threshold. A non-binder which scores a value equal or higher than the threshold

is labeled as a false positive (FP). Two commonly used performance measures are sensitivity and specificity and can be defined as follows:

$$Sensitivity = \frac{TP}{TP + FN} \quad (2.3.1)$$

$$Specificity = \frac{TN}{TN + FP} \quad (2.3.2)$$

Area Under Receiver Operator Characteristics Curve

Receiver operator characteristics curves (ROC) are used to overcome the dependence of performance measures on the parameters of the method. For example, the above measures depend largely on the chosen threshold values to separate predictive positives and negatives. This can be overcome by plotting the ROC curve by evaluating sensitivity and specificity values taken for a range of thresholds. The area under the curve (AUC) measures the power of the prediction method and used as a robust measure. While an AUC value close to one indicate a very high correlation, a value close to zero indicates negative correlation, and a value of 0.5 indicates that the prediction is equivalent to random guessing. In general an $AUC > 0.7$ indicates useful performance and a value > 0.85 a good prediction of a particular method [85]. The AUC values are good for objective comparison of different methods because they are independent of underlying parameters of different methods.

Mathew's Correlation Coefficient

Mathew's Correlation Coefficient ($\mathcal{C}$) is defined as:

$$\mathcal{C} = \frac{TP \times TN - FP \times FN}{\sqrt{((TP + FN)(TN + FP)(TP + FP)(TN + FP))}} \quad (2.3.3)$$

The Mathew's correlation coefficient ranges from $-1 \leq \mathcal{C} \leq 1$. A value of $\mathcal{C} = 1$ indicates the best possible prediction, in that every peptide was correctly predicted, and only true binding quality of peptides were predicted. On the other hand, a Matthew's correlation coefficient of $\mathcal{C} = -1$ indicates the worst possible prediction (or anti-correlation), where not a single peptide was correctly predicted. A Matthew's correlation coefficient of $\mathcal{C} = 0$ is equivalent to random predictions.

F-Measure

The F-Measure is first introduced in information retrieval and attempts to capture the trade-off between precision and recall. The concepts of precision and recall directly correspond to specificity and sensitivity. The F-measure is described as below.

$$F = \frac{2 \times \text{specificity} \times \text{sensitivity}}{\text{specificity} + \text{sensitivity}} \quad (2.3.4)$$

Equal Error Rate

The equal error rate (EER) where false acceptance rate (FAR) and false rejection rate (FRR) intersect, helps to determine the best threshold for classifying two classes. In the context of peptide binding prediction, the two classes are defined as binders and non-binders.

$$FAR = \frac{FP}{TN + FP} \quad (2.3.5)$$

$$FRR = \frac{FN}{TP + FN} \quad (2.3.6)$$

2.4 Evolutionary Approach

In this section, we give a general introduction to evolutionary approaches in the context of constrained and unconstrained optimization problems. In Chapter 3 and Chapter 4 we give existing evolutionary approaches for MHC-peptide binding prediction.

The goal of an optimization algorithm is to search for an optimal solution to a given objective, often subjecting to a set of constraints. An optimization problem consists of an objective function, f a set of decision variables $\mathbf{q}$ where $f(\mathbf{q})$ quantifies the quality of the candidate solution $\mathbf{q}$, and a set of constraints. The objective function represents the quantity to be optimized (minimized or maximized) and the constraints restrict the space of the possible solutions.

Most optimization problems often consists of boundary constraints that represent the domain of each variable. Problems that do not restrict the space of the solutions, except for boundary constraints are known as unconstrained optimization problems [86]. For those problems, the entire search space $\mathcal{S}$ is considered as the *feasible space* $\mathcal{F}$. However, many real-world problems require constraints be placed on the search space, excluding certain candidate solutions from being considered as solutions. As such, these constraints specify regions of the space that are infeasible. The aim of the constrained optimization technique is therefore to find solutions that do not lie in infeasible regions by satisfying all the stipulated constraints. A solution that satisfies all the constraints defined by the objective functions is called a *feasible solution*. Assuming a minimization problem, definitions of both unconstrained and constrained optimization problems are given below.

Unconstrained Optimization:

$$\begin{aligned} & \text{minimize} \quad f(\mathbf{q}), \quad \mathbf{q} = (\mathbf{q}_1, \dots, \mathbf{q}_M) \\ & \text{subject to} \quad \mathbf{q}_j \in \text{dom}(\mathbf{q}_j) \end{aligned}$$

where $\mathbf{q} \in \mathcal{F} = \mathcal{S}$, and $\text{dom}(\cdot)$ is the domain of the variable.

Constrained Optimization:

$$\begin{aligned} & \text{minimize} \quad f(\mathbf{q}), \quad \mathbf{q} = (\mathbf{q}_1, \dots, \mathbf{q}_M) \\ & \text{subject to} \quad h_i(\mathbf{q}) \leq 0, \quad i = 1, \dots, n_h \\ & \quad \quad \quad g_i(\mathbf{q}) = 0, \quad i = n_h + 1, \dots, n_h + n_g \\ & \quad \quad \quad q_j \in \text{dom}(q_j) \end{aligned}$$

where h_i and g_i represent inequality and equality constraints respectively, while n_h and n_g depict the number of inequality and equality constraints, and $\mathbf{q} \in \mathcal{F} \subseteq \mathcal{S}$. Figure 2.6 illustrates the feasible space and infeasible space for a given objective function when constraints are stipulated [86].

Many real world problems involve constrained optimization and the most popular approach to solve such problems are by using penalty functions due to their inherent simplicity [87]. Penalty functions can be applied to both linear and non-linear constraint problems. However, due to the generic nature of such methods, results obtained are not always satisfactory. The main drawback associated with penalty function approach is determination of the appropriate penalty parameters to guide the search towards the constrained optimum.

Figure 2.6: Illustration of a constrained optimization problem

2.4.1 Evolutionary Algorithms (EA)

Evolutionary Algorithms (EA) are based on the principles of biological evolution and are often applied to solve complex search and optimization problems. EA have been shown to be effective for noisy, multi-modal, discontinuous, and non-differentiable problems, compared to classical methods of optimization. Further, EA are more effective for linear, unimodal, quadratic, and strongly convex problems. In contrast to classical optimization where a search starts from a single point, an EA begins its search process from a diverse set of initial solutions or populations. This allows EA to search a large area of the search space in parallel in an effective way. The search is guided by the fitness values of individuals in the population.

The basic steps of EA include representation of the population with chromosome

Table 2.2: Pseudocode for evolutionary algorithm

begin: Evolutionary Algorithm
Let $t=0$ be the generation counter;
Create and initialize M -dimensional population: $\mathbf{q}(0)$;
repeat
Evaluate the fitness, $f(\mathbf{q}_i)$, of each individual, $\mathbf{q}_i$, in the population, $\mathbf{q}(t)$;
Perform cross-over to produce offspring;
Perform mutation on offspring;
Select population $\mathbf{q}(t+1)$ of new generation;
Advance to the new generation, i.e. $t = t+1$;
until terminating condition is true;
end

(binary or real valued), evaluation of fitness (objective) of individuals in a population, and subjecting the population to genetic operators such as reproduction operator, crossover operator and mutation operator. During a single iteration, each chromosome is evaluated against the goodness of the solution by computing the fitness followed by the generation of new chromosomes by subjecting the population to genetic operators, each new chromosome is paired and compared with old chromosomes from which it was produced, according to a selection rule and undergo changes defined by the genetic operators. Additional features such as elitism, constraint-handling, and multi-objective optimization are available for finding a better spread of solutions depending on the type of problem needs solving.

2.4.2 Multi-Objective Optimization

Multi-objective optimization algorithms are used when problems require simultaneous optimization of multiple competing objective functions confining very different aspects of the solutions space, conflicting or otherwise [40][88][89] [90]. Let $\mathcal{S} \subseteq \Re^M$ denotes

the M -dimensional search space (or decision space) and $\mathcal{F} \subseteq \mathcal{S}$ the space of the feasible solutions. Given a population of solutions $\mathbf{q} = (\mathbf{q}_1, \dots, \mathbf{q}_M) \in \mathcal{S}$, a single objective function f_i , is defined as $f_i : \mathbb{R}^M \rightarrow \mathbb{R}^r$. Let $\mathbf{f}(\mathbf{q}) = (f_1(\mathbf{q}), f_2(\mathbf{q}), \dots, f_r(\mathbf{q})) \subseteq \mathbb{R}^r$ where objective vector contains values of $r(\geq 2)$ objective functions. Using notations described above, the above multi-objective problem is defined as follows:

$$\begin{aligned} & \text{minimize} \quad \mathbf{f}(\mathbf{q}), \quad \mathbf{q} = (\mathbf{q}_1, \dots, \mathbf{q}_M) \\ & \text{subject to} \quad h_i(\mathbf{q}) \leq 0, \quad i = 1, \dots, n_h \\ & \quad \quad \quad g_i(\mathbf{q}) = 0, \quad i = n_h + 1, \dots, n_h + n_g \\ & \quad \quad \quad \mathbf{q} \in [\mathbf{q}_{min}, \mathbf{q}_{max}]^M \end{aligned}$$

where h_i and g_i represent inequality and equality constraints respectively, and $\mathbf{q} \in [\mathbf{q}_{min}, \mathbf{q}_{max}]$ represents the boundary constraints. The solutions, $\mathbf{q}^*$ to multi-objective problem are in the feasible space where $\mathbf{q}^* \in \mathcal{F}$.

The presence of conflicting objectives in multi-objective optimization problems may require maximizing one objective while minimizing another. Although such a scenario may lead to deterioration of results of one objective while improving the results of the other, trade-offs among such conflicting objectives are present. A potential set of solutions is determined to balance these trade-offs. Such solutions are referred to as *non-dominated solutions* or the *Pareto optimal set*. A solution is said to *dominate* another if it is better or equal with respect to all objectives and strictly better in at least one objective [40]. Often, there are more than one non-dominated solutions collectively known as the Pareto front as shown in Figure 2.7. The objective of multi-objective optimization algorithms is therefore to seek a Pareto optimal set of solutions instead of a single solution. Two main goals of MO optimization algorithms

are:

1. To converge on the Pareto optimal front
2. To Maintain as diverse distribution as possible

Figure 2.7: Illustration of Pareto optimal front

Definition 1:

A solution vector, $\mathbf{q}_1$ *dominates* solution vector, $\mathbf{q}_2$ (denoted by $\mathbf{q}_1 \prec \mathbf{q}_2$), if and only if,

1. $\mathbf{q}_1$ is not worse than $\mathbf{q}_2$ in all objectives, i.e. $f_i(\mathbf{q}_1) \leq f_i(\mathbf{q}_2)$, $\forall_i = 1, \dots, r$, and
2. $\mathbf{q}_1$ is strictly better than $\mathbf{q}_2$ in at least one objective, i.e. $\exists_i = 1, \dots, r :$
 $f_i(\mathbf{q}_1) < f_i(\mathbf{q}_2)$

Figure 2.8 illustrated the concept of dominance for a two objective function, $\mathbf{f}(\mathbf{q}) = (f_1(\mathbf{q}), f_2(\mathbf{q}))$. The colored area depicts the area of objective vectors dominated by $\mathbf{f}$.

Figure 2.8: Illustration of dominance

Definition 2:

A solution vector, $\mathbf{q}^* \in \mathcal{F}$ is Pareto-optimal if there does not exist a solution vector, $\mathbf{q} \neq \mathbf{q}^* \in \mathcal{F}$ that dominates the solution. i.e. $\nexists_i = f_i(\mathbf{q}) < f_i(\mathbf{q}^*)$. An objective vector, $\mathbf{f}^*(\mathbf{q})$, is Pareto-optimal if $\mathbf{q}$ is Pareto-optimal.

Definition 3:

The set of all Pareto-optimal decision vectors from the Pareto-optimal set, P^* . i.e. $P^* = \{\mathbf{q}^* \in \mathcal{F} \mid \nexists \mathbf{q} \in \mathcal{F} : \mathbf{q} \prec \mathbf{q}^*\}$. The Pareto-optimal set contains the set of balanced trade-off solutions for the multi-objective problem. The corresponding objective vectors are referred to as the Pareto-optimal front.

Definition 4:

Given an objective vector, $\mathbf{f}(\mathbf{q})$, and the Pareto-optimal solution set, P^* , then the Pareto-optimal front $\mathcal{P}$ is defined as $\mathcal{P} = \{\mathbf{f}(\mathbf{q}^*) = (f_1(\mathbf{q}^*), f_2(\mathbf{q}^*), \dots, f_i(\mathbf{q}^*)) \mid \mathbf{q}^* \in \mathbf{q}\}$. The Pareto front contains all the objective vectors corresponding to decision vectors which are not dominated by any other decision vector.

A pareto front returns a range of solutions as feasible solutions. The selected best solution offers trade-offs between the objective functions. In practice, prior knowledge is often used in selecting the best solution with the aid of a decision-maker who chooses a solution that is Pareto optimal. Methods of decision-making (before, after, or during search) have been extensively investigated in under multi-criterion decision making (MCDM) research [91, 92].

In our research, we experimented with various EAs for single and multi-objective function optimization. Among them, we observed the potential of Genetic Annealing Algorithm, (GAA) and Non-Sorting Genetic Algorithm II (NSGA-II), an implementation of multi-objective EA [40] for predicting binding peptides or motifs to MHC class II molecules. In chapter 3 and 4, we develop EA approaches for self-discovery and guided-discovery of motifs, and compare their performances with existing approaches.

Chapter 3

Self-Discovery of Motifs for Predicting MHC Peptide Binding

Peptide binding prediction to MHC molecules can be seen as a discovery of the presence of a motif in the peptide that binds to MHC molecules. Approaches of finding motifs in biosequences can be categorized into three main groups, namely, enumeration, heuristics and evolutionary computation [93]. While enumeration approaches adopt a consensus model as their representation of a motif, probability weight matrix is the model choice of heuristic approaches. Evolutionary computation on the other hand evolves to find the optimal motif by using a consensus or weight matrix model [94]. Other advantages of using EA algorithms are:

- Population approaches suit well for finding multiple solutions
- Implicit parallelism in EAs enables parallel search
- Multiple applications of classical methods do not constitute a parallel search

Herein we look at the problem of identifying the motif in a set of peptides (short sequences of variable lengths) as an optimization problem. We demonstrate the suitability of EA in finding an optimal alignment so that the motif that best describes the

peptide dataset is identified. Although earlier approaches formulated the problem as a single objective optimization problem by using the accuracy of motif, in practice, the motif detection problems involve more than one objective function which need simultaneous optimization. We therefore use MOEA to find a motif that characterizes a binding peptide. Recent advances and growing interest in EA and MOEA have also motivated us for exploitation of its usage for predicting peptides that bind to MHC molecules.

3.1 Previous Work

The goal of this section is to present and review EA driven methods for identifying peptide binding motifs for predicting binding peptides. Applications of EA for motif detection have revolved mainly around transcription factor binding site (TFBS) [95][96][97][98] and only a few have been developed for peptide binding prediction [76][70]. In an earlier study, Brusic *et al.* had used simple Genetic Algorithm (GA) to obtain an alignment for peptide binding predictions [70][26]. Another approach named Motif Discovery with Genetic Algorithm (MDGA) had been developed for predicting binding sites in homologous genes [99]. The ant colony search approach introduced by Karpenko et al. attempts to find an optimal local alignment for a given set of peptides in order to derive a position specific scoring matrix for peptide binding prediction [76]. In what follows, we discuss EA based methods relevant for peptide binding prediction, their strengths and weaknesses.

3.1.1 Prediction of MHC class II peptides using EA and ANN (PERUN)

The objective of this approach was to design a method that could combine experimental data and prior knowledge for the prediction of MHC class II binding data. A semi-automatic hybrid approach was adopted. In PERUN, Brusica *et al.* combined experimental binding data and expert knowledge by using an evolutionary algorithm (EA) [26]. The knowledge of primary anchor positions in reported binding motifs [22] was used to fix the position one corresponding to the primary anchor in each matrix, while the rest of the matrix was subject to the application of the EA to obtain an ensemble of alignment matrices representing putative binding cores. Once a peptide alignment was obtained from EA, an ANN classifier was used to classify binders and non-binders. Three levels of binding (high, moderate and low) have been experimented and assessed. The implemented system had shown promising results on a single HLA-DR4(B1*0401) dataset and was suggestive of its applicability for predicting peptides that bind to other HLA-DR molecules due to high structural and functional similarity of allelic variants of HLA-DR molecules. However, the results reported show that the predictive performance of PERUN is comparable to that of the quantitative matrix of Hammer *et al.* [100] and is slightly better than that of the binding motif reported by Rammensee *et al.* [22].

3.1.2 Motif Discovery using a Genetic Algorithm (MDGA)

Che *et al.* had applied a genetic algorithm to identify transcription factor binding sites (TFBS) [99]. Although there is no direct association between TFBS detection

and peptide binding site detection, we discuss MDGA in here in the context of motif discovery. In MDGA, an individual population represented by a set of possible starting locations of the binding sites on different homologous sequences. The fitness value of each individual was estimated by taking the sum of the information content (IC) for each column in the alignment of its binding sites. A penalty scheme has been implemented to penalize the individuals that have lower similarity in the alignment of their binding sites. Individuals with highly conserved binding sites were selected for future processing. After aligning the binding sites, a fitness function is used to evaluate the IC on a certain column and the background frequency of the same residue. The IC values were calculated over all four nucleotides. Pseudo counts have been used to overcome the problems that can arise from having zero observable nucleotide counts. MDGA has been tested on two datasets with experimentally validated motifs. The initial experiments performed with MDGA have indicated that the algorithm can achieve high prediction accuracy with less computational time. However, degradation of prediction accuracy when target sequences become very long was envisaged.

3.1.3 Ant Colony Search Strategy

Ant colony search strategy approach based on the ant colony system (ACS) was proposed by Karpenko *et al.* for the prediction of MHC class II binders [76]. This approach is somewhat similar to the behavior of Gibbs sampler [9]. This method uses a collective optimized search strategy of ants for the selection of the alignment which was subsequently used for the derivation of a position specific scoring matrix. The ACS concept has been derived from the behavior of ants that succeed in finding the nearest path from the nest to the food sources through communicating via

a collective memory of pheromone trails. The pheromone guiding process adopted by ants is defined in terms of information content of log-odds matrix estimated from the alignment. An objective function has been defined to estimate the sum of the relative entropies associated with an alignment. This function aims at an estimate of the significance of an alignment found by an ant through the comparison of the information content with a null model, one that is defined in terms of the background amino acid frequencies. This form of evaluation for an alignment has been used extensively in studies associated with multiple sequence alignment. An alignment is identified by the positions in sequences located by ants collectively, leading to rapid discovery of a multiple alignment. The performance comparison of ACS algorithm with the Gibbs Sampler approach has indicated comparable results for the benchmark dataset, HLA-DRB1*0401 available from the MHC-Bench [101]. However, it may be possible to improve the performance of the ACS approach by incorporating multiple objectives describing other aspects of the problem to the current objective function. The present implementation of the ACS algorithm presents a relatively slower convergence compared to that of the Gibbs method.

3.2 Motivation

Self-discovery is the process of finding the instances of a motif in a single or multiple sequences without any prior knowledge about their location or compositions, and then determining the consensus of the motif. Presence of short sequence motifs may not be apparent in an alignment of a related sequence set, especially when sequence set consists of short peptides. Although there are many computational methods available for self-discovering motifs without using prior information, the majority has focused

on DNA sequences and only a few is applicable to peptide sequences [9][10][102][76].

The necessity to obtain an alignment prior to the detection of motifs has been a major bottleneck in developing self-discovery approaches [13, 26, 30, 1, 24]. Often, an alignment is obtained based on prior knowledge such as anchor position specific residues and therefore the accuracy of the alignment heavily relies on this information [26, 24]. It is also worthwhile to note here that predictive models developed by leveraging on information inferred from small datasets may demonstrate inherent bias towards their derivative datasets and hence lack the generalization abilities. In addition, prediction accuracies of models built upon prior information often inherit aberrations in such information.

Noise and errors in datasets have an undesirable effect on building predictive models. In building a prediction model, it is therefore vital to choose an algorithm that is capable of handling noisy and missing information, thereby eliminating or minimizing errors introduced by such data samples. Another preferred characteristics of a prediction algorithm is the ability to explore a large space in finding potential solutions when searching for unknowns. In view of above requirements, we chose EA as our search mechanism due to its proven record of optimization capabilities in order to find effective algorithms to discover binding peptides to MHC molecules, which can handle noisy environment, incorporate biological evolutionary information, and explore a large search space for discovering an optimal solution. In this chapter, we propose a technique for self-discovery of motifs of peptides possessing different lengths by using single-objective GAA and MOEA. GAA based algorithm demonstrates a single objective optimization while MOEA shows the efficacy of motif detection by allowing multiple criteria. In what follows, GAA and MOEA based approaches for

self-discovery of motifs are presented.

3.3 Problem Representation

3.3.1 Motif Representation and Peptide Scoring

A k -mer motif of amino acids is characterized by a PSSM $Q = \{q_{ia}\}_{k \times 20}$ where q_{ia} denotes the binding strength of the site i when it is occupied by amino acid a . The binding score of a putative motif is computed by adding the binding scores assigned to 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 sequence $x_i = (x_{i1}, x_{i2}, \dots, x_{in_i})$ is determined by the maximum value of binding scores computed for all k -mer subsequences in x_i :

$$s_i = \max_j \{s_{ij} : j = 1, 2, \dots, n_i - k + 1\} \quad (3.3.1)$$

where s_{ij} denotes the binding score of the subsequence beginning at location j of the sequence, which is given by

$$s_{ij} = \sum_{l=1,2,\dots,k} q_{(j+l),x_{i(j+l)}} \quad (3.3.2)$$

and assuming that only one motif instance exists in every sequence, the location j^* of the motif is given by

$$j^* = \arg \max_j \{s_{ij} : j = 1, 2, \dots, n_i - k + 1\} \quad (3.3.3)$$

That is, the most likely motif instance of sequence x_i , say m_i , is given by the sequence $m_i = (x_{ij^*}, x_{ij^*+1}, \dots, x_{ij^*+k-1})$.

3.3.2 Representing PSSM in the EA Population

The elements of the PSSM Q are represented as $\mathcal{N} = 20k$ tuples $(q_{ia}, : i = 1, \dots, k; a \in \Omega)$ where Ω represents the amino acid alphabet. Each element in $\mathcal{N}$ is represented using a binary word of size ϑ so that $q_{ia} \in [0, 2^\vartheta - 1]$. The k -mer motif is therefore represented by an individual of $\mathcal{N}\vartheta$ long string in EA. Let the population at t -th iteration of the evolution is denoted by $\mathbf{q}(t) = \{\mathbf{q}_1(t), \mathbf{q}_2(t), \dots, \mathbf{q}_M(t)\}$ where $\mathbf{q}_j(t)$ represents an individual in the population of size M . During the GA evolution, the M population of binary strings, each representing a PSSM as described above is subjected to genetic operators (e.g. selection, crossover, mutation) in search for an optimal PSSM solution for a given objective function. The optimal solution is subsequently used to score peptides as described in 3.3.1.

3.4 GAA based Motif Discovery

The genetic annealing algorithm (GAA), is a hybrid of genetic algorithm [103] and simulated annealing [104][105]. It has demonstrated as a powerful tool in finding a solution when there is no exact optimal solution or in situations where the solution space is large [106][107]. The GAA incorporates simulated annealing [108] [104][105] into the crossover process of the GA [103], thereby combining the advantages of the both algorithms [106][107]. The strategy behind simulated annealing is to allow non greedy movements so that the solution can jump out of local minima.

In GAA, offspring produced in crossover between two parents are evaluated for their fitness, and highly fit offspring replace the parents for the next round of crossovers.

Table 3.1: Pseudocode demonstrating Genetic Annealing Algorithm

begin: Genetic Annealing Algorithm
Initialize the generation counter, $t=0$;
Create and initialize population, $\mathbf{q}(0)$;
Set the initial temperature, $T(0)$;
repeat
Evaluate the fitness $f(\mathbf{q}_i)$ for each individual $\mathbf{q}_i$ in the population $\mathbf{q}(t)$;
Initialize the iteration counter, $I=0$;
repeat
Perform annealing cross-over on $\mathbf{q}(t)$;
Accept offspring, if better than parent, else select according to Eq. 3.4.1;
Advance to the new iteration, $I=I+1$;
until terminating condition is true;
Advance to the new generation, $t= t+1$;
Reduce temperature, γ , i.e. $T = \gamma \cdot T$
until convergence;
end

Less fit offspring than their parents only survive with a selection probability characterized by Boltzmann distribution:

$$P_s(f) = \frac{1}{Z} \exp \left(-\frac{\Delta f}{KT} \right) \quad (3.4.1)$$

where Δf is the difference of fitnesses between the new chromosome and the old chromosome, $K > 0$ is the Boltzman constant, T is the temperature and Z is the normalizing function. After a new population is formed, the temperature of the population is lowered by a fraction γ and this procedure continues iteratively until the termination criterion is met.

The pseudocode of GAA is in given in Table 3.1. In the first step, a predefined number of individuals are recruited. Next, the fitness of each individual in the population is estimated. This is followed by the annealing crossover which operates for a

predefined number of iterations. During each iteration, the entire population is subjected to crossover process by selecting a highly fit individual (say parent-1) from the population and allowing it to mate with a partner (say parent-2) selected randomly from the population to produce two offspring. The fitnesses of the two new offspring (say offspring-1 and offspring-2) are evaluated. If the fitness of offspring-1 is better than offspring-2, the offspring-1 is compared with parent-1. 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. The entire crossover process is repeated with another partner selected at random from the population 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 [107]. After the predefined number of iterations is reached, the temperature is reduced, and the entire process is repeated for the new generation of the population. In GAA, $\mathcal{N}$ tuples in the position specific weight matrix Q are represented by integers.

3.4.1 Fitness Function

The fitness function is designed to arrive at a consensus of the motifs from all available experimental datasets and formulated to optimize the specificity and sensitivity of overall predictions in the training set by taking into account the binding information of peptides while maximizing the prediction accuracies.

A solution is evaluated based on its ability to maximize the accuracy of the identification of true binders (TP) and true non-binders (TN), and high marginal score

in separating binders and non-binders. In other words, a highly probable candidate solution must produce a lower binding score when tested on a non-binding peptide than on a binding peptide so that the behavior of the derived consensus motif becomes highly specific. These two prerequisites are introduced as a weighted sum of two objective functions linearly combined to estimate the fitness function. We try to minimize the sum of false positives (FP) and false negatives (FN) and the ratio between the average cumulative sum of non-binders and binders as a way of widening the gap between scores obtained by binders and non-binders. This is achieved by two fitness functions: f_1 to minimize the sum of false positives (FP) and false negatives (FN), and f_2 to minimize the ratio between the average cumulative scores of non-binders and binders. The fitness function f computed on the dataset over an iteration is given by the weighted sum of f_1 and f_2 as

$$f = f_1 + \kappa_2 f_2 \quad (3.4.2)$$

where

$$f_1 = \text{FN} + \kappa_1 \text{FP} \quad (3.4.3)$$

$$f_2 = \frac{N_b}{N_{nb}} \frac{\sum_{i=1}^N s(m_i) \delta(v_i = nb)}{\sum_{i=1}^N s(m_i) \delta(v_i = b)} \quad (3.4.4)$$

where $s(m_i)$ denotes the score computed for the most likely motif instance m_i of sequence x_i of the training dataset and the Kronecker δ is one when the argument is satisfied and otherwise is zero. N_b and N_{nb} are the total counts of binders and non-binders in the training dataset. The constant κ_1 ($> N_b/N_{nb}$ for $N_b > N_{nb}$, or $> N_{nb}/N_b$ for $N_{nb} > N_b$) was empirically determined to minimize the number of false positives. Initially, we could set the boundaries of the constant κ_1 as above. This

was followed by experimenting with a few different values for κ_1 within this range, for a subset of the training dataset divided into two sets; estimation and validation datasets. While estimation set was used to estimate the parameter, the validation set was used to evaluate the performance under the estimated parameters. Similar approach was adopted to estimate the combination operator κ_2 .

3.5 MOEA based Motif Discovery

The motif discovery algorithm with MOEA is driven by Non-Sorting Genetic Algorithm II (NSGA-II). In the algorithms the $\mathcal{N}$ tuples in the position specific weight matrix Q are represented by real numbers. Two fitness functions and two constraints are defined to facilitate the motif search process.

3.5.1 Non-Sorting Genetic Algorithm-II

NSGA-II has been proposed to alleviate number of drawbacks associated in MOEA, which uses non-dominated sorting and sharing. In particular, the use of MOEA is limited by: (1) $O(N_f M^3)$ computational complexity (N_f - number of objectives, M - population size) (2) lack of elitism, and (3) the need of a sharing parameter [40]. The elitism operator in EA prevents the loss of fit individuals encountered in earlier generations by allowing earlier solutions to survive in the subsequent generations, and also enables speeding up the performance. The diversity in a population is preserved by using the concept sharing, which requires the specification of a sharing parameter that determines the extent of sharing preferred in a problem. However, parameter less diversity preservation mechanism is desirable. The NSGA-II incorporates several

new genetic mechanisms to enhance classical MOEA for better convergence; fast non-dominated sorting with reduced $O(N_f M^2)$ computational complexity, elitism, diversity preservation by replacing the sharing function with crowded-comparison, and constraint handling.

Non-dominated Sorting

In non-dominated sorting, the solutions of the first non-dominated front are identified by comparing each solution with every other solution that requires $O(N_f M)$ number of comparisons per solution, and the complexity increases to $O(N_f M^2)$ in finding all non-dominated solutions in the first front. For worst case situations where there are M fronts with one solution per front, the number of computations requires becomes $O(N_f M^3)$. A simple non-dominated sorting procedure can be described as follows:

- Identify the best non-dominated set, $\mathcal{J}$

step-1: Let $i = 1$, $\mathcal{Q} = \emptyset$, $\mathbf{q}_i$ represent a motif solution.

step-2: For a solution $\mathbf{q}_j \in \mathbf{q}$ where $(j \neq i)$, check if $(\mathbf{q}_j \prec \mathbf{q}_i)$.

If yes, go to *step-4*.

step-3: If more solutions are left in $\mathbf{q}$ then $j = j + 1$; go to *step-2*; otherwise

$$\mathcal{Q} = \mathcal{Q} \cup \{\mathbf{q}_i\}.$$

step-4: $i = i + 1$; if $i \leq N$; go to *step-2*; otherwise STOP. $\mathcal{Q} = \mathcal{J}$.

- Discard them from the population
- Identify the next best non-dominated set
- Continue until all solutions are classified

Fast Non-dominated Sorting

The $O(N_f M^3)$ complexity of the non-dominated sorting is computationally expensive for large population sizes. This large computational complexity is due to the comparisons involved in the non-dominated sorting in every generation. Fast non-dominated sorting on the other hand has a computational complexity of $O(N_f M^2)$.

In fast non-dominated sorting, a population is subjected to several rounds of non-dominated sorting. That is, all the non-dominated individuals are identified and assigned the same fitness value until a new set of non-dominated solutions is found. The solutions found in subsequent rounds are assigned fitness values lower than those in the previous rounds. This process continues until the whole population is partitioned into non-dominated fronts with diverse fitness values. The fast non-dominated sorting algorithm is given in Table 3.2.

Diversity Preservation

Two main goals in MOEA are to obtain a set of solutions that converge on the Pareto optimal front and to maintain a diverse set of solutions. Here we use a crowded-comparison approach to maintain a set of diverse solutions [40]. The application of the crowded comparison operator is preceded by obtaining an estimation of the density of solutions surrounding a particular solution in the current population. This is carried out by calculating the average distance, $\mathbf{q}^{\text{dist}}$ of two points on either side of a solution along each of the objectives. The value, $\mathbf{q}^{\text{dist}}$ called crowding distance is an estimate of the perimeter of the *cuboid* formed by using the nearest neighbors as illustrated in Fig. 3.1. As shown, the crowding distance of the i -th solution in its front is the average side length of the cuboid shown with a dashed box.

Table 3.2: Pseudocode illustrating the fast non-dominated sorting algorithm.

begin: Fast Non-dominated Sorting Algorithm:	
$\mathbf{q}_i$ and $\mathbf{q}_j$ be motif solutions in the population, $\mathbf{q}$	
$n_{\mathbf{q}_i}$ be the domination count;	
i.e. number of solutions which dominate the solution $\mathbf{q}_i$	
$S_{\mathbf{q}_i}$ be a set of solutions that the solution $\mathbf{q}_i$ dominates.	
for each $\mathbf{q}_i \in \mathbf{q}$	
$n_{\mathbf{q}_i} = 0$	
$S_{\mathbf{q}_i} = \emptyset$	
for each $\mathbf{q}_j \in \mathbf{q}$	
if $\mathbf{q}_i \prec \mathbf{q}_j$ then	; $\mathbf{q}_i$ dominates $\mathbf{q}_j$
$S_{\mathbf{q}_i} = S_{\mathbf{q}_i} \cup \mathbf{q}_j$	; add $\mathbf{q}_j$ to the set of solutions dominated by $\mathbf{q}_i$
elseif $\mathbf{q}_j \prec \mathbf{q}_i$ then	
$n_{\mathbf{q}_i} = n_{\mathbf{q}_i} + 1$	; increment the domination counter of $\mathbf{q}_i$
if $n_{\mathbf{q}_i} = 0$ then	; $\mathbf{q}_i$ belongs to the first front, F_1
$\mathbf{q}_i^{\text{rank}} = 1$	
$F_1 = F_1 \cup \{\mathbf{q}_i\}$	
$t=1$	; initialize the front counter
while $F_t \neq \emptyset$	
$\mathcal{Q} = \emptyset$	; store the members of the next front
for each $\mathbf{q}_i \in F_t$	
for each $\mathbf{q}_j \in S_{\mathbf{q}_i}$	
$n_{\mathbf{q}_j} = n_{\mathbf{q}_j} - 1$	
if $n_{\mathbf{q}_j} = 0$ then	; $\mathbf{q}_j$ belongs to the next front
$\mathbf{q}_j^{\text{rank}} = t + 1$	
$\mathcal{Q} = \mathcal{Q} \cup \{\mathbf{q}_j\}$	
$t = t + 1$	
$F_t = \mathcal{Q}$	
end

Figure 3.1: Illustration of crowding distance estimation. Points marked in circles are solutions of the same non-dominated front.

The crowding distance computation involves following steps:

1. Sorting the population in ascending order according to each objective function.
2. Assigning distance value to solutions from each objective.
3. Calculating overall crowding distance values from the individual distance corresponding to each objective.

The algorithm for estimating crowding distance for all solutions in the non-dominated set $\mathcal{J}$ is given in Table 3.3.

Crowded Comparison Operator

At various stages of the algorithm, the selection process is guided by the crowded comparison operator to achieve a uniformly distributed Pareto optimal front. The crowded comparison operator, $\succ$ is defined as follows by assuming that every individual $\mathbf{q}_i$ in the population has two attributes; non-domination rank $\mathbf{q}_i^{\text{rank}}$ and crowding

Table 3.3: Algorithm for estimating crowding distance

begin: Crowding distance assignment:	
$\mathcal{J}[\mathbf{q}_i].c$ be the c -th objective function value of the individual $\mathbf{q}_i$	
in the set $\mathcal{J}$	
$f_c^{\max}$ maximum value of the c -th objective function	
$f_c^{\min}$ minimum value of the c -th objective function	
$df = f_c^{\max} - f_c^{\min}$	
$n = \mathcal{J} $	;number of solutions in $\mathcal{J}$,
for each $\mathbf{q}_i$, set $\mathcal{J}[\mathbf{q}_i]_{\text{dist}} = 0$	;initialize distance
for each objective c	
$\mathcal{J} = \text{sort}(\mathcal{J}, c)$	;sort using each obj. value
$\mathcal{J}[1]_{\text{dist}} = \mathcal{J}[n]_{\text{dist}} = \infty$	;select all boundary points
for $\mathbf{q}_i = 2 : (n - 1)$	;for all other points
$d\mathcal{J} = (\mathcal{J}[\mathbf{q}_i + 1].c - \mathcal{J}[\mathbf{q}_i - 1].c)$	
$\mathcal{J}[\mathbf{q}_i]_{\text{dist}} = \mathcal{J}[\mathbf{q}_i]_{\text{dist}} + d\mathcal{J}/df$	

distance $\mathbf{q}_i^{\text{dist}}$.

$$\begin{aligned} & \mathbf{q}_i \succ \mathbf{q}_j && \text{if } (\mathbf{q}_i^{\text{rank}} < \mathbf{q}_j^{\text{rank}}) \\ & \text{or } ((\mathbf{q}_i^{\text{rank}} = \mathbf{q}_j^{\text{rank}}) \text{ and } (\mathbf{q}_i^{\text{dist}} > \mathbf{q}_j^{\text{dist}})) \end{aligned}$$

Whereas the lower (better) rank is preferred between two solutions with different non-domination ranks, the solution located in a lesser crowded region is preferred if both solutions are on the same front.

Binary Tournament Selection

During tournament selection process, two or more individuals are randomly selected and their fitness values are compared. The best one is then released to the mating pool and the rest is returned to the population.

Simulated Binary Crossover

Simulated binary crossover (SBX) is a real parameter recombination operator which is commonly used with EA. SBX operator simulates the behavior of the single-point crossover operator on binary strings [109]. The SBX operator uses two parents and apply the combining operator variable by variable to create two offspring solutions. The combining operator involves a parameter, known as the distribution index η_c , which is fixed to a non negative value throughout a simulation run. Whereas for a chosen large value of η_c , the resulting offspring solutions are close to the parent solutions, and for a small value of η_c , solutions away from parents are often be created. As such, η_c has a direct effect in controlling the spread of offspring solutions. Since a search process of finding the optimal solution of a function landscape largely depends on controlling the spread (diversity) of offspring solutions compared to the selection pressure introduced by the chosen selection operation, fixing the η_c parameter to an appropriate value is an important task[109].

The procedure of computing the offspring-1, $\mathbf{q}_1(t+1)$ and offspring-2, $\mathbf{q}_2(t+1)$ from the parent solutions $\mathbf{q}_1(t)$ $\mathbf{q}_2(t)$ can be described as follows. A spread factor ϵ is defined as the ratio of the absolute difference in offspring values to that of the parents:

$$\epsilon = \left| \frac{\mathbf{q}_2(t+1) - \mathbf{q}_1(t+1)}{\mathbf{q}_2(t) - \mathbf{q}_1(t)} \right| \quad (3.5.1)$$

In the process of SBX, first, a random number between 0 and 1 is generated. This is followed by estimating the ordinate ϵ' from the probability distribution function as given in Eq. 3.5.2 so that area under the probability curve from zero to ϵ' is equal to

the chosen random number.

$$P(\epsilon) = \begin{cases} 0.5(\eta_c + 1)\epsilon^{\eta_c}, & \text{if } \epsilon \leq 1; \\ 0.5(\eta_c + 1)\frac{1}{\epsilon^{\eta_c+2}}, & \text{otherwise.} \end{cases} \quad (3.5.2)$$

A large value of η_c gives a higher probability for creating solutions near the parent thereby allowing a focused search and a small value of η_c allows distant solutions to be selected as offspring enabling a diverse search. Once ϵ' is determined from the probability distribution, the offsprings are calculated as follows:

$$\mathbf{q}_1(t+1) = 0.5 \left[(1 + \epsilon')\mathbf{q}_1(t) + (1 - \epsilon')\mathbf{q}_2(t) \right] \quad (3.5.3)$$

$$\mathbf{q}_2(t+1) = 0.5 \left[(1 - \epsilon')\mathbf{q}_1(t) + (1 + \epsilon')\mathbf{q}_2(t) \right] \quad (3.5.4)$$

Polynomial Mutation

The parameters of the evaluation function of each individual are tuned by using SBX together with a real-coded mutation operator known as polynomial mutation [110]. The mutated individual, $\mathbf{q}_i'(t+1)$ can be given as follow:

$$\mathbf{q}_i'(t+1) = \mathbf{q}_i(t+1) + (\mathbf{q}^U - \mathbf{q}^L) \varsigma_i \quad (3.5.5)$$

where $\mathbf{q}^L$ and $\mathbf{q}^U$ are the lower and upper boundaries of $\mathbf{q}$. And the parameter ς_i is calculated from the polynomial distribution:

$$P(\varsigma_i) = 0.5(\eta_m + 1)(1 - |\varsigma_i|^{\eta_m}) \quad (3.5.6)$$

$$\varsigma_i = \begin{cases} (2r_i)^{1/(\eta_m+1)} - 1, & \text{if } r_i < 0.5; \\ 1 - [2(1 - r_i)]^{1/(\eta_m+1)}, & \text{if } r_i \geq 0.5 \end{cases}$$

where η_m is the mutation probability and r_i is the random number in the range [0 1].

Constraint Handling Strategy

The suggested constraint-handling strategy in constrained NSGA-II eliminates the need of having a penalty parameter, and uses binary tournament selection where two solutions are selected from the population and the better solution is chosen for further processing. Under a given set of constraints, each solution can be either feasible or infeasible whereby at most three situations can be defined [40]: 1) both solutions are feasible; 2) one is feasible and other is not; and 3) both are infeasible. In order to handle the above three cases, in the context of MOEA, a set of rules is imposed as below:

- Select a solution with better objective function value by using crowded comparison operator.
- Select a feasible solution.
- Select the solution with smaller overall constraint violation.

The definition of domination between two putative motif solutions, $\mathbf{q}_i$ $\mathbf{q}_j$ where the solution $\mathbf{q}_i$ is said to constrained dominate a solution $\mathbf{q}_j$, if any of the following conditions is satisfied.

1. Solution $\mathbf{q}_i$ is feasible and solution $\mathbf{q}_j$ is not.
2. Solutions $\mathbf{q}_i$ and $\mathbf{q}_j$ are both infeasible, but solution $\mathbf{q}_i$ has a smaller overall constraint violation.
3. Solutions $\mathbf{q}_i$ and $\mathbf{q}_j$ are feasible and solution $\mathbf{q}_i$ dominates solution $\mathbf{q}_j$.

3.5.2 Objective Functions and Constraints

The two objective functions are defined as follows:

$$f_1 = \text{FN} + \kappa_1 \text{FP} \quad (3.5.7)$$

$$f_2 = \frac{N_b}{N_{nb}} \frac{\sum_{i=1}^N s(m_i) \delta(v_i = nb)}{\sum_{i=1}^N s(m_i) \delta(v_i = b)} \quad (3.5.8)$$

where $s(m_i)$ denotes the score computed for the most likely motif instance m_i of sequence x_i of the training dataset and the Kronecker δ is one when the argument is satisfied and otherwise is zero. N_b and N_{nb} are the total counts of binders and non-binders in the training dataset. The constant κ_1 ($> N_b/N_{nb}$ for $N_b > N_{nb}$, or $> N_{nb}/N_b$ for $N_{nb} > N_b$) was empirically determined to minimize the number of false positives. Initially, we could set the boundaries of the constant κ_1 as above. This was followed by experimenting with a few different values for κ_1 within this range, for a subset of the training dataset divided into two sets; estimation and validation datasets. While estimation set was used to estimate the parameter, the validation set was used to evaluate the performance under the estimated parameters. Similar approach was adopted to estimate the combination operator κ_2 .

The Eqs. 3.5.7 and 3.5.8 are subjected to the following two constraints in search of feasible motif solutions:

$$\frac{\text{FP}}{N_{nb}} \leq \frac{1}{\alpha_1} \quad (3.5.9)$$

$$\frac{\text{FN}}{N_b} \leq \frac{1}{\alpha_2} \quad (3.5.10)$$

The two parameters α_1 ($\ll N_{nb}$) and α_2 ($\ll N_b$) are set to minimize FP and FN rates, respectively. Given the training set, a few trial runs with different initializations are

necessary to determine the best values of α_1 and α_2 . A trial run is carried out by selecting a subset of the training set and setting the boundaries of the parameters to values where α_1 ($\ll N_{nb}$) and α_2 ($\ll N_b$) initially. This is followed by experimenting with a few different values of α_1 and α_2 within this range by fixing the value of α_1 and varying the value of α_2 until the best performance is obtained in terms of AUC for the chosen subset. If none of the solutions (PSSM) representing motifs satisfies the constraints defined above, MOEA reports no feasible solutions.

3.6 MHC-Peptide Binding Data

3.6.1 I-A^{g7} Dataset

The MHC class II molecule I-A^{g7} is known as the strongest genetic contributor to type 1 diabetes (T1DM)[111]. The non-obese diabetic (NOD) mouse is best known as a model for spontaneous T1DM although it is also prone to a range of other autoimmune diseases [112]. More than 100 different therapeutic manipulations that prevent T1DM in NOD mice have been demonstrated to date [113], whereas other manipulations in the NOD mouse are known to induce additional autoimmune disorders [112]. Because of this broad range of applications, the NOD mouse is amongst the most important animal models for the study of autoimmunity.

NOD mice have the H-2^{g7} haplotype: K^d, D^b, I-A^{g7}, and I-E^{null}. Of all, the class II I-A^{g7} appears to bind a large number of peptides with low affinity and also appears to be a promiscuous peptide binder [112][113]. As a result, defining a class II peptide binding motif for I-A^{g7} allele of the NOD has been more difficult than for its class I alleles. Several inconsistent I-A^{g7} binding motifs have been reported

[114]. The inconsistencies between individual I-A^{g7} binding motifs indicate biases and over-fitting to small experimental datasets from which they were derived.

For our analysis, several I-A^{g7} datasets were extracted from literature [52][115][116][117][118][119][120] and from Brusic, V.(unpublished data). The numbers of binders and non-binders in each dataset are given in Table 3.4. The datasets consist of short peptides ranging from 9-30 amino acids in length. Their binding affinities had been experimentally determined by independent studies and classified as binders or non-binders based on IC50 values according to the following scheme [116]: good binder (IC50_i≤100nM); weak binder (IC50=2000nM); non-binder (IC50_i≥50000nM). The datasets in [52][115][116][117][118][119][120] and Brusic's unpublished dataset were combined into a single training dataset and curated by removing duplicates and redundancy as follows: if a binder is a subsequence of another binder sequence, the longer binder sequence is discarded; if a non-binder is a subsequence of another non-binder, the shorter subsequence is discarded.

Let the curated whole dataset be referred to as *training* dataset here onwards and it be denoted by $D = \{(x_i, v_i) : i = 1, 2, \dots, N\}$ where N is the number of total peptide sequences 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 non-binder (nb). The number of peptides in the training set $N = 438$ in which the number of binders $N_b = 304$ and the number of non-binders $N_{nb} = 134$.

An independent dataset was generated from binders of Stratmann dataset [121] consisting of a diverse set of I-A^{g7} binding peptides with their binding affinities, to find the test accuracies in predicting binders and non-binders. The Stratmann dataset was balanced with randomly generated 9-mer non-binders so that for the

testing dataset, $N_b = N_{nb} = 112$. Twenty five such balanced test datasets were assembled by generating random samples starting from different seeds and adding them to the Stratmann dataset.

Table 3.4: Information on I-A^{g7} related peptide binding datasets. Unavailable information is indicated by ‘-’

Dataset	Non-binders	Binders	Reference
Reizis	21	33	[115]
Harrison	19	157	[116]
Gregori	31	109	[118]
Latek	8	37	[117]
Corper	35	13	[119]
MHCPEP	-	176	[52]
Yu	16	10	[120]
Brusic	37	-	[unpublished]

3.6.2 HLA-DRB1*0401 Benchmark Dataset

The HLA-DRB1*0401 benchmark dataset referred to here onwards as BM-Set1, consists of one training set and ten benchmark datasets and had been used to benchmark a number of motif finding algorithms [102][76][75][1].

The training set consisting of binders and non-binders was assembled as follows: an ensemble of 532 unique binding peptides were extracted from SYFPEITHI [34] and MHCPEP [52] databases, and a set of 177 unique non-binders was extracted from the MHCBN database [56]. The non-binder dataset consists of peptides of different lengths vary between 9-20 amino acids. The binder dataset was pre-processed by removing peptides that did not allow a hydrophobic residue at P1 position of

all putative 9-mer binding cores and unnatural peptides containing more than 75% alanine [1]. The pre-processed binder set has 456 unique peptides with a length distribution ranging from 9 to 30 amino acid residues. To overcome the imbalance between the number of binders and non-binders, the non-binder set was extrapolated. This was done by extracting single amino acid overlapping 9-mers from peptides longer than 9 amino acids and subsequently removing the duplicates.

Of the ten benchmark datasets, eight datasets were taken from the MHC-bench as described in [101]. The other two datasets were extracted from experiments described by Southwood [122] and Geluk [123]. An affinity $IC_{50}=1000nM$ was taken as the threshold for peptide binding as described in [122]. Homology reduction was carried out on all the benchmark datasets in order to reduce the chance of over-fitting by evaluating the prediction performance on peptide data included in the training. With homology reduction, the peptides in non-redundant validation sets had sequence identities less than 90% over an alignment of 9 amino acids in the training set. The number of binders and non-binders in the original and non-redundant datasets are given in Table 3.5.

3.6.3 IEDB Benchmark Dataset

We tested our method on a second benchmark dataset, BM-Set2 comprising of three mouse alleles and 13 HLA alleles made available at [124]. These quantitative peptide datasets had been extracted from the IEDB database[68]. The IEDB benchmark dataset referred here onwards as BM-Set2 consists of three mouse allele datasets and 13 HLA class II allele datasets.

The number of binders and non-binders in each dataset is given in Table 3.6.

Table 3.5: Description of the benchmark dataset, BM-SET1.

Benchmark Dataset	Original		Non-Redundant	
	Binder	Non-Binder	Binder	Non-Binder
Set1	694	323	248	283
Set2	381	292	161	255
Set3a	373	217	151	204
Set3b	279	216	128	197
Set4a	323	323	120	283
Set4b	292	292	120	255
Set5a	70	47	65	45
Set5b	48	37	47	37
Southwood	16	6	15	6
Geluk	22	83	19	80

The DRB3-0101 allele dataset was excluded from the benchmark dataset because of the significant imbalance between binders and non-binders (3 binders and 99 non-binders). For each allele dataset is divided into five files: c000, c001, c002, c003, and c004 for cross-validation. During cross-validation, one file is used as the evaluation dataset while the other four files were combined to as the training dataset. The binding affinity values (in this case log50k transformed values) greater than 0.426 are classified as binders.

3.7 Experiments

A number of experiments is carried out. First, we derive a motif for I-A^{g7} training dataset by using GAA based approach, and measure the performance on an independent test dataset, consisting of a number of balanced test datasets. The performance

Table 3.6: Description of BM-SET2

Type	Allele	Binders	Non-binders
Mouse	I-A ^b	43	33
	I-A ^d	56	286
	I-A ^s	35	91
HLA	DRB1-0101	920	283
	DRB1-0301	65	409
	DRB1-0401	209	248
	DRB1-0404	74	94
	DRB1-0405	88	83
	DRB1-0701	125	185
	DRB1-0802	58	116
	DRB1-0901	47	70
	DRB1-1101	95	264
	DRB1-1302	101	78
	DRB1-1501	188	177
	DRB4-0101	74	107
	DRB5-0101	112	231

is measured by using the AUC measure by averaging over the balanced test datasets. Next, we derive a motif by using the MOEA and the motif derived is evaluated by carrying out seven-fold cross-validation on the training dataset. This is followed by measuring the performance of MOEA self-discovered motif against other computational motif discovery approaches such as MEME, RANKPEP, and experimental motifs on the balanced I-A^{g7} test datasets by averaging the performance. In addition, we measure and compare the performances of the GAA and MOEA motifs discovered by the self-discovery approach on the BM-Set1 and BM-Set2 against other existing methods. By using an equal error rate (EER) plot, we demonstrate the selection of the best prediction threshold, which minimizes the false rejection rate (FRR) and false acceptance rate (FAR), for a given prediction model.

3.8 Results

3.8.1 Performance Comparison

Figure 3.2 illustrates the convergence of candidate solutions to a Pareto optimal front. The feasible solutions reported are evaluated for their AUC values on the training dataset and the PSSM which yields the highest AUC is selected as the consensus motif for the allele of interest.

The performance of the chosen PSSM on the testing peptide datasets was measured by estimating the AUC values. Let the score estimated for the binding core k in the peptide x_i be s_i . The binding status whether a binder (b) or non-binder (nb),

Figure 3.2: Illustration of convergence to a Pareto optimal front.

is determined according to a threshold t as follows:

$$v_i = \begin{cases} \text{b} & \text{if } s_i \geq t; \\ \text{nb} & \text{if } s_i < t. \end{cases}$$

The threshold t is determined from an EER plot drawn between far and frr. The threshold value at which both far and frr intercept is chosen as the best threshold for discriminating binders and non-binders.

On I-A^{g7} datasets

Table 3.7 shows cross-validation performance of self-discovered motif. In a particular cross-validation run, one experimental dataset was excluded and the motif was derived using the information of the remaining datasets. The motif was tested for predicting binders and non-binders of the left-out dataset.

Table 3.7: Seven-fold cross-validation performance of self-discovered motif on the training dataset

Dataset	AUC value
Reizis	75.0
Harrison	75.0
Gregori	77.0
Latek	93.0
Corper	70.0
MHCPEP	75.0
Yu	75.0

As seen in Table 3.7, by achieving AUC values greater than 0.7 for all cross-validation runs, MOEA derived motifs demonstrate better generalization capabilities compared to experimentally determined motifs.

Figure 3.3 illustrates the true boundary of the Pareto optimal front obtained for the I-A^{g7} training dataset. Motifs derived on the training dataset were next tested on the independent test dataset - a balanced set generated from Stratmann dataset. The results reported are based on the average AUC values over all balanced test sets. Figure 3.4 shows comparison of performances of motifs derived by the self-discovered and by earlier motif prediction approaches such as MEME [10], RANKPEP [61] and experimental motifs [114][125][115][116][117] [118][34]. The self-discovery approach uses only the binding information. An increase of 4-10% in predictive performance is observed with MOEA derived motif over the other approaches.

Figure 3.3: Illustration of the true boundary obtained for I-A^{g7} training dataset.

On Benchmark datasets

Comparison of performance of self-discovered motif for BM-Set1 dataset with enhanced Gibbs sampler obtained from the authors [1], TEPITOPE [72], SVRMHC [24], and ARB [32], is given in Table 3.8. As seen, MOEA shows comparable or superior performance with Gibbs sampler on all datasets except for the Southwood dataset. Out of the ten non-redundant (NR) datasets, the MOEA outperformed Gibbs sampler, TEPITOPE, SVRMHC and ARB by seven, nine, eight and ten datasets, respectively and is indicative of its better generalization ability. In the case of both original and NR Southwood datasets, the SVRMHC approach outperformed all the other methods. The SVRMHC approach has been trained on the AntiJen database (<http://www.jenner.ac.uk/antijen>), of which, Southwood dataset is a part of.

Table 3.8: Comparison of performances of the self-discovery methods of motif detection on BM-Set1 (DRB1*0401). Performances are measured in AUC values and the best values are shown in bold. †These values are based on smaller dataset sizes as SVRMHC didn't predict values for some of the peptides. The values from the Gibbs sampler were estimated from the matrix provided by the authors in [1].

Dataset		Performance (AUC%)				
		†SVRMHC	Gibbs	ARB	TEPITOPE	GAA MOEA (self-guided)
Original	set1	71.1	79.9	66.6	76.0	77.7 76.0
	set2	65.2	76.6	65.3	73.6	76.4 76.5
	set3a	62.6	74.0	65.2	73.0	71.9 73.3
	set3b	61.8	75.1	66.6	75.0	74.2 75.2
	set4a	70.6	78.8	66.8	74.8	76.6 74.8
	set4b	66.4	77.0	66.1	74.8	77.0 77.0
	set5a	55.3	60.4	53.9	65.3	76.1 77.7
	set5b	60.6	62.1	57.9	67.9	76.5 74.8
	Southwood	91.2	86.2	51.4	49.0	84.3 78.4
NR	Geluk	69.7	72.3	68.2	71.0	70.4 78.6
	set1	61.9	67.3	57.2	59.4	64.0 58.7
	set2	58.1	66.5	64.0	65.3	68.1 68.5
	set3a	57.8	59.8	60.0	59.8	62.6 66.0
	set3b	57.7	69.2	66.9	69.9	70.5 71.3
	set4a	59.7	67.1	57.5	57.3	64.3 59.9
	set4b	57.7	66.9	65.1	65.5	68.6 69.0
	set5a	54.4	60.1	53.6	64.6	75.9 79.0
	set5b	59.3	61.0	57.2	67.1	75.9 74.3
	Southwood	91.7	85.0	67.1	50.5	83.3 77.0
	Geluk	65.5	69.7	51.0	67.0	68.1 76.8

Figure 3.4: Comparison of performance of self-discovery motifs against MEME, RANKPEP, and experimental motifs on the balanced I-A^{g7} test datasets (the performance was averaged over 25 test datasets)

The performance of MOEA on BM-Set2 was compared with Gibbs sampler [1], TEPITOPE [72], SVRMHC [24], ARB [32] and NetMHCII [33]. Each allele dataset was subjected to five-fold cross-validation and the results are given in Table 3.9. The present method shows comparable or superior performance on majority of allele datasets compared to Gibbs sampler, SVRMHC, TEPITOPE, and NetMHCII. A fair comparison of ARB method cannot be drawn because the method has been trained on quantitative data obtained from IEDB [32].

Table 3.9: Comparison of performances of the self-discovery method of motif detection on BM-Set2. Performance is measured in AUC values. The best values displayed in bold are obtained by comparing all the methods listed in the table except the ARB method as ARB method has been trained on quantitative data obtained from IEDB and therefore a fair conclusion can not be drawn. The values from the Gibbs sampler were estimated from the matrix provided by the authors in [1].

Allele	Performance (AUC%)						
	SVRMHC	Gibbs	ARB	TEPITOPE	NetMHCII	GAA	MOEA (self-guided)
I-A ^b	-	-	66.2	-	90.8	81.9	91.9
I-A ^d	-	-	81.9	-	81.8	75.6	85.5
I-A ^s	-	-	-	-	89.8	74.9	88.9
DRB1-0101	62.3	67.6	66.6	64.7	71.6	60.8	65.1
DRB1-0301	-	72.2	79.9	73.4	76.5	72.2	77.8
DRB1-0401	73.9	75.9	73.7	75.4	75.8	68.4	72.5
DRB1-0404	-	74.3	78.8	82.9	78.5	76.2	78.6
DRB1-0405	70.1	72.4	72.4	79.0	73.5	74.5	75.6
DRB1-0701	-	69.5	74.9	76.8	78.7	69.0	73.5
DRB1-0802	-	72.1	80.3	76.9	75.6	68.9	77.3
DRB1-0901	-	73.4	71.1	-	77.5	71.0	71.2
DRB1-1101	-	71.5	72.7	71.0	73.4	72.8	75.9
DRB1-1302	-	71.6	91.7	72.0	81.8	75.0	82.0
DRB1-1501	73.0	67.2	79.2	72.6	73.6	72.4	74.3
DRB4-0101	-	74.2	80.0	-	73.6	72.8	75.9
DRB5-0101	64.9	61.8	67.7	65.3	66.4	69.6	66.0

3.8.2 EER Plots

An EER plot facilitates choosing the threshold that minimizes FAR and FRR upon which classification of the two classes are carried out. An ROC plot drawn at various thresholds is often used to estimate the sensitivity and specificity values of the predictive model with respect to different thresholds. Illustrated in Figure 3.5 below is the EER plot that determines the best threshold for the PSMM derived by using GAA, for the training dataset in BM-Set1.

Figure 3.5: Illustration of choosing the best threshold from an EER plot. Threshold which corresponds to the point where FAR and FRR intercept is selected as the threshold that minimizes the error in prediction.

3.8.3 Run Time Comparison

Even though EA are generally known to be computationally intensive, training for derivation of scoring matrices can be performed off-line and the prediction engines

can be provided through web services (e.g. SVRMHC, RANKPEP, ARB etc.) The computational times taken by the GAA and MOEA were slightly higher than the Gibbs sampler but were within the range of 7-20 minutes. Choosing EA methods over the EM is a trade-off between accuracy and running time. One other disadvantage of using EM for motif detection in peptide binding is that the algorithm is trained only with the positives samples (binders). As such, EM-based methods are more suitable for learning datasets with strong motifs. On the other hand EA based methods have the advantage of learning characteristics of both positives and negatives and therefore may be able to adopt a more discriminative approach during the exploration of the solution space.

3.8.4 Parameter Settings

The values of the parameters used in of GAA and MOEA based motif detection algorithms are given in Table 3.10.

Table 3.10: Parameter settings for GAA and MOEA.

Parameter	GAA	MOEA
Population size (M)	1000	500
Binary word size (ϑ)	7	7
Number of generations	40	300
Number of iterations in each generation	20	-
Initial Temperature (T_0)	0.2	-
Temperature reduction factor (γ)	0.9	-
Mutation probability (p_m)	0.035	0.005
Crossover probability (p_c)	-	0.9
Crossover distribution index	-	15 $\sim$ 20
Mutation distribution index	-	30 $\sim$ 50
κ_1	-	1 $\sim$ 2.5
α_1	-	5.0 $\sim$ 6.0
α_2	-	1.0 $\sim$ 2.0

Chapter 4

Guided-Discovery of Motifs

Presence of multiple and diverse binding motifs for a single MHC allele has been confirmed by reports based on Edman sequencing and mass spectrometric analysis [126]. Each peptide motif represents a subset of the population and fails to elucidate the variability of the population from which the subset is drawn. As a result, often times, with the experimentally determined peptides binding to MHC molecules, the binding motifs describing different composites of the data population are also available. This fact has not been investigated or used by earlier computational approaches to detect binding peptides, or motifs in general. Therefore, we introduce the new problem of guided-discovery of motifs in this chapter.

In guided-discovery, the discovery of motifs is guided by the knowledge of experimentally determined motifs, which comes with the binding data. The experimental motifs and associated information act as prior knowledge. Such information include anchor residue specific information obtained from X-ray crystallography that relates to molecule structure of an allele, entropy information from protein databases etc. In the past, prediction accuracy of motif detection has been enhanced by using anchor positions as prior information [1].

Motifs obtained experimentally using small datasets may be sub optimal in the sense they may lack the generalization capabilities. Predictions made with such motif instances on sequences in the same family and for previously unseen datasets show significantly lower accuracies. However, these motifs elucidate crucial residue information (preferred, tolerated or non-tolerated at a specific position) with respect to the binding locations [116][115][22] [121].

High binding peptides are often tolerant to a wide range of non-preferred amino acids at anchor positions [127]. Thus, the methods that rely fully on the anchor position specific residues for determining binding peptides may result in poor prediction of motif instances. Nevertheless, such information is valuable for a guided approach to seek a consensus motif for a specific allele. As such, here we incorporate the knowledge from experimental motifs explicitly into the motif discovery process. Here, in this chapter we introduce three methods for guided-discovery using MOEA approach: one uses the experimental motif patterns, the second uses computationally predicted motif patterns in the absence of experimental motifs or when the results achieved from computationally predicted motifs lack predictive power, and the other uses the anchor positions of MHC class II alleles.

4.1 Motivation

Often, the prior knowledge on anchor position specific information extracted from X-ray crystallography experiments is used to obtain an alignment for MHC class II variable length peptides [1][24][25][127]. Such prior knowledge of motifs is generally available through wet-lab experimental studies only for small sample datasets of specific alleles [22][115][116][121]. Therefore, these motifs lack generalization ability

on their own and are poor performers on new datasets. However, these sub motifs are good sources of information of binding sites and can be exploited in the guided-discovery of a consensus motif with better generalization abilities. In the absence of the experimental sub motifs, computationally predicted motifs through other available methods can be used to guide the motif discovery process to search a consensus motif.

In addition to the prior knowledge extracted from experimental motifs, there exist (or able to extract) information specific to alleles and their respective molecules and datasets. Such information includes anchor positions, evolutionary distances, and physico-chemical properties of binding site specific residues. Thus, these can be used as important properties to guide motif discovery process. Guided-discovery of motifs is introduced for the first time here. In this chapter, we propose a MOEA based technique for guided-discovery of motifs and demonstrate its application to experimental data carrying different types of prior information.

Motif discovery using probabilistic approaches often relies on positive data alone (often due to lack of negative data), and only minimal efforts have been made to enrich such processes with prior or already available information on motifs. In this chapter, we introduce an MOEA based approach referred to as *positive modeling*, which uses apriori knowledge implicitly in the form of evolutionary information for discovering binding motifs. In what follows, two guided-discovery approaches by using MOEA are presented. While one approach uses both positive and negative samples and is guided by the experimental motifs, the other uses only positive samples and is guided by the anchor positions.

4.2 Guided-Discovery by Experimental Motifs

In this algorithm, we presume that experimentally determined motifs are available along with the experimental datasets. We demonstrate the approach using the discovery of binding motif for I-A^{g7} molecule.

4.2.1 I-A^{g7} Experimental Motifs

A set of experimentally validated I-A^{g7} motifs [34][114][115][116][117][118][125] derived largely from uncorrelated datasets [115][116][117][118] was extracted and is illustrated in Table 4.1 along with the distribution of binders and non-binders in each dataset is used in the current experiments. Experimental motifs are described by the

Table 4.1: Information on I-A^{g7} related peptide binding datasets and motifs. Unavailable information is indicated by ‘-’. A motif is denoted by $m()$, for example $m(\text{Reizis})$ = motif by Reizis.

Dataset	Experimental Motif	Non-binders	Binders	Reference
Reizis	$m(\text{Reizis})$	21	33	[115]
Harrison	$m(\text{Harrison})$	19	157	[116]
Gregori	$m(\text{Gregori})$	31	109	[118]
Latek	$m(\text{Latek})$	8	37	[117]
-	$m(\text{Rammensee})$	-	-	[34]
-	$m(\text{Reich})$	-	-	[114]
-	$m(\text{Amor})$	-	-	[125]
Corper	-	35	13	[119]
MHCPEP	-	-	176	[52]
Yu	-	16	10	[120]
Brusic	-	37	-	[unpublished]

anchor positions and binding affinities of amino acids of the motif. The residues which

contribute significantly to the peptide binding are called primary anchor residues and positions they reside are called anchor positions. An amino acid occupying a specific position within a motif is characterized as well tolerated, weakly tolerated, or non-tolerated based on its involvement in the binding process.

The amino acids at primary and secondary anchor positions as well as forbidden amino acids at specific positions describe each motif extracted. We interpret these information as well tolerated, weakly tolerated and non-tolerated amino acids. While the presence of well-tolerated amino acids within the k -mer complements the binding process, the non-tolerated amino acids in these positions weaken the binding. Motifs described according to these characteristics are illustrated in Tables 4.2, 4.3, 4.4, 4.5, 4.6, 4.7 and 4.8. However, predictions made with these motifs demonstrate their lack of generalization power by giving higher prediction scores for datasets similar to their derivative datasets and lower scores for other datasets as seen in Table 4.11.

Table 4.2: The description of experimentally determined I-A^{g7} 9-mer peptide binding motif by Reizis: each position accommodates a well-tolerated, weakly-tolerated, or non-tolerated amino acid. The positions P4, P6 and P9 are the primary anchor positions where binding is highly likely to occur.

Position	Well-Tolerated	Weak-Tolerated	Non-Tolerated
P1	VEQMHLPD	-	R
P2	-	-	-
P3	-	-	-
P4	ILPV	HY	QEK
P5	-	-	-
P6	ATSNV	-	LYQK
P7	QVYLHINRF	-	-
P8	-	-	-
P9	ED	SM	LYTQK

Table 4.3: Harrison motif (Harrison et al., 1997) In Harrison motif, the positions P6 and P9 are primary anchors. Positions P3 and P8 are considered secondary anchor positions.

Position	Well-Tolerated	Weak-Tolerated	Non-Tolerated
P1	MADKLQY	P	X
P2	KADQY	LP	X
P3	RAKLQY	DGPWS	EF
P4	HAKLQY	DP	X
P5	GAQY	DKLP	X
P6	LIMV	DFHKNTY	AEGPQRSW
P7	DAEKLPQRY	X	X
P8	NAGPS	DEFKLQR	WY
P9	YFKR	DHLMNQW	AEGIPSTV

Table 4.4: Gregori motif (Gregori et al., 2000) In Gregori motif, the positions P6 and P9 are primary anchors. Positions P3, P4, P7 and P8 are considered secondary anchor positions.

Position	Well-Tolerated	Weak-Tolerated	Non-Tolerated
P1	X	X	X
P2	X	R	X
P3	S	IA	X
P4	KSA	LR	X
P5	RSGQ	PWT	X
P6	VPK	LYIR	X
P7	GARVW	QST	X
P8	GPBW	QSV	X
P9	RYKL	PQ	X

Table 4.5: Latek motif (Latek et al., 2000) The positions P1, P4, P6 and P9 are considered as primary anchors in Latek motif. Position P7 is considered a secondary anchor position

Position	Well-Tolerated	Weak-Tolerated	Non-Tolerated
P1	RAKFVSIMH	X	D
P2	X	X	X
P3	X	X	X
P4	VLI	LR	X
P5	X	X	X
P6	ATGVN	L	EY
P7	X	X	X
P8	GPBW	QSV	X
P9	GDEA	K	X

Table 4.6: Rammensee motif (Rammensee et al., 1999) In Rammensee motif, the positions P1, P4, P6 and P9 are primary anchors. Position P7 is considered a secondary anchor position.

Position	Well-Tolerated	Weak-Tolerated	Non-Tolerated
P1	KHSAV	RT	X
P2	X	X	X
P3	X	X	X
P4	L	IVM	X
P5	X	X	X
P6	VA	T	X
P7	X	X	X
P8	X	X	X
P9	DSE	X	X

Table 4.7: Reich motif (Reich et al., 1994) The Reich motif given in the paper (Reich et al., 1994) is based on MSA 560-574. This motif has been modified for the present study by taking a 9-mer sequence (565-573) from the consensus sequence. The modified motif is chosen based on the above defined Reizis and Rammensee motifs.

Position	Well-Tolerated	Weak-Tolerated	Non-Tolerated
P1	KHSAV	RT	X
P2	X	X	X
P3	X	X	X
P4	L	IVM	X
P5	X	X	X
P6	VA	T	X
P7	X	X	X
P8	X	X	X
P9	DSE	X	X

Table 4.8: Amor motif (Amor et al., 1996) In Amor motif, four positions have been described as very important. By considering the residues for these four positions and comparing them with the other motifs, positions P6, P7, P8 and P9 were chosen.

Position	Well-Tolerated	Weak-Tolerated	Non-Tolerated
P1	X	X	X
P2	X	X	X
P3	X	X	X
P4	L	X	X
P5	X	X	X
P6	IL	X	X
P7	KRH	X	X
P8	AG	X	X
P9	LF	X	X

4.2.2 Scoring Experimental Motifs

The description of an experimental k -mer motif conveys three kinds of information at each site: (1) the amino acid occupied, (2) the tolerance level of the amino acid, and (3) the strength of binding. Let us denote a k -mer motif validated in experiment 'e' by $m(e)$ and the tolerance level of the residue at site j by ρ_j where $\rho_j \in \{\text{well, weak, unknown, non-tolerated}\}$. The binding strength of site j is denoted by $\sigma_j \in \{\text{primary-anchor, secondary-anchor, other}\}$. In order to obtain quantitative measures for tolerance level (ρ_j) and binding strength (σ_j) we replaced the above qualitative measures with quantified values given in Table 4.9.

Then, the binding score for a k -mer experimental motif is given by

$$s(m(e)) = \sum_{j=1}^k \rho_j \sigma_j \quad (4.2.1)$$

Table 4.9: Quantitative measures defined for the tolerance level and the binding strength in experimental motifs.

Measure	Parameter	Value
Tolerance level	Primary anchor	4
	Secondary anchor	2
	Other position	1
Binding strength	Well-tolerated	4
	Weakly-tolerated	2
	Unknown	0
	Non-tolerated	-4

4.2.3 Representation of Experimental Motifs

The values in the PSSM corresponding to the elements in the experimental motifs are set to values within the same range $[0, 2^\vartheta - 1]$ as before. The following procedure is adopted to determine the elements of the experimental motif, $Q(m(e))$: a well tolerated amino acid at an anchor position of the motif receives the highest possible score of $2^\vartheta - 1$; the lowest score of zero is assigned to a non-tolerated residue; weakly tolerated residues and residues at secondary anchor positions receive of $(2^\vartheta - 1)/2$; and all the other unknown positions receives a score of $(2^\vartheta - 1)/3$.

4.2.4 Fitness Functions

Two objective functions f_1 and f_2 are proposed: f_1 to optimizes the specificity and sensitivity of the prediction of binders, and f_2 to best represent the characteristics of the motif that is close to the knowledge embedded in the experimental motifs:

$$f_1 = \text{FN} + \kappa_1 \text{FP} \quad (4.2.2)$$

The constant κ_1 ($> N_b/N_{nb}$ for $N_b > N_{nb}$, or $> N_{nb}/N_b$ for $N_{nb} > N_b$) was empirically determined to minimize the number of false positives.

$$f_2 = \sum_e |\hat{Q} - Q(m(e))| \quad (4.2.3)$$

where $\hat{Q}$ denotes the estimated PSSM of the motif. The MOEA minimizes the objective functions given in Eqs. 4.2.2 and 4.2.3, subjected to the two constraints given in Eqs. 4.2.4 and 4.2.5.

$$\frac{FP}{N_{nb}} \leq \frac{1}{\alpha_1} \quad (4.2.4)$$

$$\frac{FN}{N_b} \leq \frac{1}{\alpha_2} \quad (4.2.5)$$

The two parameters α_1 ($\ll N_{nb}$) and α_2 ($\ll N_b$) are set to minimize FP and FN rates, respectively. Given the training set, a few trial runs with different initializations are necessary to determine the best values of α_1 and α_2 . A trial run is carried out by selecting a subset of the training set and setting the boundaries of the parameters to values where α_1 ($\ll N_{nb}$) and α_2 ($\ll N_b$) initially. This is followed by experimenting with a few different values of α_1 and α_2 within this range by fixing the value of α_1 and varying the value of α_2 until the best performance is obtained in terms of AUC for the chosen subset. If none of the position specific weight matrices representing motifs satisfies the above constraints, MOEA reports no feasible solutions.

The summation in Eq. 4.2.3 is taken over all the experimental motifs and $|\hat{Q} - Q(m(e))|$ is the sum of squares of differences between individual elements of weight matrices $\hat{Q}$ and $Q(m(e))$. The knowledge of the experimental motif is incorporated to the consensus motif adaptively with the distance function used in f_2 .

4.3 Guided-Discovery by Computationally Predicted Motifs

Here, we demonstrate the guided-discovery approach to motif detection by fusing motifs predicted by computational methods. The tenet is to use computationally predicted motifs to guide the motif discovery process in the absence of experimentally determined motifs or when the results achieved from computationally predicted motifs lack predictive power. Herein, we will demonstrate the power of the fusion of knowledge from computationally predicted motifs in discovering a consensus or better motif with improved generalization ability.

4.3.1 Representation and Scoring of Computationally Predicted Motifs

Computationally predicted motifs are presented either as regular expression or as PSSMs. Regular expressions are represented by PSSMs similar to the representation previously adopted for experimental motifs. Amino acids corresponding to positions with high information content or probability are labeled as well-tolerant. Other amino acids are subsequently labeled as weak- or non-tolerant, depending on the level of significance achieved for being in specific positions. PSSMs derived are then scaled to maintain consistency. The scoring of the motifs is carried out by using the Eq. 4.2.1 similar to the case of experimental motifs.

We used three allele datasets as given in [33] namely, DRB1-0401, DRB1-0701, and DRB1-1501 for this experiment. These alleles are chosen due to reasonable number of binders and non-binder present in their respective datasets so that an independent

evaluation can be carried out while minimizing any biases that may arise when dealing with datasets with fewer number of samples. As for the computational motif discovery approaches, MEME and Gibbs sampler are chosen to predict motifs. Resulted motifs were then used to guide the MOEA guided-discovery approach by using the objective functions described in Eqs. 4.2.2 and 4.2.3.

4.4 Guided-Discovery by Anchor Positions

In searching motifs, probabilistic approaches often employ positive data (binders) alone by leaving out characteristics possessed by the negative samples [1] [10]. Herein, we propose and explore a predictive model based on positive samples alone which is driven by an EA. The objective functions are formulated by using relative entropy and the evolutionary distance among amino acids of an alignment. Whereas relative entropy, a measurement of uncertainty, is often used to analyze sequence features and alignments, to measure sequence conservation, the BLOSUM62 substitution matrix is a matrix suitable for modeling evolutionary problems [128]. The EA process is guided by the anchor positions illustrated in Fig.4.1.

As anchor positions are known to influence peptide-MHC binding, a higher weightage is given to such positions during the estimation of evolutionary distance. In order to reduce the sequence redundancy in an alignment, sequence clustering was followed by sequence weighting. The obtained best alignment is used to derive the PSSM and is subsequently used to predict binding peptides. The work flow of the guided-discovery by anchor positions is illustrated in Figure 4.2.

Figure 4.1: An illustration of putative 9 aa binding cores within a peptide sequence, s_i . Highlighted positions P1,P4,P6,P7,and P9 indicate anchor positions within a putative binding core associated with I-A^b.

4.4.1 Clustering and Weighting

Sequence clustering and weighting were performed according to [129] to reduce redundancy of the dataset. A set of sequences with sequence identity greater or equal to 62% forms a cluster. Cluster assignment is followed by the sequences weighting. Sequence weighting reduces over-representation of sequences in an alignment. A peptide x of length k -mer in cluster c is assigned a weight, $w_x = 1/n_c$, when n_c is the number of sequences in cluster c .

4.4.2 Pseudo-Count Correction

To avoid division by zero in calculating PSSM a pseudo-count correction was performed [130], which uses apriori knowledge of amino acid relationships represented by substitution matrices. For a given column, let pseudo-count frequencies be g_{ia} , for

amino acid a at position i , which is calculated according to Eq. 4.4.1 where q_{ib} , q_{0b} , and q_{ab} represent observed frequency of amino acid b in position i , background frequency of amino acid b , and the target frequency implicit in the substitution matrix (the frequency by which amino acid a is aligned to amino acid b), respectively.

$$g_{ia} = \sum_{b \in \Omega} \frac{q_{ib}q_{ab}}{q_{0b}} = \sum_{b \in \Omega} q_{ib}q_{a|b} \quad (4.4.1)$$

where $q_{a|b}$ is the conditional probability derived from the BLOSUM62 substitution matrix. The effective amino acid frequencies were then determined according to [130] by applying weight on pseudo-count correction as below:

$$\tilde{q}_{ia} = \frac{\mu_1 q_{ib} + \mu_2 q_{ia}}{\mu_1 + \mu_2} \quad (4.4.2)$$

Where μ_1 and μ_2 represent the effective sequence number and an arbitrary weight on the pseudo-count correction, respectively. If the number of peptide clusters generated be N_c , the value of $\mu_1 = N_c - 1$. An empirically determined suitable setting for μ_2 is given as 10 in [130].

4.4.3 Representing an Individual in a Population

Each individual (solution) in the evolving population represents possible starting positions of binding cores within each experimentally determined binding peptide in the positive training dataset. An individual is represented by a concatenated string of starting positions of all peptides in an alignment. The bit size for representing each starting position is determined as below. Given a peptide of length l , the number of k -mer peptides that can be derived from l is $d=l-k+1$. Hence, the starting positions

are located between 0 and $d-1$ where each peptide is overlapped by a single amino-acid. The bit size, ϑ is chosen such that $d < \min(2^\vartheta)$ whereby all k -mer peptide positions in the peptide are taken care of.

4.4.4 Fitness Functions

Based on the starting positions embedded in an individual, an alignment is generated for the N peptides. Let x_{j_1} and x_{j_2} be two peptide sequences where $x_{j_1} = (x_{j_1 1}, x_{j_1 2}, \dots, x_{j_1 n_{j_1}})$ and $x_{j_2} = (x_{j_2 1}, x_{j_2 2}, \dots, x_{j_2 n_{j_2}})$.

The fitness of the alignment is then evaluated by f_3 estimating the evolutionary distance, and f_4 estimating the entropy at anchor positions as given by the Eq. 4.4.3 and Eq. 4.4.5 below.

$$f_3 = \sum_{\substack{j_1 = 1 \\ j_2 = j_1 + 1}}^N \sum_{i=k} w_j B(x_{j_1 i}, x_{j_2 i}) \quad (4.4.3)$$

The evolutionary distance between two peptide sequences x_{j_1} and x_{j_2} at anchor positions $k' = \{P1, P4, P6, P7, P9\}$ of the k binding core in the alignment is calculated as below, where $B(x_{j_1 i}, x_{j_2 i})$ is the score estimates from the BLOSUM62 substitution matrix for amino acids $x_{j_1 i}$ and $x_{j_2 i}$. The weighting factor w_j is given in Eq. 4.4.4

$$w_j = \begin{cases} w, & \text{if an anchor position where } j = k'; \\ 1.0, & \text{otherwise.} \end{cases} \quad (4.4.4)$$

Next, we introduce another fitness function f_4 , which is equal to the sum of the relative entropy or the Kullback-Leibler distance in an alignment window [1, 43, 76].

$$f_4 = \sum_{j \in k} \sum_a q_{ja} \log \frac{\tilde{q}_{ja}}{q_{0a}} \quad (4.4.5)$$

where q_{ja} is the frequency of amino acid a occupying at position j in the alignment, $\tilde{q}_{ja}$ is the frequency of pseudo-count and sequence weight corrected amino acid a at position j , and q_{0a} is the background frequency of amino acid a . A number of different approaches are available for estimating background frequencies, also known as *background model* or *null model*: amino acid distribution in the SWISS-PROT database [131], a flat distribution where all amino acid frequencies are equal to $1/20$, or an amino acid distribution estimated from a non-binder dataset.

The best candidates comprise of individuals that maximize Eq. 4.4.3 and minimize Eq.4.4.5 simultaneously. An alignment that best optimize Eq. 4.4.3 and Eq.4.4.5 is used to build the PSSM, and subsequently use for predicting binders in the testing datasets.

4.5 Experiments and Results

In what follows, the description of experiments carried out and results obtained with the guided-discovery of motifs by using experimental motifs and anchor positions are given. While experimental motifs are incorporated into the guided-discovery approach to find a motif for I-A^{g7} dataset, the anchor positions are used to guide the discovery process to determine a motif for the I-A^b dataset.

4.5.1 I-A^{g7} dataset

Guided-discovery with experimental motifs is carried out with the I-A^{g7} datasets and the experimental motifs given in Table 4.10. The same training dataset, independent test dataset generated from binders of Stratmann dataset [121] I-A^{g7} binding peptides,

Figure 4.2: The flow chart of the motif discovery process guided by the anchor positions. Anchor positions are given more weightage during the estimation of fitness functions. An individual (solution) in the population individual (solution) represents starting positions of all putative binding cores in the positive sample set. EA operators represent crossover and mutation, respectively.

and the 25 balanced datasets used to measure the test accuracies of the self-discovery approach are used for measuring the performance of the motif derived from the guided-discovery approach as well. Figure 4.3 illustrates the true boundary of the Pareto optimal front obtained for the $I-A^{g7}$ training dataset.

We perform seven-fold cross-validation on the training dataset by using the motif obtained by incorporating the experimental motifs into the guided-discovery process. A computational validation of the accuracy of experimentally determined $I-A^{g7}$ motifs

on multiple datasets described in Table 4.1 is carried out next. In order to establish the behavior of experimentally determined motifs, we then measure the performance of each motif on the dataset that has been used to derive the motif as well as on other datasets. We also investigate guiding the motif discovery process with computationally predicted motifs by using probabilistic methods, in the absence of the experimental motifs. The next experiment is carried out to measure the performance of the predictive model devised using positive samples alone and guided by the anchor positions.

The cross-validation performance of the motif obtained from the guided-discovery process on I-A^{g7} datasets is given in Table 4.10. In a particular cross-validation run, one experimental dataset was excluded and the motif was derived using the information of the remaining datasets.

Table 4.10: Seven-fold cross-validation performance of guided-discovery motif on the training dataset.

Dataset	AUC value
Reizis	77.0
Harrison	74.0
Gregori	81.0
Latek	83.0
Corper	72.0
MHCPEP	77.0
Yu	71.0

As seen in Table 4.10, by achieving AUC values greater than 0.7 for all cross-validation runs, MOEA derived motifs demonstrate better generalization capabilities compared to experimentally determined motifs.

Figure 4.3: Illustration of the true boundary obtained for I-A^{g7} training dataset from guided discovery approach.

Validation of I-A^{g7} Motifs

The computational analysis on the multiple datasets indicate that experimental motifs satisfy only a subset of rules describing the optimal motif. From the validation results given in Table 4.11, the presence of diagonal dominance is evident in Reizis, Harrison, Gregori, and Latek motifs. Those motifs have yielded better performance for their respective datasets that have been used to derive the motifs, than for other datasets. This demonstrates biases that could arise when deriving motifs from datasets smaller in size or high in similarity. As seen, experimental motifs derived under such conditions lack generalization ability and therefore do not well characterize I-A^{g7} peptide binding to MHC molecules.

Table 4.11: Performance measured by AUC of experimentally determined I-A^{g7} motifs on their own datasets and other experimental datasets.

Experimental motif	Performance (AUC)						
	Dataset						
	Reizis	Harrison	Gregori	Latek	Corper	MHCPEP	Yu
$m(\text{Reizis})$	95.0	68.0	74.0	95.0	50.0	59.0	48.0
$m(\text{Harrison})$	75.0	88.0	69.0	64.0	53.0	72.0	33.0
$m(\text{Gregori})$	64.0	68.0	71.0	73.0	40.0	64.0	61.0
$m(\text{Latek})$	66.0	72.0	80.0	95.0	64.0	52.0	75.0
$m(\text{Rammensee})$	49.0	64.0	76.0	82.0	60.0	48.0	43.0
$m(\text{Reich})$	55.0	64.0	69.0	58.0	56.0	47.0	50.0
$m(\text{Amor})$	69.0	54.0	66.0	70.0	56.0	66.0	40.0

Performance Comparison

Motifs derived on the training dataset were next tested on the independent test dataset - a balanced set generated from Stratmann dataset. The results reported are based on the average AUC values over all balanced test sets. Figure 4.4 shows the comparison of performances of motifs derived by the self-discovered and by earlier motif prediction approaches such as MEME [10], RANKPEP [61] and experimental motifs [34][114][115][116][117][118][125]. The binding scores of I-A^{g7} experimental motifs were computed using Eq. 4.2.1. The values of the parameters used in the motif discovery process guided by the experimental motifs are given in Table 4.12 and the plot depicting AUC values obtained during the parameter estimation for α_1 and α_2 is given in Fig. 4.5.

Figure 4.4: Comparison of performance of guided-discovery motifs against MEME, RANKPEP, and experimental motifs on the balanced I-A^{g7} test datasets (the performance was averaged over 25 test datasets).

4.5.2 I-A^b Positive Modeling dataset

A dataset comprises of I-A^b peptides is used for the positive modeling. Peptide sequences and their binding affinities are obtained from SYFPEITHI [34], MHCPEP [52], AntiJen [55] and EPIMHC [57] databases. The extracted dataset consists of 251 unique binders with a length distribution ranging from 9 to 24 amino acid residues and 58 non-binders. Binder set was divided into two sets, training and test set so that there is no overlap between the two datasets. While training set, referred to as pm-training consists of 167 binders, the testing dataset (pm-testing) consists of 84 binders and 58 non-binders.

The performance of the motif derived by using positive samples alone, and guided

Table 4.12: Parameter setting for guided-discovery with experimental motifs.

Parameter	Setting
Population size (M)	500
Binary word size (ϑ)	7
Number of generations	300
Crossover probability (p_c)	0.9
Mutation probability (p_m)	0.005
Crossover distribution index	15 $\sim$ 20
Mutation distribution index	30 $\sim$ 50
κ_1	1 $\sim$ 2.5
α_1	5.0 $\sim$ 6.0
α_2	1.0 $\sim$ 2.0

by the anchor positions was compared with motifs derived from MEME [15][61][132] approaches, and self-discovered motif by using both positive and negative samples. The AUC performance is illustrated in the Table 4.13. The values of the parameters used in the motif discovery process guided by the anchor positions are given in Table 4.15.

Table 4.13: Comparison of performance of guided-discovery using positive modeling of motifs against MEME, Gibbs sampler, MOEA-self discovered motif, and RANKPEP on I-A^b dataset

Approach	AUC
Positive modeling	79.0
MEME	71.0
Gibbs sampler	86.0
RANKPEP	91.0
MOEA-self	96.0

Figure 4.5: Illustration of AUC values obtained during the parameter estimation for α_1 and α_2 . A subset of the I-A^{g7} training dataset was used for the parameter estimation. Discontinued curves indicate that there were no feasible solutions for given α_1 and α_2 values. Parameter estimation was performed by fixing one parameter (α_1) and evaluating AUC values for a range of values of the other parameter (α_2).

4.5.3 Performance of Guided-Discovery by Computationally Predicted Motifs

Here, the motif discovery process is guided by computationally predicted motifs. For our experiments, we utilize the motifs predicted by Gibbs sampler and MEME methods. Three MHC class II allele datasets, namely DRB1*0401, DRB1*1501, and DRB1*0701 datasets were used in the experiments. These datasets were chosen because they consist of reasonable number of binders and non-binders to form independent training and validation datasets for each computational prediction method. From each allele dataset, three mutually exclusive datasets were formed to obtain a

motif from each method as illustrated in Fig. 4.6. The training datasets of MEME and Gibbs sampler contain only binders, whereas the MOEA training set consists of both binders and non-binders. A testing set consisting of both binders and non-binders was formed with no overlap between training and the testing datasets. The testing set was used to measure performance of all three methods. Each method was trained three times by swapping training datasets, and the motif identified was tested on the testing dataset. The average of the performance is given in the Table 4.14.

Figure 4.6: An illustration of MOEA based motif discovery process guided by computationally predicted motifs. As an example, Motifs predicted by MEME and Gibbs approach were used to guide the motif discovery process. Three mutually exclusive binding datasets were used to train MEME, Gibbs, and MOEA respectively. While binding datasets B1 and B2 were used for training MEME and Gibbs, MOEA was trained with B3 binder set combined with a non-binder dataset, NB_x. An independent test dataset comprising binder set B and non-binder dataset NB_y was used for testing where AUC values were estimated for the test dataset from each of the discovered motif $m(\cdot)$.

Table 4.14: Comparison of performance of guided-discovery by using computationally predicted motifs. Same dataset was used to measure the performance of all three methods: MEME, Gibbs sampler, MOEA-guided discovery

Method	Average AUC value		
	DRB1*0401	DRB1*1501	DRB1*0701
MEME	49.0	65.7	57.7
Gibbs sampler	67.0	75.3	59.7
MOEA-guided	70.0	76.9	68.8

As illustrated in the Table 4.14, for all three alleles, the motifs discovered by using the guided-discovery principle outperformed predictions made with motifs obtained from independent prediction methods. These results demonstrate the power of fusing fragmented knowledge available from computationally predicted motifs in discovering a motif with improved power to better discriminate binders from non-binders. The parameter settings used in guided-discovery by computationally predicted motifs are same as the parameters given in Table 4.12.

Table 4.15: Parameter setting for guided-discovery with anchor positions.

Parameter	Setting
Population size (M)	1000
Binary word size (ϑ)	7
Number of generations	500
Crossover probability (p_c)	0.9
Mutation probability (p_m)	0.0004
Crossover distribution index	15 ~ 20
Mutation distribution index	30 ~ 50
Pseudo-count correction (β)	10.0
Weighting factor for anchor positions (w)	2.0

Chapter 5

Characterization of Motifs

X-ray crystallography studies of peptide-MHC complexes have revealed important insights into selectivity of binding peptides to different HLA alleles [8][21][29]. At the binding cleft of a peptide-MHC complex, some of the pockets have favored specific physicochemical properties over others, demonstrating their affinity towards specific amino acids. Amino acid compositions and their abundances of respective physicochemical properties therefore facilitate inferring characteristics such as affinity of binding peptides, which in turn is useful in understanding binding mechanisms and in designing or predicting peptides bind to MHC molecules.

Peptides are highly biased in terms of their position-dependent amino acid compositions and often favor hydrophobic sequences. This bias is more prominent at anchor positions which tend to accommodate only specific types of amino acids [127]. Despite the highly specific nature of binding peptides, studies have shown that high binding peptides are tolerant to a wide range of non-preferred amino acids making characterization of motifs a challenging task [127]. Despite the fact that physicochemical properties entail characteristics of binding peptides of MHC molecules, there have been no conclusive and comprehensive computational efforts to use such properties to

systematically characterize binding peptides. Therefore, in this chapter, we present techniques for characterizing motifs in order to understand their underlying binding mechanisms.

5.1 Previous Work

Majority of present methods for characterization of binding motifs has involved painstaking evaluation of their physicochemical properties by experimental means [117][119][133]. These methods include crystallography studies for obtaining three-dimensional structures, isolation of naturally or synthetically processed peptides from antigen presenting cells (APC), and subsequent manual inspection and mapping of residues in core binding regions of peptides for the purpose of evaluating the significance of binding [133]. Mutant peptides are also sometime used to observe the influence of the mapping of a specific amino acids occupying the region of the binding core. Other experimental methods involve using phage displayed peptide libraries for characterizing biochemical properties of molecules [121].

Descriptions of physicochemical properties at binding sites of binding peptides to specific alleles have been investigated with the aid of models of three dimensional quantitative structure activity relationship (3D-QSAR). Recently introduced computational models known as comparative molecular similarity indices analysis (CoMSIA), using 3D-QSAR, have been used to asses the contributions of physicochemical properties in peptide binding to HLA-A*0201 allele [134]. The CoMSIA method uses fields based on similarity indices such as steric, electrostatic, hydrophobic, and hydrogen bond donor and acceptor properties that describe similarities and differences among ligands, evaluate physicochemical requirements at each position of the peptide

structure, and decide the preferred amino acid sequences for high-affinity binding to HLA-A*0201 molecule [134]. The results have shown that hydrophobic interactions play a dominant role in peptide binding to MHC molecules. Coefficient contour maps of five physicochemical properties involved in the interactions between peptide and MHC molecule, generated by the CoMSIA model, have been used to exhibit favorable or unfavorable physicochemical properties affecting affinity of binding.

Though 3D-QSAR models are highly explanatory, the design of such models relies on a plethora of assumptions and the quality of these models are only as good as the quality of the data used in deriving the models. Another experimental study carried out on characterizing HLA-DQ8 binding motifs with the help of amino acid distributions at pockets of the binding core has yielded results consistent with the structural aspects of respective molecules [135]. However, as described, these methods are exploratory and carried out on a specific allele instead of a large scale analysis involving a wide range of alleles.

5.2 Motivation

Polymorphic residues that clustered within and around the binding groove provide the variations in shape and chemical properties which account for specific peptide-binding motifs identified by each MHC allele [51]. Therefore study of physicochemical properties of binding sites will facilitate understanding of immune responses and prediction of immunogenicity of MHC-binding peptides. Binding sites show allele-specific preferences of physicochemical properties of specific residues. Physicochemical properties are observed by analyzing crystal structures and peptid-binding experiments

determining binding specificities of residues at anchor positions. Amino acid properties have been used in the past to encode binding data to enhance the performance of classifiers [136][137][138]. In [136] Wen Liu et al. have used 11-factor encoding scheme, which takes into account a number of important physicochemical parameters such as hydrophobicity scale, polarity, isoelectric point, and accessible surface area to describe a peptide. Cui et al. [137][138] have used feature vector representing sequence derived structural and physicochemical properties such as hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure, and solvent accessibility of a peptide. Despite all evidences and earlier studies, physicochemical properties have not been extensively used and systematically quantified to find or characterize binding motifs in a large scale study.

In this chapter, we propose baseline methodologies aiming at characterization of motifs to explore and understand the relative abundance of physicochemical properties of amino acids residing at the binding sites of computationally predicted binders. Our proposed characterization is based on physicochemical properties of amino acids such as hydropathy (hydrophobic, hydrophilic, neutral) polarity (polar, non-polar), charge, and volume (very small, small, medium, large, very large). These parameters were chosen based on the peptide binding studies [115][116][117][133][139] and the analysis of I-A^{g7} crystal structure [119]. The abundance of these parameters are used to investigate binding sites predicted by both self- and guided-discovery on I-A^{g7} datasets. Moreover, we investigate the conservation of the I-A^{g7} binding site by comparing entropy at the binding site against a random binding site, and by using motif logos. Provided for comparison are motif logos derived for the BM-Set2 consisting of sixteen different alleles.

Figure 5.1: Illustration of the chemical structure of an amino acid.

5.3 Amino Acid Properties

Amino acids are building blocks of proteins. Three sequential bases of DNA or codons code for a single amino acid. There are twenty different amino acids in the alphabet: {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y}. The remaining characters of the English alphabet are sometimes used as ambiguity codes, for example, X stands for any amino acid.

The chemical structure of an amino acid is given in Fig. 5.1. Each amino acid has at least one amine and one carboxylic acid functional group. The R group is referred to as the amino acid side chain or residue. Due to the variations in the structures of different R groups, different amino acids demonstrate different physical and chemical properties.

Figure 5.2 illustrates the Venn diagram showing the relationships of the 20 amino acids [140] based on their various structural and functional properties. Amino acids are broadly classified into four classes, determined mainly based on their side chains:

Figure 5.2: Venn-diagram indicating physicochemical properties of amino acid residues.

(1) non-polar and neutral, (2) polar and neutral, (3) acidic and polar and, and (4) basic and polar. Depending on the polarity of the side chain, amino acids vary in their hydrophilic or hydrophobic characteristics. These properties are important in predicting protein structures and protein-protein interactions. The distribution of hydrophilic and hydrophobic amino acid determines the tertiary structure of the protein. And their physical location on the outside structure of the proteins influences their quaternary structure.

Table 5.1 illustrates values of the most frequently used physicochemical properties, hydropathy, polarity, and bulkiness according to the scales described by Kyte-Doolittle [2], Grantham [3], and Zimmerman[4], respectively.

In a protein, hydrophobic amino acids are more likely to be located at the protein interior whereas hydrophilic amino acid are more accessible to solvent, favoring interaction with water. Highly hydrophobic residues include V, I, L, M, F, W, and C. Amino acids A, Y, H, T, S, P, and G are less hydrophobic or indifferent to water. Arginine (R) and K are residues that are partly hydrophobic. Polar amino acids are those with side-chains that prefer to reside in an aqueous (i.e. water) environment. Therefore, they are exposed at the surface of a protein. Dominant polar amino acids are R, K, D, E, N, and Q. Other less polar or indifferent amino acids are: H, A, Y, T, S, P and G. Whereas negatively charged amino acids include E and D, the positively charged ones include R, K, and H occasionally acts as positively charged. Five amino acid volume classes can be defined based on the level of amino acid volumes in angstrom³ ($\text{\AA}^3$) [141]: very small (60-90 $\text{\AA}^3$; G, A, S); small (108-117 $\text{\AA}^3$; C, D, P, N, T); medium (138-154 $\text{\AA}^3$; E, V, Q, H); large (162-174 $\text{\AA}^3$; M, I, L, K, R); and very large (189-228 $\text{\AA}^3$; F, Y, W).

Table 5.1: Physicochemical properties of Amino Acids: the scales illustrated for hydropathy, polarity, bulkiness are Kyte-Doolittle [2], Grantham [3], and Zimmerman[4], respectively.

Residue			Hydropathy	Polarity	Bulkiness
Alanine	Ala	A	1.800	8.100	11.500
Arginine	Arg	R	-4.500	10.500	14.280
Asparagine	Asn	N	-3.500	11.600	12.820
Aspartic Acid	Asp	D	-3.500	13.000	11.680
Cysteine	Cys	C	2.500	5.500	13.460
Glutamine	Gln	Q	-3.500	10.500	14.450
Glutamic Acid	Glu	E	-3.500	12.300	13.570
Glycine	Gly	G	-0.400	9.000	3.400
Histidine	His	H	-3.200	10.400	13.690
Isoleucine	Ile	I	4.500	5.200	21.400
Leucine	Leu	L	3.800	4.900	21.400
Lysine	Lys	K	-3.900	11.300	15.710
Methionine	Met	M	1.900	5.700	16.250
Phenyl-alanine	Phe	F	2.800	5.200	19.800
Proline	Pro	P	-1.600	8.000	17.430
Serine	Ser	S	-0.800	9.200	9.470
Threonine	Thr	T	-0.700	8.600	15.770
Trypto-phan	Trp	W	-0.900	5.400	21.670
Tyrosine	Tyr	Y	-1.300	6.200	18.030
Valine	Val	V	4.200	5.900	21.570

There exist a number of scales for assigning numerical values describing different chemical and physical properties of amino acids. The most frequently used physicochemical property scales are hydrophobicity, hydrophilicity, bulkiness, molecular weight, polarity, volume, surface area, and refractivity. A bulky side chain of an amino acid at a specific position is known to have a marked effect on the conformation and thus the functions of a peptide. The variations of hydrophilic or hydrophobic character in amino acids depend on the polarity of the side chain of amino acids. For instance, the hydrophobicity scale gives the relative hydrophobicity values among amino acids; more positive values indicate stronger hydrophobicity while hydrophilic amino acids have negative values. A detail list of different scales for the above listed parameters can be found in [142].

5.4 Methods

Often, computational methods for characterization of motifs involve estimating entropy of motifs and displaying their characteristics with motif logos. Here, we discuss these methods in details and demonstrate their suitability and limitations to characterize motifs. Experimental approaches for detection of peptide binding often use physicochemical properties to elucidate binding motifs. Understanding properties and characteristics of amino acids involved in binding is an important aspect of defining rules governing selectivity and specificity of peptide binding.

For our analysis, we use properties such as hydropathy (hydrophobic, hydrophilic, neutral), polarity (polar, non-polar, charge, uncharge), and volume (very large, large, medium, small, very small) to quantitate the characteristics of binding cores and compare them with experimental findings. Evaluation is carried out by measuring

average relative abundances of each amino acid properties described above relative to each position within the 9-mer binding core.

We propose to generate profiles of physicochemical properties to compare the properties of amino acids at each position within predicted binding cores relative to the properties of randomly selected 9-mers from training data. This method finds the changes (peaks) that occur at the binding core, which show significantly different properties compared to those observed in randomly selected 9-mers.

5.4.1 Entropy of a Motif

The uncertainty of a random variable is usually defined by its entropy, which measures the amount of information required to describe the variable [143]. The entropy H (or Shannon Entropy) of the site i of an aligned set of sequences are given by

$$H_i = - \sum_{a_i \in \Omega} p_{ia} \log_2 p_{ia} \quad (5.4.1)$$

where p_{ia} is the probability of amino acid a at site i .

The relative entropy measures the distance between two probability distributions. Given the background distribution, p_{i0} and the distribution of motif site i , ($p_{ia} : a_i \in \Omega$), then relative entropy at site i is given by

$$H(p_{ia} \parallel p_{i0}) = \sum_a p_{ia} \log_2 \frac{p_{ia}}{p_{0a}} \quad (5.4.2)$$

Given an alignment for a binding site, the entropy of the alignment gives the amount of variability through a column in the binding site [144].

An entropy of a motif m of length k can be defined as follows:

$$H(m) = - \sum_{i=1}^k \sum_{a_i \in \Omega} p_{ia} \log_2 p_{ia} \quad (5.4.3)$$

This entropy of a motif can be used to estimate the goodness of a motif.

5.4.2 Motif Logos

A motif logo is a pattern that describes sequence conservation measured by using the information content. The information content is often defined as the relative entropy between the amino acid distribution of the motif (binding site), and the background distribution. When the chosen background distribution is other than uniform probability distribution, the logos are called Kullback-Leibler logos.

The sequence conservation at a specific position in an alignment is defined as the difference between the maximum possible entropy and the entropy of the observed symbol distribution for site i [144]:

$$H_{max} - H_i = \log_2 |\Omega| - \left(- \sum_{a \in \Omega} p_{ia} \log_2 p_{ia} \right) \quad (5.4.4)$$

where f_a is the observed frequency of symbol a at a specific position in the binding site. The maximum sequence conservation per site for protein is $\log_2 |\Omega| = 4.32$ bits.

Following steps were adopted for generating logos of the motifs: By using the estimated PSSM, binding scores of all peptides in the training dataset were estimated. From each peptide, the 9-mer which scored the highest was extracted. This was followed by the estimation of specificity and sensitivity values for a number of thresholds. A threshold, t was then chosen where specificity=sensitivity=81% or greater. At the chosen threshold value t , an alignment of the 9-mer binding cores was obtained for the actual binders detected as binders (true positives). These 9-mer binders were then input to the web server at [145] to obtain the motif logos. A logo consists of stacks of letters per position in the binding site. The overall height of each stack indicates the sequence conservation at that position, measured in bits. The height of

symbols within the stack indicate the relative frequency of the corresponding amino acid at that position.

5.4.3 Characterization Based on Binding Sites

Using the scoring matrix of motif of an allele, binding peptides in the training set could be extracted. The 9-mer binding peptides which scored high and predicted as binders were isolated and subjected to duplicate removal. This was followed by estimating the average relative abundance of physicochemical properties at each position within the 9-mer alignment. The profiles of the abundance of physicochemical properties observed at positions P1 to P9 are used to characterize binding sites.

5.4.4 Characterization by Bootstrapping of Binding Cores

To evaluate the differences among the predicted binding sites of an allele against randomly selected binding cores of the same allele in the training dataset, 9-mer cores from peptides longer than 9-mer in the binder dataset were extracted from the training set. The extracted 9-mer peptides were then curated by removing duplicates. Thereafter, the average relative abundance of physicochemical properties at each position within the 9-mer core were estimated to generate characteristic profiles for the binding core region. These profiles were subsequently used to evaluate and compare different motifs.

5.5 Experiments and Results

5.5.1 Characterization of I-A^{g7} Predicted Binding Sites

Due to the presence of candidate motifs from several experiments, we first characterize the motifs of the I-A^{g7} allele. The results of the characterization based on binding sites for I-A^{g7} are shown: physicochemical properties favored at each position of predicted binding peptides are illustrated in Figures 5.3, 5.4, and 5.5. As shown in Fig. 5.3, the position P4 of the predicted binders appears to have a higher concentration of hydrophobic residues compared to other positions [50][34]. P4 position is also shown to highly disfavor charged and hydrophilic residues [50] and is evident in Fig. 5.3 and Fig. 5.4. P9 of binding peptides favored the non-polar residues over polar, and charged over uncharged residues as shown in Fig. 5.4. With respect to the relative abundance of the volume of amino acids, positions P1 and P8 of binders indicate large residues and very small residues in P9 (see Fig. 5.5).

The study on the physicochemical properties of the binders predicted by the derived motif supported the evidence given in numerous previous experimental studies on I-A^{g7} binding motifs [117][119][133][139]. Moreover, the anchor positions of the binding sites determined are also in close agreement with the previous studies. As seen from the results, binding motif of I-A^{g7} is degenerate allowing binding to a wide selection of peptides.

Figure 5.3: Illustration of relative abundance of hydrophobic residues at the predicted I-A^{g7} binding site.

Figure 5.4: Illustration of relative abundance of polarity at binding positions of predicted I-A^{g7} binding site.

Figure 5.5: Illustration of relative abundance of volume of residues at the predicted I-A^{g7} binding site.

5.5.2 Characterization of I-A^{g7} by Random Selection of Binding Cores

The results obtained for the characterization by bootstrapping of binding cores are provided in this section. The plots illustrating relative abundance of properties in randomly selected 9-mer binding cores are given in Fig. 5.6, 5.7, and 5.8. Less variations of physicochemical properties were observed among positions P1-P9 of the randomly selected binding cores. Figure 5.9 illustrates the comparison of variation of relative abundance of physicochemical properties along 9-mer positions of (a) a predicted binder and (b) a random selection of a binding core.

Figure 5.6: Illustration of relative abundance of hydrophobic residues at randomly selected 9-mers of I-A^{g7} training dataset.

Figure 5.7: Illustration of relative abundance of polarity at randomly selected 9-mers of I-A^{g7} training dataset.

Figure 5.8: Illustration of relative abundance of volume of residues at randomly selected 9-mers of I-A^{g7} training dataset.

Figure 5.9: Comparison of variation of relative abundance of physicochemical properties along 9-mer positions of (a) a predicted binder (b) a random selection

Figure 5.10: Illustration of entropy at each position at the predicted binding site and at the randomly selected 9-mer binding core in the I-A^{g7} training dataset.

5.5.3 Analysis of Entropy and Motif Logos

The entropy plot obtained for the alignment of predicted binding sites and the random 9-mer binding cores selected from the training dataset is given in Fig. 5.10. As seen, the positions P3, P4, P7, P9 of the predicted binders exhibit less uncertainty compared to the positions P1, P2, P5, P6, and P8. The identification of the importance of P9 together with P4 and P7 as decisive binding positions, and the degeneracy of P1 corroborates with the earlier findings [115][116] [119]. Even though P6 has also been previously identified as a possible anchor position, it was not so obvious in the present analysis. The level of uncertainty at the positions within the randomly selected 9-mer cores was higher than that of the predicted binding cores, indicating lack of information content in them.

The binding motifs derived from self-discovery and guided-discovery for I-A^{g7} datasets are illustrated as motif logos in Fig. 5.11. Figures 5.12, 5.13, 5.14, 5.15

Figure 5.11: Motif logos derived for I-A^{g7} mouse allele (a)self-discovery (b)guided-discovery.

and 5.16 to depicts the sequence logo plots derived for the three mouse alleles and thirteen HLA alleles described in the BM-Set2 described in Table 3.6.

Figure 5.12: Motif logos derived for mouse alleles (a)I-A^b (b)I-A^d and (c)I-A^s.

Figure 5.13: Motif logos derived for HLA alleles in the BM-Set2 (a)DRB1-0101 (b)DRB1-0301 (c)DRB1-0401 and (d)DRB1-0404.

Figure 5.14: Motif logos derived for HLA alleles in the BM-Set2 (a)DRB1-0405 (b)DRB1-0701 (c)DRB1-0802 and (d)DRB1-0901.

Figure 5.16: Motif logos derived for HLA alleles in the BM-Set2 (a)DRB4-0101 and (b) DRB5-0101.

5.6 Validation of Physicochemical Profiles

To understand the binding mechanism of MHC-peptide binding motifs, we presented techniques that use physicochemical properties of amino acids at binding sites to characterize binding motifs. The proposed methods estimate the average of abundance of number of chemical properties at each position within an alignment of the 9-mer cores, obtained by using PSSM derived for an allele. This gives a profile of motifs characterizing their binding abilities. Subsequently, we propose to perform similar analysis on randomly selected 9-mer binding cores to characterize the motifs, mimicking motifs happen to exist by chance. In the following, we show how characteristics profiles of derived motifs and randomly selected motifs are used evaluate their differences. This will show the ability of our characterization approach to differentiate real motifs from randomly occurring motifs.

The profiles of relative abundance of properties at the predicted binding sites and randomly selected 9-mer binding cores of I-A^{g7} in the training dataset were derived. The number of 9-mer samples used in each case was 195 and 212, respectively. The average abundances of physicochemical properties at positions P1 to p9 were estimated and these values then used to plot the corresponding physicochemical profiles. Similarly, profiles were generated for all 12 different physicochemical properties. Figs. 5.17, 5.18 illustrate the characteristic profiles estimated for hydropathy. The estimates of polarity are given in Fig. 5.19 and 5.20. Figures corresponding to the volumes are given in Fig. 5.21, 5.22, and Fig. 5.23, respectively. The minimum and the maximum differences between the values obtained for the majority of the peaks of the predicted binding core positions and the randomly selected binding cores are in the range of 1.5 - 10 times, respectively.

As seen in the figures, no significant variation of specific physicochemical properties were observed at positions P1-P9 of random binding cores, compared to variations of the profiles at the positions of predicted binding sites. This demonstrates the ability of the proposed approach to capture information embedded at binding sites with the aid of physicochemical properties.

We also compared the physicochemical profiles against the motif logos derived for I-A^{g7} given in Fig. 5.11. These motif logos indicate a strong presence of amino acid R, which is a dominant polar in P1. This is observed in P1 in the profile depicted in Fig. 5.19. The dominance of charged amino acids at P2 is illustrated in Fig. 5.20. The position P3 favors non-polar and very small amino acids according to Figs. 5.19(b) and 5.23(a) which is in agreement with the observed dominant amino acid G in the motif logo. The anchor position P4 highly favors hydrophobic and non-polar amino acids, in this case L, and is evident in Figs. 5.17(a) and 5.19(b). Amino acids D and N are dominant in positions P5 and P6, and their characteristics are observed in the respective positions in Figs. 5.23(a) and 5.19(a). Tyrosine is dominant at position P7 and its characteristics are reflected in the profiles given in Figs. 5.17(b) and 5.21 (b). The position P8 is dominated by R and the associated characteristics are observable in Figs. 5.20(a) and 5.21(b). In line with the motif logo where G is dominant, the characteristics hydrophilic and non-polar are dominant at the position P9 as illustrates in Figs. 5.17(b) and 5.19(b).

As seen, physicochemical profiles corroborates with observations made with other motif characterization approaches. It is also worthy of note here that the proposed characterization approach can be easily applied to generate physicochemical profiles for other alleles. However, caution must be exercised in interpreting results as the

(a)

(b)

Figure 5.17: Illustration of relative abundance of residues at predicted binding core (BC) and randomly selected 9-mer core (RBC) in the I-A^{g7} training set (a) hydrophobic (b) hydrophilic

reliability of the characterization of motifs also depends on the size of datasets utilized in an experiment. The smaller datasets may not be representative enough to draw reliable conclusions.

Figure 5.18: Comparison of relative abundance of neutral residues between MHC-peptide binding sites and 9-mers randomly selected from the I-A^{g7} training dataset

Figure 5.19: Illustration of relative abundance of residues at predicted binding core (BC) and randomly selected 9-mer core (RBC) in the I-A^{g7} training set (a) polar (b) non-polar

Figure 5.20: Illustration of relative abundance of residues at predicted binding core (BC) and randomly selected 9-mer core (RBC) in the I-A^{g7} training set (a) charge (b) uncharge

Figure 5.21: Illustration of relative abundance of residues at predicted binding core (BC) and randomly selected 9-mer core (RBC) in the I-A^{g7} training set (a) verylarge (b) large

Figure 5.22: Comparison of relative abundance of medium residues between MHC-peptide binding sites and 9-mers randomly selected from the I-A^{g7} training dataset

(a)

(b)

Figure 5.23: Illustration of relative abundance of residues at predicted binding core (BC) and randomly selected 9-mer core (RBC) in the I-A^{g7} training set (a) verysmall (b) small

Chapter 6

Conclusion

6.1 Summary

Peptides that bind to specific MHC molecules and trigger T cell recognition (T cell epitopes) are able to stimulate immune responses and therefore are potential targets for development of vaccines and immunotherapy. Identification of such peptides is of paramount importance for diagnosis and treatment of many diseases. Computational approaches for their prediction can cut down the need of tedious and costly wet-lab experiments. As such, tools for predicting T cell epitopes have played an important role in advancing peptide based vaccine discovery. Over the past decade, researchers have developed various approaches to predict peptides that allow binding to MHC molecules. However, these prediction models often possess limitations due to the challenges posed by the characteristics of MHC molecules and the composition of binding peptides.

In this thesis, we introduced novel evolutionary algorithmic approaches for predicting peptides binding to MHC class II molecules. The problem of peptide binding

prediction was presented in the context of motif detection. The new concept - guided-discovery of motifs was introduced to find better motifs than those experimentally determined or computationally predicted by different techniques. Then, multi-objective evolutionary algorithms were proposed for both self-discovery and guided-discovery of motifs. The proposed guided-discovery approach uses information of experimentally found motifs or computationally predicted motifs. Unlike previous approaches, the present approach for self-discovery of motifs, we used the information available from both binders as well as non-binders.

For self-discovery and guided-discovery of motifs, we proposed several EA techniques: for self-discovery (1) single-objective genetic annealing algorithm, which incorporates annealing into the cross-over operator to allow moves resulting in solutions worse than the current solution in order to avoid getting trapped in a local minima resulting a better solution. (2) multi-objective EA, which uses objective functions designed to improve sensitivity and specificity while increasing the degree of separability of binders and non-binders. The proposed methods for guided-discovery are: (1) multi-objective positive sample modeling, for predicting binding peptides was investigated for its suitability as a prediction mechanism when only positive samples are available; the search is supported by analyzing entropy and evolutionary distance at key binding positions of MHC class II molecule discovered by xray-crystallography. (2) multi-objective EA for motif discovery by fusing experimentally predicted motifs was explored for its ability to discover a consensus motif when experimentally determined motifs lack coherence. (3) multi-objective EA for motif discovery by fusing computationally predicted motifs was explored to determine its ability to discover a

motif having better predictive power than motifs derived from computational methods that are of low accuracy. We showed applicability of the above approaches in various situations and their efficacy by comparing their performances with existing techniques.

The single objective GAA performed comparatively to existing techniques in detecting peptides that bind to MHC class II molecules. The motifs derived by MOEA self-discovery process for sixteen different alleles in the benchmark datasets against other methods such as enhanced Gibbs sampler, TEPITOPE, SVRMHC, NetMHCII, and ARB showed comparable or superior results. The proposed self-discovery approach using both binder and non-binder information often performed better than that of the probabilistic approaches using only binder information for discovery of binding motifs. The performance of the positive sample modeling was unremarkable for the tested dataset, for which, Gibbs sampler, and RANKPEP reported better results, whereas MEME reported the lowest.

MOEA guided-discovery process was tested on I-A^{g7} molecule, a promiscuous binder which had difficulty in determining a consensus motif. The motif discovery process was guided by a number of experimentally determined motifs. Final results showed an increase of 4-10% in predictive accuracy over other approaches, suggesting the efficiency of the guided-discovery algorithm in situations where determining a consensus is difficult purely by experimental studies. The MOEA guided-discovery approach driven by computationally predicted motifs outperformed when methods use individual computationally predicted motifs, on all three datasets tested. In brief, by using prior information of experimental motifs, the guided-discovery approach

achieves prediction accuracies higher than those obtained by any one of the experimental methods. Similarly, enhanced performance in motif detection is achieved by fusing computationally predicted motifs, compared to respective individual computational predictive approaches.

All in all, the results obtained showed that the motifs discovered by the proposed MOEA-based self-discovery and guided-discovery algorithms are better than earlier methods in predicting binding sites. This shows the applicability of the proposed methods in finding binding motifs in a wide range of alleles. The main advantage of the proposed guided-motif discovery approach over other methods is that its ability to fuse prior information while exploring a large solution space, providing more opportunities to find an optimal or near-optimal solution.

Lastly, methods to characterize motifs at the predicted binding sites by using physicochemical properties was proposed. There have been no comprehensive and proper way to characterize motifs in terms of their physicochemical properties. The characterization of predicted binding cores against randomly selected (bootstrapped) binding cores demonstrated significant variability (peaks) at specific positions within the 9-mer peptides. The minimum and maximum differences between values obtained for majority of peaks are in the range of 1.5 - 10 times, respectively. On the contrary to MHC class I molecules, where anchor positions are highly conserved allowing accommodation of only a few different amino acids, the anchor positions of MHC class II allow a wide range of amino acids, thereby introducing high degeneracy. This is evident in the illustrations depicted by the binding motif logos. The entropy values at each position of the predicted binding core were consistently higher than the values obtained for the positions at the random binding core illustrating the differentiating

power of the characteristics of the predicted binding core.

6.2 Discussion

The likelihood of finding an optimal motif by MOEA is higher than those by using a local or greedy search because of the stochastic nature of EA. The proposed approach learns from the characteristics of both binders and non-binders in the training dataset compared to methods use information only from binders to determine motifs. Moreover, ranges of the parameters involved in MOEA are known, so the parameters of the fitness functions are quickly estimated in a few cross-validation runs. Furthermore, unlike the earlier methods, the MOEA-based self-discovery and guided-discovery methods do not rely on any prior information such as anchor positions to obtain an alignment, prior distributions, etc. Given sufficient data samples representing both binders and non-binders, the method could be applicable to find motifs in other types of molecules.

Although there are several methods available for predicting peptide binding to MHC molecules, the performance of each of these methods may vary significantly across MHC class types and allele types, and also depend on the characteristics of datasets. This makes it difficult to choose the best prediction method without an empirical validation of a number of models. There can also be instances where the available prediction models may be incapable of providing the desired results due to various reasons such as deficiencies in model scalability, aberrations or variable length peptides in the datasets, difficult datasets due to errors and biases or degenerate nature of some motifs. Another aspect that hinders the refinement of computational prediction methods is the lack of access to wet-lab facilities, especially in the case of

MHC class II. Once 9-mer binding cores are identified from the longer peptides, those must be validated by experimental methods for accuracy, so that the artifacts can be easily detected and subsequently the algorithms can be retrained to rectify for any errors and anomalies.

Herein, we showed that motif detection algorithms based on evolutionary computation can play a pivotal role in predicting peptide binders with good accuracy in the event when enumeration and stochastic methods failed to produce reasonable prediction accuracies. Even though EAs are generally known to be computationally intensive, training for derivation of scoring matrices can be performed off-line and the prediction engines can be provided through web services. As seen in Tables 3.8 and 3.9, a single method do not always perform well on all types of allele datasets. The present methods showed higher accuracy in detecting motifs on majority of MHC alleles in the benchmark datasets. Therefore, we believe that MOEA-based methods could provide a general algorithmic framework for efficiently determining motifs in a wide range of MHC molecules.

In immunology, accuracy and speed in predicting binding peptides is of paramount importance. Computationally predicted binders do subsequently need to be validated with wet-lab experiments. By using computational predictions as an initial step, high cost involved in initial screening and time-consuming clinical testing can be significantly reduced. Towards this end, the proposed MOEA methods present a promising way to predict peptides that bind to MHC class II alleles including promiscuous and low affinity peptide binders.

6.3 Future Work

The research work performed in this thesis can be extended in many directions. A potential future direction would be to integrate additional information such as peptide length and peptide flanking regions (PFR). Such information has been shown to have the potential to enhance motif detection. This would lead to further improvement of the performance of the present algorithm.

Another extension would be to computationally predict the presence of multiple registers within binding peptides and examine their contribution to peptide binding. Multiple registers have been identified in high and medium binding peptides through structure based prediction approaches. New objective functions and constraints can be defined to allow the detection of multiple registers in peptides. However, as mentioned earlier, computationally predicted regions must be experimentally verified to establish their validity.

Better and improved prediction approaches can be envisaged with the availability of growing numbers of experimentally validated peptides from publicly available databases. The existing methods can be retrained on these incoming new datasets to improve their predictive power. Moreover, large scale benchmark studies, as provided here help analyzing the performance of methods so that the best prediction method for an allele can be decided. Another avenue open for exploration for improved T cell epitope prediction is to integrate MHC-binding prediction with protease cleavage and TAP transport as suggested in [146] for class I.

Another aspect that needs to be looked into is to improve MOEA based prediction algorithms so that the empirical testing required to determine the coefficients of the fitness function is minimized. New objective functions describing other feature

spaces such as three-dimensional structural parameters such as isotropic surface area, electronic charge index, etc and properties of the side chains of amino acids in of binding peptides can be defined and combined with the existing objective functions.

It has observed that computational complexity of NSGA algorithm increases with the population size due to the complexity involved in the non-dominated sorting takes place in every generation. Therefore, another direction for future research is to find ways to minimize the computational complexity involved in the NSGA-II algorithm. Such an approach would be very efficient specially for datasets containing peptides of longer lengths.

As discussed above, this research could be further strengthened by preforming a more detailed mathematical and computational analysis of the algorithms proposed in the thesis.

Bibliography

- [1] 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*, 20(9):1388–1397, 2004.
- [2] J. Kyte and R. Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132, 1982 May 5.
- [3] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864, 1974.
- [4] J. Zimmerman, N. Eliezer, and R. Simha. The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol*, 21(2):170–201, 1968 Nov.
- [5] H. Mamitsuka. Predicting peptides that bind to mhc molecules using supervised learning of hidden markov models. *PROTEINS: Structure, Function, and Genetics*, 33:460–474, 1998.
- [6] L.J. Stern and D.C. Wiley. Antigenic peptide binding by class I and class II histocompatibility proteins. *Behring Inst Mitt*, 94:1–10, 1994.
- [7] M. Bouvier and D. C. Wiley. Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules. *Science*, 265(5170):398–402, 1994.

- [8] 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*, 74(5):929–937, 1993.
- [9] 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*, 262(5131):208–214, 1993.
- [10] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51–80, 1995.
- [11] I. Rigoutsos and A. Floratos. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, 14:55–67, 1998.
- [12] S. Henikoff, J. G. Henikoff, W. J. Alford, and S. Pietrokovski. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, 163:GC17–GC26, 1995.
- [13] I. Jonassen, J. F. Collins, and D. G. Higgins. Finding flexible patterns in unaligned protein sequences. *Protein Sci*, 4:1587–1595, 1995.
- [14] M. P. Styczynski K. L. Jensen, I. Rigoutsos, and G. N. Stephanopoulos. A generic motif discovery algorithm for sequential data. *Bioinformatics*, 22(1):21–28, 2006.
- [15] T. Bailey, C. Elkan, and B. Noble. Multiple Em for motif elicitation. <http://meme.sdsc.edu/meme/website/meme.html>.
- [16] M. Tompa, N. Li, T. L. Bailey, and et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature-biotechnology*, 23(1):137–144, 2005.

- [17] J. Hu, B. Li, and D. Kihara. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research*, 33(15):4899–4913, 2005.
- [18] G. Zhang, A.M. Khan, K.N. Srinivasan, J.T. August, and V. Brusic. Neural models for predicting viral vaccine targets. *JBioinformatics and Computational Biology*, 3(5):1207–1225, 2005.
- [19] D.R. Flower. Towards in silico prediction of immunogenic epitopes. *TRENDS in Immunology*, 24(12):667–674, 2003.
- [20] J. C. Tong, T. W. Tan, and S. Ranganathan. Methods and protocols for prediction of immunogenic epitopes. *Briefings in Bioinformatics*, 8(2):96–108, 2006.
- [21] D. R. Madden. The three-dimensional structure of peptide-MHC complexes. *Annu Rev Immunol*, 13:587–622, 1995.
- [22] H.G. Rammensee, T. Friede, and S. Stevanović. MHC ligands and peptide motifs: first listing. *Immunogenetics*, 41(4):178–228, 1995.
- [23] E. Y. Jones, L. Fugger, J. L. Strominger, and C. Siebold. Mhc class II proteins and disease: a structural perspective. *Nat Rev Immunology*, 6:271–282, 2006.
- [24] J. Wan, W. Liu, Q. Xu, D. R. Flower, and T. Li. SVRMHC prediction server for MHC-binding peptides. *BMC Bioinformatics*, 7:463, 2006.
- [25] I. A. Doytchinova and D. R. Flower. Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. *Bioinformatics*, 19(17):2263–2270, 2003.
- [26] V. Brusic, R. George, M. Honeyman, and L. C. Harrison. Prediction of MHC class II-binding peptides and artificial neural network. *Bioinformatics*, 14:121–130, 1998.

- [27] A. Sette. Tools of the trade in vaccine design. *Science*, 290(5499):2074b–2075, 2000.
- [28] P. R. Mallios. Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics*, 15(6):432–439, 1999.
- [29] 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*, 351(6324):290–296, 1991.
- [30] P. R. Mallios. Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics*, 17(10):942–948, 2001.
- [31] 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*, 94(3):264–270, 2002.
- [32] H. Bui, J. Sidney, B. Peters, M. Sathiamurthy, A. Sinichi, K. Purton, B. R. Moth, F. V. Chisari, D. I. Watkins, and A. Sette. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*, 57(5):304–314, 2005.
- [33] M. Nielsen, C. Lundegaard, and O. Lund. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*, 8(238), 2007.
- [34] H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanovic. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219, 1999.

- [35] A. J. Godkin, K. J. Smith, A. Willis, M. V. Tejada-Simon, and J. Zhang. Naturally processed hla class ii peptides reveal highly conserved immunogenic flanking region sequence preferences that reflect antigen processing rather than peptide-mhc interactions. *J Immunol*, 166:6720–6727, 2001.
- [36] I. A. Doytchinova and D. R. Flower. A comparative molecular similarity index analysis (CoMSIA) study identifies an HLA-A2 binding supermotif. *J Comput Aided Mol Des*, 16(8-9):535–544, 2002.
- [37] I. A. Doytchinova and D. R. Flower. Additive method for the prediction of protein-peptide binding affinity. application to the MHC class Imolecule HLA-A*0201. *J Proteome Res*, 1(2):263–272, 2002.
- [38] T. Krink, F. Bodgan, G. B. Fogel, and R. Thomson. Noisy optimization problems - a particular challenge for differential evolution? In *Proc. Sixth Congress on Evolutionary Computation (CEC-2004)*, IEEE Press, 2004.
- [39] D. V. Arnold and H. G. Beyer. A comparison of evolution strategies with other direct search methods on the presence of noise. *computational optimization and Applications*, 24:135–159, 2001.
- [40] K. Deb, A. Pratap, S. Agrawal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm:NSGA-II. *IEEE Trans on Evolutionary Computation*, 6(2):182–197., 2002.
- [41] J. Knowles and D. Corn. The pareto archived evolution strategy: A new baseline algorithm for multiobjective optimization. In *Proceedings of the 1999 Congress on Evolutionary Computation*, pages 98–105, IEEE Press,Piscataway, NJ, 1999.
- [42] E. Zitzler. Evolutionary algorithms for multiobjective optimization:methods and applications. In *Doctoral dissertation ETH 13398*, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, 1999.

- [43] O. Lund, M. Nielsen, C. Lundegaard, C. Kesmir, and S. Brunak. *Immunological Bioinformatics*. The MIT Press, Cambridge, Massachusetts, 2005.
- [44] P. Lydyard, A. Whelan, and M. Fanger. *Immunology, vol. 2nd edition*. Taylor and Francis, 2004.
- [45] J. D. Roberts, D. Niedzwiecki, and W. E. Carson. Phase 2 study of the g209-2M melanoma peptide vaccine and low-dose interleukin-2 in advanced melanoma: Cancer and leukemia group B 509901. *J Immunotherapy*, 29:95–101, 2006.
- [46] D. N. Bourdette, E. Edmonds, and C. Smith. A highly immunogenic trivalent t cell receptor peptide vaccine for multiple sclerosis. *Mult Scler*, 11:552–561, 2005.
- [47] J. A. Lopez, C. Weilenman, and R. Audran. A synthetic malaria vaccine elicits a potent CD8() and CD4() t lymphocyte immune response in humans. implications for vaccination strategies. *EurJ Immunol*, 31:1989–1998, 2001.
- [48] K.L. Knutson, K. Schiffman, and M.L. Disis. Immunization with a HER-2/neu helper peptide vaccine generates HER-2/neu CD8 T-cell immunity in cancer patients. *J. Clin. Invest*, 107:477–484, 2001.
- [49] I. Roitt. *Essential Immunology*. Blackwell Science Ltd; 9 Rev Ed edition, 1997.
- [50] J.A.Todd, H. Acha-Orbea, J. I. Bell, N. Chao, Z. Fronek, C. O. Jacob, M. McDermott, A. A. Sinha, L. Timmerman, and L. Steinman et al. A molecular basis for MHC class II associated autoimmunity. *Science*, 240(4855):1003–1009, 1998.
- [51] M. G. Rudolph, R. L. Stanfield, and I. A. Wilson. How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol*, 24:419–466, 2006.
- [52] V. Brusica, R. George, and L. C. Harrison. MHCPEP, a database of MHC-binding peptides: update. *Nucleic Acids Res*, 26(1):368–371, 1998.

- [53] A. Sette, W. Fleri, B. Peters, M. Sathiamurthy, H. H. Bui, and S. Wilson. A roadmap for the immunomics of category A-C pathogens. *Immunity*, 22(2):155–161, 2005.
- [54] C. P. Toseland, D. J. Clayton, H. McSparron, S. L. Hemsley, M. J. Blythe, K. Paine, I.A. Doytchinova, P. Guan, C. K. Hattotuwigama, and D. R. Flower. Antijen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res*, 1(1):4, 2005.
- [55] M. J. Blythe, I. A. Doytchinova, and D. R. Flower. Jenpep: a database of quantitative functional peptide data for immunology. *Bioinformatics*, 18:434–439, 2002.
- [56] M. Bhasin, H. Singh, and G. P. S. Raghava. MHCBN: A comprehensive database of MHC binding and non-binding peptides. *Bioinformatics*, 19:665–666, 2003.
- [57] P. A. Reche, H. Zhang, J. P. Glutting, and E. L. Reinherz. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. *Bioinformatics*, 21:2140–2141, 2005.
- [58] R. M. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press Cambridge, 1998.
- [59] M. Gribskov, A. D. McLachlan, and D. Eisenberg. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci.*, 84(13):4355–8, 1987.
- [60] J. S. Liu and C. E. Lawrence A. F. Neuwald. Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *J. Amer. Statis. Assoc.*, 90:1156–1170, 1995.

- [61] 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*, 56(6):405–419, 2004.
- [62] S. Henikoff and. Block+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, 15:471–479, 1999.
- [63] J. S. Liu. The collapsed gibbs sampler in bayesian computations with applications to gene regulation problem. *J. Amer. Statist. Assoc.*, 1995.
- [64] J. S. Liu and C. E. Lawrence. Unified gibbs method for biological sequence analysis. In *Proc. of the Biometrics Section, Amer. Statist. Assoc.*, pages 194–199, 1996.
- [65] G. Thijs and. A higher order background model improves the detection of regulatory elements by gibbs sampling. *Bioinformatics*, 17(12):1113–1122, 2001.
- [66] 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*, 9(2):447–464, 2002.
- [67] F. Dellaert. The expectation maximization algorithm. In *Technical Report Number GIT-GVU-02-20*, College of computing, Georgia Institute of technology, 2002.
- [68] A. Sette A, W. Fleri, B. Peters, M. Sathiamurthy, H.H. Bui, and S. Wilson. A roadmap for the immunomics of category A-C pathogens. *Immunity*, 22(2):155–161, 2005.
- [69] K. Gulukota, J. Sidney, A. Sette, and C. DeLisi. Two complementary methods for predicting peptide binding major histocompatibility complex molecules. *J Mol Biol*, 267:1258–1267, 1997.

- [70] V. Brusica, C. Schonbach, M. Takiguchi, V. Ciesielski V, and L. C. Harrison. Application of genetic search in derivation of matrix models of peptide binding to MHC molecules. In *Proc Int Conf Intell Syst Mol Biol*, pages 75–83, 1997.
- [71] K. C. Parker, M. A. Bednarek, and J. E. Coligan. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*, 152:163–175, 1994.
- [72] T. Sturniolo, E. Bono, D. Jiayi, L. Raddrizzani L, O. Tuereci, U. Sahin, M. Braxenthaler, F. Gallazzi, M. Protti, F. Sinigaglia, and et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nature Biotech*, 17(6):555–561, 1999.
- [73] F. R. Burden and D. A. Winkler. Predictive bayesian neural network models of MHC class II peptide binding. *J Mol Graph Model*, 23(6):481–489, 2005.
- [74] L. Bisset and W. Fierz. Using a neural network to identify potential hla-dr1 binding sites within proteins. *J Mol Recognition*, 6:41–48, 1994.
- [75] M. Bhasin and G. P. Raghava. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics*, 20(3):421–423, 2004.
- [76] O. Karpenko, J. Shi, and Y. Dai. Prediction of MHC class II binders using the ant colony search strategy. *Artif Intell Medicine*, 35(1-2):47–56, 2005 Sep-Oct.
- [77] Y. Altuvia, A. Sette, J. Sidney, S. Southwood, and H. Margalit. A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. *Hum Immunol*, 58:1–11, 1997.
- [78] H. H. Bui, A. J. Schiewe, H. von Grafenstein, and I. S. Haworth. Structural prediction of peptides binding to MHC class I molecules. *Proteins*, 63:43–52, 2006.

- [79] T. Lengauer I. Antes, S. W. Siu. DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. *Bioinformatics*, 22(14):e16–24, 2006.
- [80] O. Schueler-Furman, Y. Altuvia, A. Sette, and H. Margalit. Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci*, 9(9):1838–1846, 2000.
- [81] O. Schueler-Furman, R. Elber, and H. Margalit. Knowledgebased structure prediction of MHC class I bound peptides: a study of 23 complexes. *Fold Des*, 3:549–564, 1998.
- [82] J. C. Tong, T. W. Tan, and S. Ranganathan. Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci*, 13:2523–2532, 2004.
- [83] D. Hammerstrom. Neural networks at work. *IEEE Spectrum*, 30(6):4355–8, 1993.
- [84] B. Peters, W. Tong, J. Sidney, and et al. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics*, 19:1765–72, 2003.
- [85] J. A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4957):1285–1293, 1988.
- [86] A. P. Engelbrecht. *Fundamentals of computational swarm intelligence*:. Wiley, 2005.
- [87] C. A. Coello. Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art. *Computer Methods in Applied Mechanics and Engineering*, 191(11):1245–1287, 2002.

- [88] E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: A comparative case study and the strength of pareto approach. *IEEE Trans on Evolutionary Computation*, 3:257–271, 1999.
- [89] J. D. Knowles. Approximating the nondominant front using the Pareto archived evolution strategy. *Evolutionary Computation, MIT Press edn*, 8:49–172, Summer 2000.
- [90] C. Fonseca and P. J. Fleming. Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization. In *fifth Intl conference on Genetic Algorithms*, pages 416–423, San Mateo, CA, Morgan Kauffman, 1993.
- [91] J. S. Dyer, P. C. Fishburn, R. E. Steuer, J. Wallenius, and S. Zionts. Multiple criteria decision making, multiattribute utility theory: the next ten years. *Management Science*, 38(5):645–654, 1992.
- [92] V. Belton and T. J. Stewart. Multiple criteria decision analysis: an integrated approach. Springer-Verlag, Berlin, Germany, 2002.
- [93] M. A. Lones and A. M. Tyrrell. The evolutionary computation approach to motif discovery in biological sequences. *Genetic and Evolutionary Computation*, pages 1–11, 2005.
- [94] D. Corne, A. Meade, and R. Sibly. Evolving core promoter signal motifs. *IEEE Congress on Evolutionary Computation*, pages 1162–1169., 2001.
- [95] R. G. Beiko and R. L. Charlebois. GANN: genetic algorithm neural networks for the detection of conserved combinations of features in DNA. *BMC Bioinformatics*, 6(1):36, 2005.
- [96] 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*, 32(13):3826–3835, 2004.

- [97] F. Liu, J. Tsai, R. Chen, S. Chen, and S. Shih. FMGA: Finding motifs by genetic algorithm. In *IEEE BIBE*, 2004.
- [98] N. Lo, S. Changchien, Y. Chang, and T. Lu. Human promoter prediction based on sorted consensus sequence patterns by genetic algorithms. In *Intl congr on Biological and Medical Engineering*, pages 111–112, 2002.
- [99] D. S. Che, Y. L. Song, and K. Rasheed. MDGA: motif discovery using a genetic algorithm. In *GECCO*, pages 447–452, 2005.
- [100] J. Hammer, E. Bono, F. Gallazzi, C. Belunis, Z. Nagy, and F. Sinigaglia. Antigenic peptide binding by class I and class II histocompatibility proteins—precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J Exp Med*, 180(6):2353–2358, 1994.
- [101] G. P. S. Raghava. Mhcbench. <http://www.imtech.res.in/raghava/mhcbench>.
- [102] N. Murugan and Y. Dai. Prediction of MHC class II binding peptides based on an iterative learning model. *Immunome Res*, 1(6):10, 2005.
- [103] J. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.
- [104] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, May 1983.
- [105] G. B. Sorkin. Efficient simulated annealing on fractal energy landscapes. *Algorithmica*, 6:367–418, 1991.
- [106] S. Lim and B. Tan. Performance of the genetic annealing algorithm in DFM synthesis of dynamic musical sound samples. *J Audio Eng Soc*, 57(5), 1999.

- [107] B. Tan and S. Lim. Automated parameter optimization for double frequency modulation synthesis using the genetic annealing algorithm. *J Audio Eng Soc*, 44:3–15, 1996.
- [108] G. Fogel and D. Corne. *Evolutionary Computation in Bioinformatics*. Morgan Kaufman publishers, 2003.
- [109] K. Deb and A. Kumar. Real-coded genetic algorithms with simulated binary crossover: Studies on multi-modal and multi-objective problems. *Complex Systems*, 9(6):431–454, 1995.
- [110] K. Deb and R. B. Agrawal. Simulated binary crossover for continuous search space. *Complex Systems*, 9:115–148, 1995.
- [111] R. Tisch and H. McDevitt. Insulin-dependent diabetes mellitus. *Cell*, 85:291–297, 1996.
- [112] C. A. Aoki, A. T. Borchers, W. M. Ridgway, C. L. Keen, A. A. Ansari, and M. E. Gershwin. *Autoimmun Rev*, 4:373–379, 2005.
- [113] T. Aly, D. Devendra, and G. S. Eisenbarth. Immunotherapeutic approaches to prevent, ameliorate, and cure type 1 diabetes. *Am J Ther*, 12:481–490, 2005.
- [114] E. P. Reich, H. von Grafenstein, A. Barlow, K. E. Swenson, K. Williams K, and C. A. Janeway. Self peptides isolated from MHC glycoproteins of non-obese diabetic mice. *J Immunol*, 152:2279–2288, 1994.
- [115] 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*, 9(1):43–51, 1997.

- [116] 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*, 185(6):1013–1021, 1997.
- [117] 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*, 12(6):699–710, 2000.
- [118] 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*, 12(4):493–503, 2000.
- [119] A. Corper, T. Stratmann, V. Apostolopoulos, C. A. Scott, K. C. Garcia, A. S. Kang, I. A. Wilson, and L. Teyton. A structural framework for deciphering the link between I-Ag7 and autoimmune diabetes. *Science*, 288(5465):505–511, 2000.
- [120] 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*, 30(9):2497–2506, 2000.
- [121] T. Stratmann, V. Apostolopoulos, V. Mallet-Designé, A. L. Corper, C. A. Scott, I. A. Wilson, A. S. Kang, and L. Teyton. The I-Ag7 MHC class II molecule linked to murine diabetes is a promiscuous peptide binder. *J Immunology*, 165(6):3214–3225, 2000.

- [122] S. Southwood, J. Sidney, A. Kondo, M. del Guercio, Appella E, S. Hoffman, R. T. Kubo, R. Chestnut, H. M. Grey, and A. Sette. Several common HLA-DR types share largely overlapping peptide binding repertoires. *Immunology*, 160:3363–3373, 1998.
- [123] A. Geluk, K. van Meijgaarden, N. Schloot, J. Drijfhout, T. Ottenhoff, and B. Roep. Hla-dr binding analysis of peptides from islet antigens in iddm. *Diabetes*, 47:1584–1600, 1998.
- [124] M. Nielsen, C. Lundegaard, and O. Lund. NetMHCII. <http://www.cbs.dtu.dk/services/NetMHCII/>, 2007.
- [125] 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*, 156(8):3000–3008, 1996.
- [126] K. R. Prilliman, M. Lindsey, K. W. Jackson, J. Cole, R. Bonner, and W. H. Hildebrand. Complexity among constituents of the HLA-B*1501 peptide motif. *Immunogenetics*, 48:89–97, 1998.
- [127] I. A. Doytchinova, V. A. Walshe, N. A. Jones, S. E. Gloster, P. Borrow, and D. R. Flower. Coupling in silico and in vitro analysis of peptide-mhc binding: A bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes. *J. Immunology*, 2004.
- [128] S. Heinkoff and J. Heinkoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.*, 89:10915–10919, 1992.
- [129] U. Hobohm, M. Scharf, R. Schneider, and C. Sander. Selection of representative protein datasets. *Protein Sci.*, 1:409–417, 1992.

- [130] S. F. Altschul, T. L. Madden, A. A. Schffer¹, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.
- [131] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its sup-plement TrEMBL. *Nucleic Acids Research*, 28:45–48, 2000.
- [132] A. F. Neuwald, J. S. Liu, and C. E. Lawrence. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*, 4(8):1618–1632, 1995.
- [133] A. Suri, I. Vidavsky, K. van der Drift, O. Kanagawa O, 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*, 168(3):1235–1243, 2002.
- [134] I. A. Doytchinova and D. R. Flower. Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex: A three-dimensional quantitative structure-activity relationship study. *Proteins: Structure, Function, and Genetics*, 48:505–518, 2002.
- [135] A. Suri, J. J. Walters, M. L. Gross, and E. R. Unanue. Natural peptides selected by diabetogenic DQ8 and murine I-Ag7 molecules show common sequence specificity. *J. Clin. Invest.*, 115:2268–2276, 2005.
- [136] W. Liu, X. Meng, Q. Xu, D. R. Flower, and T. Li. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (svr) models. *BMC Bioinformatics*, 7:182, 2006.
- [137] J. Cui, L. Y. Han, H. H. Lin, Z. Q. Tang, L. Jiang, Z. W. Cao, and Y. Z. Chen. MHC-BPS: MHC-binder prediction server for identifying peptides of flexible lengths from sequence-derived physicochemical properties. *Immunogenetics*, 58:607–613, 2006.

- [138] J. Cui, L. Y. Han, H. H. Lin, H. L. Zhang, Z. Q. Tang, C. J. Zheng, Z. W. Cao, and Y. Z. Chen. Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical peroperties. *Molecular Immunology*, 44:866–877, 2007.
- [139] C. Mnz, M. Hofmann, K. Yoshida, A. K. Moustakas, H. Kikutani, S. Stevanovi, G. K. Papadopoulos, and H. Rammensee. Peptide analysis, stability studies, and structural modeling explain contradictory peptide motifs and unique properties of the NOD mouse MHC class II molecule H2-Ag7. *Eur J Immunol*, 32:2105–2116, 2002.
- [140] W. R. Taylor. The classification of amino acid conservation. *J. Theor. Biol.*, 119:206–218, 1986.
- [141] A. A. Zamyatnin. Protein volume in solution. *Prog. Biophys. Mol. Biol.*, 24:107–123, 1972.
- [142] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch. Protein identification and analysis tools on the ExPASy server. In *The Proteomics Protocols Handbook*, Humana Press, pages 571–607, John M. Walker (ed), 2005.
- [143] M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, Inc., New York, 1991.
- [144] D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, 1990.
- [145] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. Weblogo: A sequence logo generator. <http://weblogo.berkeley.edu/>, 2004.
- [146] S. Tenzer, B. Peters, S. Bulik, O. Schoor, C. Lemmel, M. M. Schatz, P. M. Kloetzel, H. G. Rammensee, H. Schild, and H. G. Holzhutter. Modeling the

mhc class i pathway by combining predictions of proteasomal cleavage, tap transport and mhc class i binding. *Cell Mol Life Sci*, 62:1025–1037, 2005.

Appendix

Guided-Discovery

I-A^{g7}

Motif derived for I-A ^{g7} by using MOEA guided-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	103.8	114.9	26.26	44.96	35.6	68.66	91.87	41.37	88.79	70.58	15.82	41.34	52.41	39	114.6	108.5	93.48	42.49	84.77	42.66
P2	76.76	41.44	65.38	19.28	60.54	19.49	123.2	42.52	70.85	41.92	67.74	26.91	44.52	64.74	92.26	78.48	82.16	77.21	39.82	75.3
P3	78.4	23.62	43.07	53.57	52.8	42.56	55.54	116.3	69.28	58.38	21.44	47.98	73.83	97.63	110	101.9	19.78	90.6	42.33	79.81
P4	120.3	39.11	40.35	42.42	78.59	19.4	42.55	105.8	43.75	93.4	22.52	54.89	110.2	30.31	106.2	98.11	70.85	83.91	105.1	43.54
P5	34.34	43.49	47.04	78.3	39.37	43.5	83.24	42.21	40.77	66.96	22.25	84.21	54.41	40.32	58.29	107	41.85	110	41.3	57.45
P6	44.73	42.14	63.01	3.29	44.4	35.94	39.31	39.85	70.17	74.13	42.79	92.1	43.06	76.39	95.15	42.8	67.5	124.4	76.89	61.1
P7	107.3	38.71	84.71	4.14	76.94	86.31	44.76	77.22	41.4	48.91	54.5	47.86	103.6	69.76	90.08	56.63	42.7	72.44	44.02	122.4
P8	71.19	43.75	90.99	37.82	42.3	24.68	79.13	64.54	101.8	78.23	30.92	47.73	103.3	50.96	41.49	55.8	31.99	98.81	115.9	92.3
P9	114.6	59.31	101	43.54	43.54	102.3	41.93	98.59	95.41	71.62	47.86	43.34	31.16	24.58	119.6	42.49	32.49	42.53	35.59	92.04

Self Discovery

I-A^{g7}

Motif derived for I-A ^{g7} by using GAA																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	123	57	18	69	9	43	23	116	66	112	48	81	35	56	64	79	120	72	119	40
P2	60	64	56	41	92	127	105	20	120	107	8	59	125	59	35	97	5	47	27	94
P3	95	65	96	98	7	41	97	121	85	117	6	102	42	127	121	88	127	112	17	82
P4	47	1	9	77	109	50	73	84	62	126	24	114	37	17	126	74	60	61	122	96
P5	63	13	85	0	17	1	60	23	29	27	71	89	56	57	127	117	61	113	39	51
P6	50	97	40	3	41	14	120	119	64	70	60	89	98	24	106	107	6	1	94	7
P7	28	119	56	65	102	81	52	102	119	90	120	27	34	82	67	4	61	127	95	122
P8	112	7	125	88	1	81	126	61	115	54	91	9	113	101	80	34	86	40	61	33
P9	117	9	68	9	80	122	74	78	63	92	6	62	100	96	96	127	48	127	121	67

Motif derived for I-A ^{g7} by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	59.20	87.70	5.95	11.00	17.70	44.40	15.90	27.50	69.80	69.00	29.90	45.60	119.00	126.00	126.00	111.00	77.00	15.30	97.20	73.50
P2	1.11	62.10	102.00	1.27	2.24	0.43	127.00	11.20	47.00	116.00	48.80	5.56	80.70	60.70	104.00	125.00	41.60	68.80	45.50	127.00
P3	29.70	8.38	22.40	7.46	19.40	94.50	52.80	116.00	3.42	124.00	45.70	36.00	75.20	56.60	116.00	117.00	95.40	104.00	27.90	85.00
P4	75.80	55.20	25.60	57.30	53.00	0.93	18.80	91.60	41.70	104.00	32.30	79.00	111.00	4.65	123.00	115.00	34.40	88.70	104.00	24.00
P5	39.60	107.00	46.20	8.08	87.60	94.40	45.60	116.00	2.18	92.30	98.60	111.00	69.00	68.40	103.00	96.90	76.60	74.20	2.22	24.60
P6	97.90	9.76	118.00	31.60	3.75	109.00	26.00	6.92	10.70	78.10	29.00	91.20	110.00	22.40	85.90	24.40	92.50	126.00	85.00	0.07
P7	114.00	104.00	32.90	46.80	20.30	106.00	50.60	125.00	15.10	54.20	16.00	17.10	86.70	80.70	30.10	127.00	20.00	102.00	6.28	117.00
P8	77.40	126.00	23.90	9.21	58.50	24.90	126.00	93.10	59.00	98.10	78.60	37.20	97.90	5.06	120.00	78.00	64.00	126.00	121.00	127.00
P9	116.00	19.50	57.40	0.75	38.40	69.70	95.80	118.00	89.30	17.70	1.44	112.00	21.20	2.06	91.90	51.90	13.80	85.30	21.30	104.00

BM-SET1 DRB1-0401 Allele

Motif derived for HLA-DRB1*0401 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	13.18	55.81	2.83	40.19	126.1	5.06	4.3	43.93	74.72	0.19	101.1	6.34	14.2	53.67	0.74	99.84	4.79	23.79	112.7	122.5
P2	3.77	8.93	35.54	5.5	121.2	9.19	7.15	24.42	5.49	71.8	9.28	3.65	2.99	2.94	11.68	88.89	4.76	124.5	121.6	126.4
P3	57.11	14.34	39.38	57.2	80.07	1.69	43.95	37.55	59.89	100.6	102.8	18.26	4.19	81.51	10.15	1.62	4.16	125.9	35.92	52.16
P4	117.6	15.76	11.65	126	7.31	2.27	27.18	15.67	101.6	91.23	2.9	125.5	1.32	68.12	44.03	0.96	17.18	14.02	97.05	7.9
P5	1.92	35.43	83.06	124.7	115.5	2.94	9.81	97.04	7.07	25.65	21.9	8.33	1.29	124.3	0.69	121.9	11.06	107	27.12	63.29
P6	1.95	71.97	99.53	2.57	71.92	1.89	21.52	8.27	3.03	29.96	15.39	126.1	33.27	31.24	1.4	70.69	122.4	111.7	2.98	3.49
P7	1.77	13.17	10.81	31.99	28.72	3.41	8.73	5.13	0.45	90.48	1.87	124.2	2.38	111.8	12.67	110.5	116.5	47.16	106.9	28.45
P8	67.16	11.56	101.5	50.61	33.46	1.93	26.66	14.03	0.29	106.2	43.27	14.9	110.4	85.5	72.52	16.65	41.41	12.85	99.89	6.88
P9	9.35	67.04	4.65	7.81	6.7	4.36	100.8	7.58	113.6	19.99	56.4	39.04	3.28	115	27.33	40.9	76.66	2.05	52	33.62

BM-SET2 Alleles

Motif derived for I-A ^b by using MOEA self-discovery																					
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
P1	1.03	52.28	25.77	0.44	2.36	1.82	6.56	0.42	3.4	16.91	122.6	7.55	37.78	0.4	0.09	0.37	0.66	40.78	7.73	125.8	
P2	0.31	107.4	3.05	0.57	118.3	1.01	96.34	0.61	0.57	1.74	0.43	11.95	5.88	1.83	125.4	0.08	0.24	4.95	19.01	2.31	
P3	10.02	34.41	0.11	11.58	124	0.4	5.31	0.45	3.37	30.46	126.2	114.4	0.01	3.5	0.13	0.53	33.19	0.47	13.54	78.81	
P4	125.2	6.84	0.18	6.31	92.24	0.92	0.12	0.54	0.31	28.44	0.86	52.04	1.36	0.06	0.02	3.08	0.69	4.05	19.33	0.13	
P5	0.98	21.16	13.23	4.22	40.03	1.11	0.15	2.58	0.09	0.22	0.06	116.8	7.36	1.27	2.28	27.85	110.5	0.51	125.4	0.2	
P6	0.08	0.03	0.14	74.34	6.36	7.35	82.47	0.18	0.18	3.86	1.19	4.77	126.6	0.36	0.23	68.48	55.1	0.03	0.71	102.2	
P7	2.16	2.94	0.77	0.89	0.06	0.47	34.23	1.81	1.28	100.8	0.32	0.8	59.18	3.85	1.75	125.2	0.52	1.45	41.9	0.52	
P8	0.01	78.26	5.39	95.4	2.77	1.04	25.91	0.06	0.62	100.8	0.85	0.37	0.58	0.48	0.06	0.62	1.76	38.34	59.94	4.96	
P9	3.47	53.29	0.31	62.03	0.43	0.67	42.64	32.78	0.02	1.44	125.4	7.58	1.39	5.51	0.24	32.96	0.33	0.37	76.63	4.03	

Motif derived for I-A ^d by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	56.77	53.35	7.31	85.38	0.53	1.27	19.72	56.55	0.4	0.2	125.7	6.19	5.3	4.18	24.89	121.8	1.05	102.6	6.44	8.8
P2	101.9	4.88	80.47	104.8	13.02	4.42	57.24	13.2	101.9	28.65	20.87	2.19	1.87	2.27	126.9	0.95	4.2	7.49	12.23	0.15
P3	121.9	50.48	108.6	55.98	7.47	19.59	1.19	31.2	48.17	69.36	92.68	98.28	69.65	3.56	67.14	74.38	85.45	0.66	53.75	41.48
P4	126.8	30.7	9.15	3.02	31.36	1.6	93.5	126.8	64.39	57.92	125.5	2.43	42.18	49.32	17.42	42.9	1.46	14.53	95.55	0.65
P5	123.6	26.64	123.1	11.13	8.48	0.19	0.54	52.42	105.1	3.16	14.66	34.37	3.04	20.79	110.5	5.37	97.53	119.9	19.59	32.46
P6	70.96	7.32	8.06	12.96	37.05	0.64	39.86	116.1	4.16	21.37	77.75	1.42	89.26	21.43	0.22	126.7	1.98	126.5	3.95	59.55
P7	1.12	0.13	95.77	0.6	111.3	3.35	48.86	15.22	51.51	68.63	3.01	17.26	0.39	53.71	5.52	106.2	18.31	126.1	89.54	114.2
P8	123.3	48.13	118.2	0.75	47.61	100.5	1.65	2.2	31.94	115.7	30.01	0.71	1.82	87.31	100.7	107.2	6.81	38.09	20.69	99.07
P9	126.5	27.77	16.43	0.68	7.03	100.5	31.89	2.28	0.29	66.59	105.9	21.34	6.37	47.56	125.6	112.3	78.89	45.85	42.52	14.64

Motif derived for I-A ^s by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	25.88	25.64	3	30.03	0.07	126.9	8.66	116.2	2.66	0.17	10.92	1.16	32.87	21.01	41.97	3.7	28.93	21.61	31.69	0.07
P2	26.01	5.36	2.75	12.38	6.34	16.21	10.97	34.43	124.4	2.11	6.03	6.12	41.54	13.95	42.91	1.01	1.35	0.79	50.98	26.82
P3	17.2	13.53	3.36	22.21	91.54	108	9.45	20.74	59.32	3.46	17.98	0.82	50.97	0.3	6.22	1.36	12.84	10.43	108.3	43.14
P4	3.67	51.56	1.67	9.03	24.89	24.4	12.45	118	0.78	103.8	23.96	0.58	12.1	17.88	45.06	0.05	0.26	29.39	22.88	2.77
P5	0.33	6.68	47.15	9.46	51.22	8.11	10.9	62.15	31.59	3.26	40.85	14.6	7.8	3.8	4.79	11.45	106.3	62.97	1.43	12.7
P6	23.45	0.7	0.78	43.38	0.3	1.02	121.6	2.15	124.4	0.39	95.92	36.66	0.63	6.2	1.52	11.5	16.6	1.67	4.69	23.11
P7	27.15	17.41	6.41	14.47	46.92	0.51	27.74	3.15	14.57	0.65	119.1	1.49	126.9	38.64	0.47	0.56	10.34	1.49	8.09	0.39
P8	44.8	2.41	126.9	1.08	8.85	0.54	1.61	16.98	1.26	0.37	1.2	64.16	126.2	5.21	1.92	0.47	11.1	1.44	7.99	4.8
P9	2.67	5.4	31.11	50.09	19.58	1.87	0.21	115.3	126.6	3.65	121.3	0.74	125.5	32.93	0.62	1.64	1.68	4.92	46.25	0.27

Motif derived for DRB1-0101 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	7.19	32.96	2.94	2.05	115.8	100.5	71.13	100.7	0.29	55.94	91.31	50.99	45.17	45.46	43.23	32.35	17.3	59.62	25.3	30.3
P2	0.47	29.34	7.02	0.06	36.82	94.45	8.47	48.83	96.7	10.61	0.72	27.64	6.81	73.73	1.35	71.84	33.84	127	82.51	3.26
P3	94.59	3.2	0.54	1.89	2.52	118.8	17.13	2.25	22.84	124.5	112	70.57	19.94	56.93	39.31	125.4	115.7	84.27	37.46	2.28
P4	63.6	0.43	4.35	36.28	0.12	24.81	126.3	9.78	0.38	87.65	121.4	0.18	124.9	45.95	93.56	27.7	50.76	16.65	17.71	27.43
P5	102.5	30.73	43.97	0.83	17.18	106.2	20.45	20.16	0.78	117.1	120.9	3.7	77.42	27.21	111.6	24.43	6.67	2.61	10.09	1.12
P6	126.4	0.26	2	13.82	35.68	81.96	5.66	18.04	2.7	11.76	125.4	8.96	2.16	6.57	45.2	118.8	121.6	88.11	9.77	33.57
P7	82.41	38.42	0.03	84.49	11.97	8.75	42.17	2.63	23.54	10.19	83.81	11.69	89.14	40.82	15.76	124.6	28.46	99.71	3.18	105.4
P8	0.11	85.73	0.81	0.39	68.73	5.28	107.1	52.87	0.14	125.6	62.61	44.74	66.25	125.5	120.4	86.13	112.6	122.8	123.9	8.46
P9	119.6	39.81	1.8	2.28	35.94	10.36	23.31	7.77	2.03	126.8	3.95	76.8	3.46	23.59	73.53	2.95	124.4	122.1	94.89	1.7

Motif derived for DRB1-0301 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	8.73	18.03	111	72.18	9.86	1.83	3.4	55.37	22.22	33.72	125.5	1.56	58.48	28.22	94.94	17.75	126.3	0.55	112.6	0.02
P2	1.41	7.08	58.56	2.2	32.03	0.99	37.87	69.59	17.35	31.89	126.5	21.82	1.93	7.6	126.8	0.04	1.98	24.06	34.45	4.98
P3	12.51	8.31	61.39	0.49	127	3.67	16.66	11	51.75	2.18	126.7	18.75	6.86	1.55	2.02	21.8	48.55	0.04	2.8	3.29
P4	88.78	0.85	117	13.15	92.58	13.7	13.62	64.74	52.84	2.11	34.14	109.4	59.87	13.12	37.08	0.31	49.59	2.03	5.35	56.83
P5	2.02	6.19	124.4	9.88	22.07	0.33	12.61	17.33	105.5	1.47	53.07	9.4	2.49	55.12	114.6	34.28	123.9	0.16	48.2	81.5
P6	62.55	8.02	117.7	118.4	20.29	1.46	5.03	11.76	13.56	52.19	4.65	7.37	24.28	64.61	109.3	2.78	121.8	5.12	8.95	0.25
P7	43.71	42.64	4.88	48.36	13.28	0.07	0.41	75.31	0.44	127	36.65	3.33	7.51	0.28	32.91	9.85	56.2	61.94	95.47	2.26
P8	126.4	47.08	2.94	0.17	13.75	0.44	51.03	0.04	38.19	101	121.4	2.4	75.96	23.01	125.6	6.62	11.72	48.98	4.79	32.08
P9	112.8	1.03	4.78	0.05	58.52	0.92	14.76	1.49	0.93	1.61	126.5	43.92	0.08	46.51	125.5	1.69	48.01	60.19	54.81	1.32

Motif derived for DRB1-0401 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	47.15	1.96	5.81	14.26	43.18	3.27	111.4	52.59	117.8	69.75	126.8	2.56	6.72	8.25	59.59	27.8	9.89	97.73	16.88	107.2
P2	0.21	32.71	52.97	12.76	119.9	6.04	83.47	7.58	119.5	25.26	94.58	8.11	1.35	24.46	108.9	62.38	31.89	110.9	14.5	123.4
P3	51.51	35.56	115.7	26.42	123.6	8.88	27.25	2.14	3.31	17.61	119.5	77.92	2.83	94.18	68.21	46.6	75.8	59.62	27.96	89.97
P4	12.96	33.55	29.04	6.02	122.3	2.74	25.06	20.96	10.55	115.6	62.07	17.74	9.69	114.9	56.03	9	123.8	1.71	119.4	107.9
P5	51.64	6.53	1.8	85.16	109.8	2.75	17.56	7.48	45.5	3.68	78.58	5.42	6.2	59.23	14.23	120.8	107.7	57.56	63.16	94.69
P6	31.1	25.29	2.67	13.26	26.11	35.11	6.89	72.38	118.2	52.11	121.7	11.7	113.3	88.24	32.69	87.35	68.5	2.21	126	89.58
P7	0.97	0.34	22.61	3.85	9.23	0.4	0.68	91.03	51.48	23.25	110.6	126	17.02	6.49	6.54	107.9	126.8	28.61	20.23	7.45
P8	113.2	2.77	8.33	1.69	17.54	4.42	0.03	3.95	24.51	106.8	126.4	83.04	101	31.23	32.58	118.5	102.8	38.73	115	32.27
P9	124	5.14	0.98	0.41	24.69	4.94	101.9	99.34	1.21	25.22	117.4	106.7	1.89	1.16	78.69	119.6	121.5	15.39	2.28	71.39

Motif derived for DRB1-0404 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	9.09	42.16	59.93	81.23	18.78	35.71	9.34	57.39	36.52	121.4	102	121.4	3.79	87.62	48.76	13.85	70.37	122.5	11.77	19.43
P2	99.89	108.7	32.8	1.4	24.62	0.46	49.91	118.1	123.1	126.9	122	49.8	125.7	2.36	58.73	78.01	1.74	120.1	0.51	13.48
P3	4.81	22.87	22.6	73.23	34.32	78.45	17.99	124.3	100.2	92.58	109.4	126.4	91.02	32.16	83.54	18.34	1.38	126.4	63.43	8.05
P4	54.34	8.72	23.81	8.91	3.52	34.74	4.36	60.18	4.3	124.9	119.5	34.04	77.67	64.12	4.51	68.23	76.84	102.3	37.76	122.6
P5	19.95	54.85	83	25.43	86.93	13.13	86.24	13.01	15.36	100.2	111.9	34.69	30.46	79.79	117.4	122.1	111.3	1.77	26.47	0.11
P6	107.6	2.66	7.86	2.47	112.7	51.72	0.33	9.51	35.93	118.1	123.6	89.06	31.44	17.65	1.87	81.31	125.1	112.2	76.01	22.82
P7	122.7	1.62	1.85	0.26	43.27	73.77	33.73	111.5	105	20.77	39.63	91.91	1.19	12.47	5.51	122.3	24.28	2.05	3.42	125.3
P8	122.6	4.2	24.03	0.91	103.4	43.21	20.95	59.15	83.75	30.99	66.32	71.38	35.01	16.77	56.66	42.15	124.7	63.4	4.3	40.67
P9	107.1	14.2	1.62	7.6	66.32	116.4	74.07	75.46	8.63	49.84	100.2	126.2	113.7	31.67	28.16	24.25	79.82	33.25	73.58	45.09

Motif derived for DRB1-0405 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	0.22	9.37	18.52	115.3	1.16	14.29	34.28	102.4	72.91	1.76	8.3	0.33	0.87	2.36	58.24	93.25	52.09	106	50.65	42.19
P2	2.36	0.76	0.13	82.66	0.35	0.09	20.33	14.7	1.65	1.25	0.12	52.89	0.97	108.7	123.7	78.05	11.99	2.33	5.61	2.86
P3	11.71	1.48	14.8	17.09	96.54	4.64	2.07	99.71	0.22	0.35	126.9	8.11	5.38	0.06	18.01	0.07	0.17	1	107.9	126.7
P4	1.12	4.17	1.26	19.31	119.5	1.39	3.15	106.5	33.47	1.42	59.23	124.9	94.05	117.6	0.11	23.37	72.92	11.31	121.7	125.3
P5	1.08	3.92	0.71	0.45	13.67	0.07	0.01	10.51	58.06	41.45	3.47	34.42	120.8	29.85	1.33	126.6	126.5	11.29	126.2	121.2
P6	0.25	14.15	50.66	0.72	32.48	0.82	6.37	105.1	26.32	2.59	82.72	55.41	79.53	115.3	1.1	47.51	95.23	9.67	123	104.8
P7	1.57	9.98	0.81	0.38	23.5	9.75	0.16	80.88	15.38	97.63	6.78	2.52	123.1	124.3	2.11	122.1	118.1	3	8.25	28.81
P8	0.28	47.26	73.76	1.61	1.19	42.45	1.28	12.48	3.33	33.81	7.12	78.9	117.6	99.14	11.13	126.5	118.5	18.67	2.24	6.08
P9	20.32	5.03	4.41	87.37	22.64	3.29	12.7	126.4	28.24	114.2	84.67	116.6	2.14	2.74	1.12	21.18	47.86	44.39	1.72	2.71

Motif derived for DRB1-0701 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	61.38	18.24	12.49	33.97	122	21.33	76.11	4.36	122.9	61.77	42.98	16.15	91.91	3.55	9.18	6.97	13.35	99.25	52.51	117.7
P2	20.5	24.2	19.27	28.12	105.8	4.66	3.82	105	98.23	52.2	117.6	62.01	94.5	35.13	88.55	35.85	108.2	89.42	63.86	123.8
P3	8.43	38.6	3.59	4.3	51.76	1.71	59.43	4.34	77.58	9.67	89.14	125.1	92.33	9.53	51.77	4.59	99.52	126.9	120.3	126.7
P4	121.6	17.72	10.31	3.77	62.29	5.36	98.41	50.2	106.4	12.2	16.52	50.26	5.27	1.87	125.8	13.66	126.1	75.87	6.51	29.72
P5	126.7	56.68	23.62	3.99	34.44	4.68	3.57	88.48	27.99	7.59	114.9	101.1	17.15	0.99	1.91	118.5	123.2	13.75	36.24	10.51
P6	99.87	25.9	1.43	2.36	32.02	26.85	49.71	3.53	33.21	50.84	37.1	53.92	7.38	9.36	112.8	105.3	120.2	2.05	118.4	114.6
P7	12.5	0.6	7.86	3.68	13.44	45.48	45.34	0.74	15.66	22.22	123.2	1.8	17.35	4.69	95.9	126.2	99.58	125.9	86.16	27.56
P8	11.81	31.45	19.24	3.6	95.75	15.64	55.69	1.04	111.2	111.8	56.57	15.54	54.25	1.53	123.7	8.06	123.5	107.2	56.06	92.98
P9	96.73	35.12	10.43	2.34	89.71	8.1	22.73	71.87	63.19	67.33	119.4	13.04	30.86	3.17	1.48	14.68	118.2	113.3	21.89	72.28

Motif derived for DRB1-0802 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	1.37	1.26	0.55	0.33	86.27	28.15	0.08	33.09	102.6	0.5	120.2	93.8	0.11	85.45	61.39	0.54	16.1	0.88	125.4	126.9
P2	0.81	13.89	2.03	0.38	0.99	1.76	110	91.1	127	2.6	90.17	72.87	0.59	0.49	39.86	51.36	25.24	1.41	124.6	6.82
P3	126.7	0.9	89.06	4.27	55.52	3.66	78.93	102.8	2.69	0.22	8	9.11	4.14	3.83	0.32	0.44	118.7	18.69	40.27	3.31
P4	2.52	47.1	1.77	1.58	0.71	0.73	0.05	0.72	74.49	126.8	100.9	1.38	1.71	0.42	5.82	2.47	85.86	70.76	126.8	34.46
P5	4.06	3.17	2.6	0.79	1.7	72.81	61.97	6.6	125.5	11	8.64	0.31	115.1	0.85	122.9	1.74	41.63	73.64	66.86	0.22
P6	14.13	0.83	19.27	1.32	13.38	0.88	1.48	74.17	69.6	113.9	51.45	50.03	26.89	9.41	125.9	4.4	12.88	0.02	46.04	9.85
P7	109.9	3.08	0.16	53.92	3.65	11.58	19.31	0.04	122	0.84	14.58	123.5	6.5	0.33	19.94	16.59	29.57	0.31	59.71	111.1
P8	34.08	0.92	0.21	0.67	100	0.05	64.85	0.43	122.3	0.64	125.9	7.6	99	126.8	0.46	0.39	10.13	1.22	62.95	1.86
P9	106.6	5.69	2.28	0.43	55.36	76	0.06	79.93	118.4	2.85	0.35	0.6	76.62	103.6	101.2	106.9	23.69	2.07	86.44	34.99

Motif derived for DRB1-0901 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	41.21	14.87	1.17	0.95	58.12	72.73	57.84	36.7	126.8	1.44	124	2.72	6.24	60.88	0.99	31.93	0.93	126.4	126.7	0.17
P2	63.48	2.32	15.31	0.83	62.96	90.67	29.38	124.8	125.8	2.85	16.17	61.45	38.28	101.2	3.53	13.02	126.8	115.1	1.46	76.75
P3	38.96	15.63	6.14	1.18	90.27	0.13	1.39	2.46	23.29	19.2	126.3	126.5	1.97	7.39	9.31	116.8	48.74	56.4	8.03	87.95
P4	122.1	3.87	73.73	82.53	1.26	65.87	110.2	0.09	30.47	26.43	0.6	0.41	47.38	0.07	125.2	126.5	1	103.9	3.6	2.27
P5	117.1	13.49	0.14	2.37	0.46	124	35.36	5.13	125.2	0.48	72.37	0.08	126.3	0.33	0.34	126.8	0.13	0.9	42.99	13.74
P6	125.4	1.03	2.92	124	35.41	1.55	49.32	6.37	18.97	51.33	6.02	17.85	0.05	18.32	0.25	17.26	126.9	80.59	115.9	49.81
P7	106.9	21.9	1.74	0.67	0.34	121.2	2.26	33.93	60.9	39.28	2.71	34.63	57.32	68.71	1.67	112.4	0.38	109.1	65.58	42.43
P8	126.8	0.23	95.45	0.52	27.08	41.16	1.67	120.8	126.7	105.1	49.71	0.57	111	33.24	1.02	7.09	101.2	1.06	74.85	40.53
P9	98.49	0.09	5.65	0.37	45.03	0.31	37.28	105.4	1.38	97.2	66.57	4.68	16.98	53.8	1.71	118.9	109.1	100.7	93.99	94.26

Motif derived for DRB1-1101 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	100.3	5.57	5.66	2.66	118.9	0.37	5.65	0.36	116.2	117.3	12.53	75.14	1.44	0.91	21.14	1.64	37.79	0.48	118.1	125
P2	110.3	13.69	1.45	0.12	2.95	0.47	70.79	100.2	126	4.03	38.72	1.34	1.92	0.22	1.26	29.64	105.2	0.87	19.58	3.95
P3	1.07	17.51	44.89	0.53	90.91	8.54	0.12	45.35	8.9	34.64	118	31.61	2.77	2.17	14.34	6.63	126	126.9	125.1	12.29
P4	49.48	3.48	2.03	2.19	1.11	8.92	103.2	4.14	2.16	84.66	123.9	4.59	0.36	45.33	110.3	116.6	33.42	28.75	67.93	4.26
P5	0.17	67.16	23.32	16.64	15.8	64.66	94.85	0.8	0.62	0.16	125.3	96.36	107.2	10.41	125.3	1.84	2	123.1	126.8	12.91
P6	112.3	4.11	0.16	28.88	112.7	15.09	85.77	24.22	30.26	114.7	124.5	3.22	0.89	1	4.89	3.81	15.47	22.28	96.55	88.36
P7	105.6	41.14	4	74.12	22.14	2.28	12.49	0.97	1.52	0.24	77.41	126.7	2.37	0.88	91.84	82.92	1.39	62.57	3.6	27.49
P8	78.87	2.45	12.69	10.01	24.04	0.98	17.1	21.65	106.5	0.5	76.01	0.77	78.69	93.3	114.5	1.38	2.85	23.22	121.1	31.53
P9	124.3	2.66	3.96	79.4	13.97	0.46	10.25	121.8	0.81	15.15	14.68	7.87	17.42	0.27	52.2	3.8	92.08	4.38	125.8	85.69

Motif derived for DRB1-1302 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	0.6	6.96	6.14	0.79	3.32	101.4	29.4	117.3	126.6	108.2	116.2	99.44	21.38	69.68	51.37	0.08	0.02	9.86	12.66	4.99
P2	3.16	19.3	7.39	2.25	124	121	3.47	0.85	91.72	75.5	1.81	122	5.33	93.16	1.06	14.06	4.89	73.15	69.7	2.21
P3	79.85	17.05	0.38	2.26	3.35	3.32	21.23	60.11	53.48	113.1	51.84	92.97	63.66	115.8	9.36	78.57	0.01	23.81	0.2	1.52
P4	48.64	0.38	63.84	4.53	1.05	0.05	1.92	116.7	0.03	100.6	126.6	126.7	10.61	9.1	0.08	33.9	24.45	78.33	2.45	84.66
P5	0.25	0.08	3.41	5.27	2.79	6.98	4.48	102	123.5	118.3	90.66	126.4	2.49	33.21	3.87	0.8	49.61	28.26	0.04	24.4
P6	0.21	21.99	0.69	1	3.39	110	42.94	126.9	17.62	20.7	123.9	126.6	50.61	1.21	1.57	112.1	37.21	13.45	0.24	3.09
P7	1.69	1.25	1.55	0.74	11.55	92.55	25.36	35.79	125.4	121.2	114.6	116.4	3.84	0.61	1.5	67.81	60.1	122.7	1.41	2.74
P8	0.04	1.49	26.83	22.23	57.33	1.76	1.34	78.65	7.36	125.9	100.2	124.4	3.13	1.27	49.13	42.37	11.1	105.9	38.87	4.37
P9	0.84	82.64	3.69	0.41	108	0.03	23.41	125.4	13.69	124.9	126.6	36.36	2.36	0.42	3.61	3.66	0.04	121.9	2.79	0.22

Motif derived for DRB1-1501 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	92.77	42.12	1.76	19.34	44.32	99.45	95.44	59.22	2.04	35.64	13.04	92.49	53.45	6.63	126.7	18.36	7.93	119.9	121.6	0.8
P2	51.55	9.92	27.79	2.38	10.34	42	126.7	104.6	102.6	53.56	17.93	47.41	0.52	111.3	11.43	0.47	10.97	103.8	6.78	57.71
P3	1.63	22.32	0.81	13.9	124.3	10.45	4.74	126.3	4.39	91.25	122.8	103.7	3.99	44.64	68.04	2.58	0.57	1.94	47.29	64.54
P4	10.76	2.02	25.33	0.54	122.7	36.25	4.83	0.93	95.48	1.63	57.9	9.87	0.23	43.98	0.74	10.13	18.96	108.4	120.6	0.25
P5	95.02	21.53	27.29	10.59	0.99	3.98	116.1	107	85.76	115.1	62.74	0.67	0.99	7.23	116.1	10.66	14.44	41.32	118.3	10.63
P6	10.33	15.04	12.27	2.27	93.96	77.22	21.33	40.57	99.08	124.8	117.8	123.8	0.85	0.28	14.26	9.75	48.52	27.51	106.4	6.02
P7	5.59	1.57	1.42	0.33	45.84	18.84	1.44	125.4	53.01	2.82	115.1	2.31	56.92	3.97	67.01	13.74	3.91	12.59	47.32	21.94
P8	3.39	30.14	84.1	0.29	12	3.89	4.31	104.9	4.88	80.56	92.63	3.75	8.57	0.29	7.75	117.6	21.18	124.2	55.28	12.09
P9	4.23	30.51	23.5	16.67	4.83	20.03	49.68	76.01	125.8	6.05	2.67	21.39	45.5	13.87	117.8	57.68	121.5	38.71	93.3	20.89

Motif derived for DRB4-0101 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	4.44	21.51	1.19	82.88	4.1	4.18	56.35	104.9	100.5	10.75	102.7	3.55	43.53	124.9	122.8	11.61	6.11	1.71	86.34	3.23
P2	4.06	2.81	45.09	0.49	10.3	9.77	20.27	75.73	74.73	118.8	78.52	57.93	56.55	119.1	2.48	4.33	4.44	62.45	111.4	5.26
P3	4.17	0.39	10.08	23.37	8.86	0.89	102.3	126.3	12.12	61.27	16.12	24.08	59.39	59.43	3.55	26.76	3.11	53.29	7.87	64.68
P4	7.63	5.5	17.69	1.54	24.81	1.44	84.99	108.1	67.12	63.04	65.29	14.31	1.09	13.47	121.4	111.9	0.86	2.84	19.09	122.2
P5	15.52	3.45	1.16	0.21	80.73	0	7.18	0.97	0.99	125.1	119	6.56	78	120.2	0.18	3.19	11.41	46.64	99.4	122.1
P6	79.6	120.4	1.27	0.12	2.49	29.35	7.03	49.7	9.78	95.32	125.7	1.98	5.95	88.8	87.05	3.42	60.48	8.03	123.5	105
P7	0.67	2.01	4.31	1	103.5	4.92	10.56	2.66	5.91	107.3	126.3	44.82	24.52	107.5	20.19	109	15	0.39	11.44	42.04
P8	9.17	80.19	7.93	4.09	99.47	0.64	46.81	94.96	0.07	126.3	116.9	124.6	90.41	126.7	9.06	3.99	0.9	2.86	14.98	1.62
P9	4.75	29.78	97.99	1.03	15.44	0.81	113.1	40.69	82.28	42.92	121.9	122.3	118.2	102.2	49.1	12	0.43	8.73	71.26	2.19

Motif derived for DRB5-0101 by using MOEA self-discovery																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	1.68	6	54.29	36.76	21.29	4.27	19.71	86.74	120.5	1.75	109.5	65.3	44.06	44.26	6.58	18.13	124.9	10.1	17.33	15.5
P2	0.6	16.79	29.04	3.22	106.2	7.56	113	80.75	55.35	126.9	120.5	116.4	90.2	7.29	3.2	9.36	42.2	1.42	4.29	84.1
P3	9.1	10.62	51.28	14.54	6.41	81.14	17.12	13.69	11.25	70.52	6.92	7.39	4.06	23.84	114.3	0.48	17.59	0.53	118.7	0.45
P4	94.63	2.76	17.45	0.64	11.63	4.29	91.55	0.9	38.42	81.49	17.79	97.89	0.19	116.8	15.99	125.4	10.19	2.72	43.26	98.3
P5	24.99	2.92	2.8	126.1	0.7	3.14	17.51	0.26	1.06	7.5	119	119.3	107.4	125.6	33.75	3.94	11.55	5.25	36.5	24.8
P6	95.16	10.75	2	8.34	46.34	1.88	87.68	96.23	92.31	3.34	15.36	1.88	24.2	116	2.33	0.92	5.39	51.58	103	33.7
P7	126	9.57	2.45	4.15	3.2	0.76	18.58	72.39	10.38	0.31	126.6	24.68	9.96	6.62	124.5	4.38	7.2	6.13	124.6	3.25
P8	0.46	75.19	98.23	1.12	126.1	2.35	53.39	0.36	123.4	0.48	5.73	19.31	30.86	24.24	11.58	102	11.1	110.2	106.8	44.7
P9	53.38	2.96	21.19	2.3	20.43	25.68	27.63	4.36	26.65	0.07	19.43	28.6	126.9	123	126	5.34	4.07	0.48	120.4	118