Published in final edited form as:
J Immunol Methods. 2011 November 30; 374(1-2): 18–25. doi:10.1016/j.jim.2011.07.007.

Dana-Farber Repository for Machine Learning in Immunology

Guang Lan Zhang[a,1], **Hong Huang Lin**[a,b,1], **Derin B. Keskin**[a], **Ellis L. Reinherz**[a], and **Vladimir Brusic**[a,*]

[a]Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston, MA 02115, USA

Abstract

The immune system is characterized by high combinatorial complexity that necessitates the use of specialized computational tools for analysis of immunological data. Machine learning (ML) algorithms are used in combination with classical experimentation for the selection of vaccine targets and in computational simulations that reduce the number of necessary experiments. The development of ML algorithms requires standardized data sets, consistent measurement methods, and uniform scales. To bridge the gap between the immunology community and the ML community, we designed a repository for machine learning in immunology named Dana-Farber Repository for Machine Learning in Immunology (DFRMLI). This repository provides standardized data sets of HLA-binding peptides with all binding affinities mapped onto a common scale. It also provides a list of experimentally validated naturally processed T cell epitopes derived from tumor or virus antigens. The DFRMLI data were preprocessed and ensure consistency, comparability, detailed descriptions, and statistically meaningful sample sizes for peptides that bind to various HLA molecules. The repository is accessible at http://bio.dfci.harvard.edu/DFRMLI/.

1. Introduction

The immune system is characterized by high combinatorial complexity mandating the use of specialized computational tools for the analysis of immunological data (Petrovsky and Brusic, 2002). The immunoinformatics applications include *in silico* tools such as computational models, prediction systems, and simulators that complement experimentation (Pappalardo *et al*., 2009). Machine learning (ML) in bioinformatics plays an important role, principally in developing accurate *in silico* methods for bioinformatics (Baldi and Brunak, 2001; Zhang and Rajapakse, 2009). The applications of ML have proven valuable in immunology; examples include the analysis of antigens (Lafuente and Reche, 2009), the analysis of allergenicity (Muh *et al*., 2009), the study of antibodies and their properties (David *et al*., 2010), the design of vaccine protocols (Palladini *et al*., 2010), and the classification of immunological profiles (Herz and Yanover, 2007). These developments help improve practical applications such as discovery, design, and optimization of vaccines. Vaccine development, however, is a complex task – the selection of components for actual vaccine formulations requires the analysis of a huge combinatorial space. With a few exceptions, the effectiveness of vaccines is limited to a subset of pathogens because of the

*Corresponding author: Tel.: +1 617 632 3824; fax: +1 617 632 3351. vladimir_brusic@dfci.harvard.edu (V. Brusic)..
[1]These authors contributed equally to this work.
[b]Current address: Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA

variability of pathogens and the variability of immune systems in humans (Brusic and August, 2004). Many of these vaccines require frequent re-formulation (as with influenza vaccines) or for some pathogens there is a complete lack of efficient and safe vaccines (such as dengue, West Nile, or HIV viruses). Finding the best combinations of components for vaccine formulation requires a significant amount of experimentation, advanced high-throughput instrumentation and bioinformatics analyses (Bambini and Rappuoli, 2009). The ML techniques play an increasingly important role both in the selection of suitable targets (molecules that are combined into vaccines) and in reducing the number of necessary experiments over the large combinatorial space of possible formulations.

1.1 Combinatorial properties of HLA and HLA-binding peptides

T cells of the immune system continually scan for the presence of foreign antigens that indicate the possible presence of invading microorganisms (viruses, bacteria, fungi, or parasites), transformed own cells (tumors), or cells from other organisms (transplants). T cell-based immune responses involve the recognition of short antigenic peptides presented by the major histocompatibility complex (MHC) molecules on the surface of target cells. T cells recognize these peptides (T cell epitopes) through highly specific interactions between T cell receptors and peptide/MHC complexes (Meuer *et al.*, 1982). Peptides presented by MHC molecules originate mainly from intracellular (MHC class I) or extracellular (class II) proteins. Typically, MHC class I mediated activation of cytotoxic T cells results in killing target cells, while MHC class II mediated activation is involved in regulation of immune responses. The ability of the immune system to respond to any given antigen varies between individuals because of differences in their human leukocyte antigen (HLA, human MHC) genes. Each human individual expresses three to six classical HLA class I molecules encoded at three loci: A, B, and C, and up to twelve classical HLA class II molecules encoded at DRA, DRB1, DRB3, DRB4, DRB5, DQA1, DQB1, DPA1, and DPB1 loci (Shiina *et al*., 2004). HLA genes demonstrate an extensive polymorphism with 4682 protein-coding variants of HLA class I genes characterized and named as of January 2011 (Robinson *et al*., 2011). These include 1519 HLA-A, 2069 -B, 1016 -C, and 78 non classical (HLA-E, -F and -G) proteins. There are 873 -DRB1, and 85 -DRB3/4/5 reported gene variants. In addition, more than 350 protein variants of HLA-DQ and -DP molecules have been characterized. Distinct protein-coding HLA alleles identified in US population by high-resolution typing include 100 HLA-A, 184 -B, 90 -C, 86 -DRB1, and 22 -DQB1 variants (Maiers *et al*., 2007). The HLA combination consisting of 1-2 HLA molecules for each classical locus (A, -B, -C, -DRBA, -DRB1, - DQA1, -DQB1, -DPA1, and -DPB1) and 0-2 molecules from three nonobligatory HLA loci (HLA-DRB3, -DRB4, and DRB5) determines both the individual's classical phenotype and the specific repertoire of antigenic targets for recognition by his/her immune system.

The majority of naturally processed HLA class I binding peptides are 8-11 amino acids long (Rammensee *et al*., 1999). These result from antigen processing pathways of target antigens (self, of foreign), including proteasome/TAP and a number of alternative pathways (Petrovsky and Brusic, 2004). Peptides that bind HLA Class II molecules are usually 12-20 amino acids long. They bind HLA class II molecules through a 9-mer binding core with flanking residues extending from the binding core outside of the HLA binding groove (Stern *et al*., 1994). It was proposed that flanking residues of HLA class II associated patterns reflect conserved antigen processing patterns (Godkin *et al*., 2001).

The total number of possible 8, 9, 10, and 11-mer peptides (search space for HLA class I binding peptides) is 2.16×10^{14}. The number of all possible 9-mer peptides (search space for binding cores of HLA class II binding peptides) is 5.12×10^{11}. The current theoretical number of HLA A-B-C-DRB1-DQB1 haplotypes, based on the number of characterized alleles, is 9.83×10^{13}. Taking into account only HLA alleles observed and confirmed by

high-resolution typing in the US population the theoretical number of HLA A-B-C-DRB1-DQB1 haplotypes is 3.13×10^9, while the current number of observed haplotypes was 7,987 as of 2007 (Maiers *et al*., 2007). The theoretical number of possible haplotypes is an overestimate since some of the combinations of HLA alleles are not possible due to linkage disequilibrium (Miretti *et al*., 2005). However, the actual observed number represents a gross underestimate – in our ongoing study (data not shown) of seven individuals only two of fourteen A-B-C-DRB1-DQB1 haplotypes were present within the list of 7987 observed haplotypes (Figure 1). These numbers indicate both the enormous combinatorial complexity of HLA/peptide interactions, as well as the discrepancy between the current information available from experiments and expected from theoretical analysis. Furthermore, the data available for the US population are not necessarily representative of the broad human population. Together, these data indicate a huge combinatorial space to be considered for vaccine development and clinical applications, such as transplant matching, drug development, and design of clinical trials.

While high-throughput methods for quantitative analysis of peptide binding to HLA molecules have been developed (*e.g.* Jiang and Boder, 2010; Harndahl *et al.*, 2011, this issue; Montero-Julian, 2011, this issue), they are limited to a small number of common HLA alleles. These methods can support mapping of hundreds to thousands of peptides making them suitable for T cell epitope mapping within complete proteomes of individual small viruses. A comprehensive mapping of T cell epitopes for a given human individual for a single small virus variant, such as influenza, requires the analysis of more than 100,000 peptide/HLA combinations. If we analyze a single virus variant for clinically relevant HLA alleles, as defined in (Maiers *et al*., 2007), there will be roughly 5 million peptide/HLA combinations. If we consider multiple variants of viruses, as well as the variation of HLA, the peptide space to be studied (search space) exceeds billions of possible assays. The gap between the experimental high-throughput methods and the size of the search space is addressed by the use of computational prediction methods for alleles and supertypes (Zhang *et al*., 2011a, this issue). Predictions of peptide/HLA binding have reached a very high level of accuracy for several HLA molecules, such as HLA-A*0201 (9 and 10-mers) (Zhang *et al*., 2011b, this issue; Lin *et al*., 2008a; Lin *et al*., 2008b). However, there is an urgent need for further improvement of *in silico* prediction tools in this area. At the moment we do not have reliable computational tools for prediction of peptide binding to the majority of known HLA molecules. For HLA class I binding peptides, the validated and accurate prediction models are lacking for all 8-mer and 11-mer and for majority of 9-mer and 10-mer peptides. For HLA class II binding peptides, the validated prediction models for HLA-DR are at best of marginal performance (Lin *et al.*, 2008b) and are completely lacking for HLA-DQ, and DP. Although significant progress has been made and *in silico* tools were shown to be as accurate as biochemical binding assays (Zhang *et al.*, 2011b), this only applies to a small set of HLA molecules. To enable full coverage of HLA targeted peptides we need further development and a systematic approach that will cover the diversity of HLA molecules and the variety of peptide lengths. Furthermore, recent studies have provided evidence that sets of naturally processed peptides identified by elution and mass spectrometry (Johnson *et al*., 2009; Fissolo *et al*., 2009; Wahl *et al*., 2010) do not fully correspond to peptides that are identified by biochemical binding assays. A significant proportion of the eluted peptides lack canonical binding motifs determined by binding assays.

The ability to precisely determine the properties of HLA-associated peptides is of critical importance for understanding the immune responses and for practical applications such as in transplantation (Ofran *et al*., 2010), vaccine and therapeutics development (Sette and Rappuoli, 2010; Cohen *et al*., 2010), or de-immunization of therapeutic proteins (De Groot *et al*., 2005). Combinatorial properties of the HLA, antigen processing and presentation constraints, diversity of experimental methods for their study, and the importance of HLA-

associated peptides for vaccine development make this field ideal for systematic application of ML methods rather than *ad hoc* solutions. However, systematic application of ML methods requires the common standards: unified nomenclature, quality control of the data, mapping of data from multiple sources to a common scale, and error and bias correction. ML repositories provide a link between application domains and the ML community, and are essential for both standardization and advanced tools development.

1.2 Machine learning repositories

General ML repositories provide a variety of data sets for the empirical analysis of various machine learning algorithms. Specialized ML repositories contain quality-controlled data sets from specific applications domains. The principal users of general ML repositories are computer scientists who develop, refine, and optimize ML algorithms, and analyze their behavior and properties under various conditions. These data sets allow performance comparison of an algorithm under different conditions and comparison of different algorithms across multiple data sets. Understanding the theoretical aspects of ML algorithms is important for the development of ML applications but alone it is not sufficient for the development of domain-specific applications. Such applications require precise tuning of ML algorithms to exclude domain-specific peculiarities of data types and models, but at the same time capture limitations, exceptions, and adequate level of complexity of the domain. An example of general ML repository is the UCI Machine Learning Repository (Frank and Asuncion, 2010). The UCI ML Repository is a collection of databases, domain theories, and data generators principally for use by the ML community for the development, analysis, and assessment of ML algorithms. It serves as a primary source of data sets for research, development and education in the ML field. Examples of data sets and repositories for ML are shown in Table 1.

Data repositories serve an important role – they provide benchmarking data sets for the development of ML applications including algorithms (search, classification, clustering, feature extraction, prediction, forecast), mathematical modeling, and quality assessment metrics. Properly designed ML repositories provide standardized data sets characterized by:

- Consistency – data represent high quality and reproducible measurements
- Comparability – data collected using different methodologies are mutually consistent and provided at similar scales of measurement
- Detailed description and availability of additional data – provided by availability of references and detailed description of data sets
- Large number of data points – sufficient number of observations are needed for statistical assessment of performance of ML algorithms.

2. DFRMLI repository

The lack of high quality data presents a major obstacle to the development of better computational solutions for immunological applications. A large number of experimentally verified HLA (and other MHC) binders and non-binders have been discovered and published in scientific literature and public databases such as SYFPEITHI (Rammensee *et al*., 1999) or IEDB (Kim *et al*., 2011, this issue; Vita *et al*., 2009). Several benchmark data sets are also available, including binding peptides to HLA class I (Peters *et al*., 2006; Lin et al., 2008a; Lundegaard *et al*., 2008) or to HLA class II (Lin *et al*., 2008b; Nielsen and Lund, 2009; Wang et al., 2010). While these databases and datasets are useful for the people working in immunoinformatics (Brusic and Petrovsky, 2003), they are still not in the format suitable for use by the mainstream ML community – those who are working in the development of advanced machine learning algorithms. To bridge the gap between

immunology and ML communities, we designed DFRMLI (Dana-Farber Repository for Machine Learning in Immunology). We have preprocessed and organized data to ensure consistency, comparability, detailed descriptions, and statistically meaningful sample sizes. The repository is accessible at http://bio.dfci.harvard.edu/DFRMLI/ and the front page is shown in Figure 2.

3. DFRMLI design

The repository is comprised of three parts. The first part “HLA Binding Peptides” provides information on peptide binding to HLA alleles, which could be used to develop computational models for HLA binding prediction. The second section “T cell epitopes” provides a list of experimentally validated T cell epitopes derived from tumor or virus antigens. It could be used to evaluate the performance of computational models in predicting T cell epitopes. The third section describes methods used for comparison and scaling of the data sets.

3.1 HLA binding peptides

This section has two parts: 1. Full records of the MHCPEP database (Brusic *et al*., 1997) and 2. Curated peptide binding information for selected common HLA variants including training and validation data sets. The MHCPEP database is now obsolete and has been replaced by the IEDB (Kim *et al*., 2011, this issue). The MHCPEP database contains 13,423 peptide sequences known to bind to MHC molecules and is included for historical reasons *i.e*. comparison of performance using historical data or lower quality data than currently available. Information of peptide sequences and MHC specificity is provided in the flat text format.

The detailed information of peptides binding to representative HLA alleles is listed in the Part 2 of HLA binding peptides section of DFRMLI. For each allele, the data are divided into training and validation sets. The recommended use is that training datasets is for development of computational models for HLA binding prediction and internal cross-validation, while the validation datasets should be used for model evaluation. The sizes of these data sets are given in Table 2.

The peptide binding information for eight HLA-I alleles (HLA-A*0101, A*0201, A*0301, A*1101, A*2402, B*0702, B*0801, B*1501) is included. Training dataset were collected from IEDB (Immune Epitope database) (Kim *et al*., 2011), CBS (Center for Biological Sequence analysis, Technical University of Denmark) data set (Nielsen *et al*., 2003), MULTIPRED data set (Zhang *et al*., 2005), or HotSpot Hunter data set (Zhang *et al*., 2008). In the training data sets for HLA-I alleles, the numbers of 9-mer peptides range from 99 (HLA-A*0301) to 3087 (HLA-A*0201) and those of 10-mer peptides range from 56 (HLA-A*0101) to 1316 (HLA-A*0201). Up to five validation datasets per HLA-I allele provide information of peptides that are not included in the training datasets. They were measured by iTopia™ Epitope Discovery System (Montero-Julian, 2011, this issue). The first dataset includes a full overlapping study of 134 9-mer peptides spanning the full length of tumor antigen survivin (Bachinsky *et al*., 2005). The second validation dataset includes 42 9-mer peptides spanning a 50 amino acids long construct containing cytomegalovirus (CMV) internal matrix protein pp65 9-mer peptides (Lin *et al*., 2008a). The third validation set includes 206 overlapping 9-mer peptides spanning tumor-associated antigen 5T4 (Shingler *et al*., 2008), as well as a 9-mer data set and a 10-mer data set used in the MLI competition (Crowe *et al*., 2011, this issue). The combined validation sets comprising all these datasets are provided for model validation use.

The peptide binding information of seven HLA-II alleles (HLA-DRB1*0101, HLA-DRB1*0301, HLA-DRB1*0401, HLA-DRB1*0701, HLA-DRB1*1101, HLA-DRB1*1301, and HLA-DRB1*1501) is also included. Training datasets for each allele were collected from the IEDB (Peters *et al*., 2005) and the benchmark data set (Wang et al., 2010). In the training data sets, the number of binding peptides is in the range of 30 (HLA-DRB1*1301) to 3882 (HLA-DRB1*0101). Peptides derived from four protein antigens including bee venom (30 18-mer peptides), LAGE-1 (17 16-mer to 19-mer peptides), lipocalin (25 16-mer peptides), and NEF (31 15-mer to 16-mer peptides), are provided as the validation set for MHC-II binding prediction (Texier *et al*., 2000; Mandic *et al*., 2003; Immonen *et al*., 2005; and Gahery *et al*., 2007). Their binding affinities to seven HLA-II molecules were measured as the concentration of peptides that prevented binding of 50% of the labeled probes.

3.2 T cell epitopes

The repository also provides several lists of T cell epitopes, which could be used to evaluate the performance of computational models to predict T cell epitopes. A total of 718 T cell epitopes derived from human tumor antigens are included. They collected from TANTIGEN (http://cvc.dfci.harvard.edu/tadb/) and are 8-31 amino acids in length and restricted by multiple HLA class I and class II alleles. Another 44 HLA-A2 restricted T cell epitopes derived from virus antigens are also provided. All of these T cell epitopes were found to be both presented by HLA A2 and recognized by T cells. Their capability to stimulate T cell was verified by various experiments. Information of antigen names, epitope sequences and references is provided. In addition, we provide the panel of 32 T cell epitopes, 8-12 amino acids in length, with sequences derived from the Cytomegalovirus, Epstein-Barr Virus and Influenza Virus (CEF). They are used as standard probes for T cell epitope studies (Currier *et al*., 2002; Nielsen *et al*., 2010).

3.3 Comparison and scaling

Immunological data reported in different studies are measured under different experimental conditions or with different reference peptides. Some assays measured peptide binding by measuring radioactive ligand (Sidney *et al*., 2001) or quantitative enzyme-linked immunosorbent assay (Sylvester-Hvid *et al*., 2002), whereas some may be represented by the percentage of binding affinity relative to control peptides (Bachinsky *et al*., 2005). The binding affinities in these studies were measured in different units. It is inappropriate to simply combine heterogeneous data without any transformation of raw binding affinities. To solve this problem, proper scaling and data transformation should be performed so that data generated from different sources could be integrated and compared. To enable inspection and comparison of predictions for different HLA alleles we scaled all the data to a common scale, e.g., from 0 to 100 using logarithm and linear transformation.

$$y_i = \begin{cases} x_i & if \quad x_i \geq 12 \quad and \quad x_i \leq 20000 \\ 12 & if \quad x_i < 12 \\ 20000 & if \quad x_i > 20000 \end{cases}$$

$$y_i^s = a + b \times \log(y_i), \quad where \quad a = \frac{\log(20000) \times 100}{\log(20000) - \log(12)} \quad b = \frac{-100}{\log(20000) - \log(12)} \tag{1}$$

$$y_i = \begin{cases} x_i & if \quad x_i \geq 10 \quad and \quad x_i \leq 100 \\ 10 & if \quad x_i < 10 \\ 100 & if \quad x_i > 100 \end{cases} \qquad y_i^s = a + b \times \ln(y_i), \quad where \quad a = -100 \quad b = \frac{100}{\ln(10)} = 43.4 \tag{2}$$

$$y_i = \begin{cases} x_i & if \quad x_i \neq 4 \\ 5 & if \quad x_i = 4 \end{cases} \qquad y_i^s = a + b \times y_i, \quad where \quad a = -12.5 \quad b = 12.5 \tag{3}$$

where y_i^S is the final scaled binding score ranging from 0 to 100, x_i is the original experimental measurements, and y_i is an intermediate value to map x_i *to* a slightly different scale. Equation (1), (2), and (3) are for mapping of measurements of binding affinities expressed as concentrations (in nM), measurements expressed as relative binding affinity (in %) to the labeled reference peptide, and measurements used by Multipred and Hotspot Hunter data sets, onto 0-100 scale. The transformation maps the scale onto intuitive values where 80-100 represent strong binding, 50-79 moderate to low binding, and 0-49 of no functional relevance.

For the assessment of classification accuracy, the commonly used measure is the area under the ROC curve (A_{ROC}) (Swets, 1988). It is convenient because it is calculated using all decision thresholds and all related decision values. The ROC curve is a plot of true positive rate TP/(TP+FN) on the vertical axis *vs.* false positive rate FP/(TN+FP) on the horizontal axis for the full range of the decision thresholds. The values AROC≥0.9 indicate excellent, 0.9>AROC≥0.8 good, 0.8>AROC≥0.7 marginal and 0.7>AROC poor predictions (Swets, 1988).

To assess the accuracy of binding affinity prediction the Pearson correlation coefficient for experimental measurements X and a prediction series Y for the studied set of peptides can be used:

$$r_{xy} = \frac{\Sigma\left(x_i - \bar{x}\right) \times \left(y_i - \bar{y}\right)}{\sqrt{\Sigma\left(x_i - \bar{x}\right)^2 \times \Sigma\left(y_i - \bar{y}\right)^2}} \tag{4}$$

where x_i and $\bar{x}$ are experimental individual and average affinities; y_i and $\bar{y}$ are individual and average peptide predictions. The range of correlation coefficient is within −1 to 1, with 1 representing a perfect positive linear relationship, −1 representing a perfect negative linear relationship, and 0 representing total lack of correlation.

4. Conclusion

Combining experimental and *in silico* methods enables systematic study of highly combinatorial problems associated with deciphering immune responses. With the advancement of experimental technologies, the amount of immunological data produced and distributed across literature and databases keep increasing exponentially. To fully utilize these data, advanced computational methods including statistical and ML algorithms must be used for development of improved bioinformatics tools. However, immunological data, like other biological data, are usually described qualitatively and these descriptions are often

ambiguous, presenting a challenge for the mainstream ML developers. The DFRMLI is designed to bridge this gap through providing a resource of well-defined and annotated immunological data that could be conveniently used by the ML community. The data was pre-processed and carefully categorized so that they can be directly used by ML practitioners. Scaling and comparison transformations were also applied to integrate data from different sources. DFRMLI has already been used in a number of reported studies (Lin *et al*., 2008a; Lin *et al*., 2008b; Singh and Mishra, 2008; Bordner and Mittelman, 2010). We plan to expand DFRMLI with other annotated data from the immunological domain.

Acknowledgments

This work was supported by NIH grants U19AI57330 and 470 U01AI90043 and a grant from DOD W81XWH-07-1-0080. We are thankful to Dr Songsak Tongchusak for providing a list of T cell epitopes that was included in DFRMLI and Tara C. Mayo for thoughtful review of the manuscript.

References

Bachinsky MM, Guillen DE, Patel SR, Singleton J, Chen C, Soltis DA, Tussey LG. Mapping and binding analysis of peptides derived from the tumor-associated antigen survivin for eight HLA alleles. Cancer Immun. 2005; 22(5):6. [PubMed: 15779886]

Baldi, P.; Brunak, S. Bioinformatics: the machine learning approach. MIT Press; 2001.

Bambini S, Rappuoli R. The use of genomics in microbial vaccine development. Drug Discov Today. 2009; 14(5-6):252. [PubMed: 19150507]

Bordner AJ, Mittelmann HD. MultiRTA: a simple yet reliable method for predicting peptide binding affinities for multiple class II MHC allotypes. BMC Bioinformatics. 2010; 11:482. [PubMed: 20868497]

Brusic V, August JT. The changing field of vaccine development in the genomics era. Pharmacogenomics. 2004; 5(6):597–600. [PubMed: 15335280]

Brusic V, Petrovsky N. Immunoinformatics – the new kid in town. NovartisFound. Symp. 2003; 254:3.

Brusic V, Rudy G, Harrison LC. MHCPEP, a database of MHC-binding peptides: update 1997. Nucleic Acids Res. 1997; 26(1):368–371. [PubMed: 9399876]

Cohen T, Moise L, Ardito M, Martin W, De Groot AS. A method for individualizing the prediction of immunogenicity of protein vaccines and biologic therapeutics: individualized T cell epitope measure (iTEM). J. Biomed. Biotechnol. 2010:961752. pii.

Crowe, et al. In preparation for this issue. J. Immunol. Methods. 2011 this issue.

Currier JR, Kuta EG, Turk E, Earhart LB, Loomis-Price L, Janetzki S, Ferrari G, Birx DL, Cox JH. A panel of MHC class I restricted viral peptides for use as a quality control for vaccine trial ELISPOT assays. J Immunol Methods. 2002; 260(1-2):157–72. [PubMed: 11792386]

David MP, Concepcion GP, Padlan EA. Using simple artificial intelligence methods for predicting amyloidogenesis in antibodies. BMC Bioinformatics. 2010; 11:79. [PubMed: 20144194]

De Groot AS, Knopp PM, Martin W. De-immunization of therapeutic proteins by T cell epitope modification. Dev Biol (Basel). 2005; 122:171–94. [PubMed: 16375261]

Fissolo N, Haag S, de Graaf KL, Drews O, Stevanovic S, Rammensee HG, Weissert R. Naturally presented peptides on major histocompatibility complex I and II molecules eluted from central nervous system of multiple sclerosis patients. Mol. Cell. Proteomics. 2009; 8(9):2090–101. [PubMed: 19531498]

Frank, A.; Asuncion, A. UCI Machine Learning Repository. University of California, School of Information and Computer Science; Irvine, CA: 2010. http://archive.ics.uci.edu/ml

Gahery H, Figueiredo S, Texier C, Pouvelle-Moratille S, Ourth L, Igea C, Surenaud M, Guillet JG, Maillere B. HLA-DR-restricted peptides identified in the Nef protein can induce HIV type 1-specific IL-2/IFN-gamma-secreting CD4+ and CD4+ /CD8+ T cells in humans after lipopeptide vaccination. AIDS Res. Hum. Retroviruses. 2007; 23(3):427–37. [PubMed: 17411376]

Godkin AJ, Smith KJ, Willis A, Tejada-Simon MV, Zhang J, Elliott T, Hill AV. aturally processed HLA class II peptides reveal highly conserved immunogenic flanking region sequence preferences

that reflect antigen processing rather than peptide-MHC interactions. J. Immunol. 2001; 166(11): 6720–7. [PubMed: 11359828]

Harndahl M, Rasmussen M, Roder G, Buus S. Real-time, high-throughput measurements of peptide-MHC-I dissociation using a scintillation proximity assay. J. Immunol. Methods. 2011 in press, this issue.

Hertz T, Yanover C. Identifying HLA supertypes by learning distance functions. Bioinformatics. 2007; 23(2):e148. [PubMed: 17237084]

Immonen A, Farci S, Taivainen A, Partanen J, Pouvelle-Moratille S, Narvanen A, Kinnunen T, Saarelainen S, Rytkonen-Nissinen M, Maillere B, Virtanen T. T cell epitope-containing peptides of the major dog allergen Can f 1 as candidates for allergen immunotherapy. J. Immunol. 2005; 175(6):3614–20. [PubMed: 16148105]

Jiang W, Boder ET. High-throughput engineering and analysis of peptide binding to class II MHC. Proc. Natl. Acad. Sci. USA. 2010; 107(30):13258–63. [PubMed: 20622157]

Johnson KL, Ovsyannikova IG, Mason CJ, Bergen HR 3rd, Poland GA. Discovery of naturally processed and HLA-presented class I peptides from vaccinia virus infection using mass spectrometry for vaccine development. Vaccine. 2009; 28(1):38–47. [PubMed: 19822231]

Kim Y, Sette A, Peters B. Applications for T-cell epitope queries and tools in the Immune Epitope Database and Analysis Resource. J Immunol Methods. 2011 in press, this issue.

Lafuente EM, Reche PA. Prediction of MHC-peptide binding: a systematic and comprehensive overview. Curr Pharm Des. 2009; 15(28):3209. [PubMed: 19860671]

Lin HH, Ray S, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. BMC Immunol. 2008a; 9:8. [PubMed: 18366636]

Lin HH, Zhang GL, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. BMC Bioinformatics. 2008b; 9(Suppl 12):S22. [PubMed: 19091022]

Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res. 2008; 1(36):W509–12. Web Server issue. [PubMed: 18463140]

Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. Hum. Immunol. 2007; 68(9):779. [PubMed: 17869653]

Mandic M, Almunia C, Vicel S, Gillet D, Janjic B, Coval K, Maillere B, Kirkwood JM, Zarour HM. The alternative open reading frame of LAGE-1 gives rise to multiple promiscuous HLA-DR-restricted epitopes recognized by T-helper 1-type tumor-reactive CD4+ T cells. Cancer Res. 2003; 63(19):6506–15. [PubMed: 14559844]

Meuer SC, Schlossman SF, Reinherz EL. Clonal analysis of human cytotoxic T lymphocytes: T4+ and T8+ effector T cells recognize products of different major histocompatibility complex regions. Proc. Natl Acad. Sci. USA. 1982; 79:4395–9. [PubMed: 6981813]

Miretti MM, Walsh EC, Ke X, Delgado M, Griffiths M, Hunt S, Morrison J, Whittaker P, Lander ES, Cardon LR, Bentley DR, Rioux JD, Beck S, Deloukas P. A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. Am. J. Hum. Genet. 2005; 76:634. [PubMed: 15747258]

Montero-Julian FA. iTOPIA Epitope Discovery System a new technology for the identification of MHC Class I epitopes. J Immunol Methods. 2011 in press, this issue.

Muh HC, Tong JC, Tammi MT. AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins. PLoS One. 2009; 4(6):e5861. [PubMed: 19516900]

Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics. 2009; 10:296. [PubMed: 19765293]

Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, Brunak S, Lund O. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. 2003; 12(5):1007–17. [PubMed: 12717023]

Nielsen JS, Wick DA, Tran E, Nelson BH, Webb JR. An in vitro-transcribed-mRNA polyepitope construct encoding 32 distinct HLA class I-restricted epitopes from CMV, EBV, and Influenza for

use as a functional control in human immune monitoring studies. J. Immunol. Methods. 2010; 360(1-2):149–56. [PubMed: 20637775]

Ofran Y, Kim HT, Brusic V, Blake L, Mandrell M, Wu CJ, Sarantopoulos S, Bellucci R, Keskin DB, Soiffer RJ, Antin JH, Ritz J. Diverse patterns of T-cell response against multiple newly identified human Y chromosome-encoded minor histocompatibility epitopes. Clin Cancer Res. 2010; 16(5): 1642–51. 2010. [PubMed: 20160060]

Palladini A, Nicoletti G, Pappalardo F, Murgo A, Grosso V, Stivani V, Ianzano ML, Antognoli A, Croci S, Landuzzi L, De Giovanni C, Nanni P, Motta S, Lollini PL. In silico modeling and in vivo efficacy of cancer-preventive vaccinations. Cancer Res. 2010; 70(20):7755–63. [PubMed: 20924100]

Pappalardo F, Halling-Brown MD, Rapin N, Zhang P, Alemani D, Emerson A, Paci P, Duroux P, Pennisi M, Palladini A, Miotto O, Churchill D, Rossi E, Shepherd AJ, Moss DS, Castiglione F, Bernaschi M, Lefranc MP, Brunak S, Motta S, Lollini PL, Basford KE, Brusic V. ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design and optimization. Brief. Bioinform. 2009; 10:330. [PubMed: 19383844]

Peters B, Bui HH, Frankild S, Nielson M, Lundegaard C, Kostem E, Basch D, Lamberth K, Harndahl M, Fleri W, Wilson SS, Sidney J, Lund O, Buus S, Sette A. A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS Comput Biol. 2006; 2(6):e65. [PubMed: 16789818]

Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko JV, Sathiamurthy M, Schoenberger S, Stewart S, Surko P, Way S, Wilson S, Sette A. The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol. 2005; 3(3):e91. [PubMed: 15760272]

Petrovsky N, Brusic V. Computational immunology: The coming of age. Immunol Cell Biol. 2002; 80(3):248–54. [PubMed: 12067412]

Petrovsky N, Brusic V. Virtual models of the HLA class I antigen processing pathway. Methods. 2004; 34(4):429. [PubMed: 15542368]

Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. 1999; 50(3-4):213–9. [PubMed: 10602881]

Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, Marsh SGE. The IMGT/HLA Database. Nucleic Acids Res. 2011; 39(Suppl 1):D1171–6. [PubMed: 21071412]

Sette A, Rappuoli R. Reverse vaccinology: developing vaccines in the era of genomics. Immunity. 2010; 33(4):530–41. [PubMed: 21029963]

Shingler WH, Chikoti P, Kingsman SM, Harrop R. Identification and functional validation of MHC class I epitopes in the tumor-associated antigen 5T4. Int Immunol. 2008; 20(8):1057–66. [PubMed: 18567615]

Shiina T, Inoko H, Kulski JK. An update of the HLA genomic region, locus information and disease associations. Tissue Antigens. 2004; 64(6):631. [PubMed: 15546336]

Sidney J, Southwood S, Oseroff C, del Guercio MF, Sette A, Grey HM. Measurement of MHC/peptide interactions by gel filtration. Curr Protoc Immunol. 2001 Chapter 18:Unit 18.3.

Singh SP, Mishra BN. Prediction of MHC binding peptide using Gibbs motif sampler, weight matrix and artificial neural network. Bioinformation. 2008; 3(4):150–5. [PubMed: 19238237]

Stern LJ, Brown JH, Jardetzky TS, Gorga JC, Urban RG, Strominger JL, Wiley DC. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. Nature. 1994; 368:215. [PubMed: 8145819]

Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988; 240(4857):1285–93. [PubMed: 3287615]

Sylvester-Hvid C, Kristensen N, Blicher T, Ferre H, Lauemoller SL, Wolf XA, Lamberth K, Nissen MH, Pedersen LO, Buus S. Establishment of a quantitative ELISA capable of determining peptide - MHC class I interaction. Tissue Antigens. 2002; 59(4):251–8. [PubMed: 12135423]

Texier C, Pouvelle S, Busson M, Herve M, Charron D, Menez A, Maillere B. HLA-DR restricted peptide candidates for bee venom immunotherapy. J. Immunol. 2000; 164(6):3177–3184. [PubMed: 10706708]

Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B. The immune epitope database 2.0. Nucleic Acids Res. 2009; 38:D854–62. Database issue. [PubMed: 19906713]

Wahl A, Schafer F, Bardet W, Hildebrand WH. HLA class I molecules reflect an altered host proteome after influenza virus infection. Hum Immunol. 2010; 71(1):14–22. [PubMed: 19748539]

Wang P, Sidney J, Kim Y, Sette A, Lund O, Nielsen M, Peters B. Peptide binding predictions for HLA DR, DP and DQ molecules. BMC Bioinformatics. 2010; 11:568. [PubMed: 21092157]

Zhang GL, Deluca DS, Keskin DB, Chitkushev L, Zlateva T, Lund O, Reinherz EL, Brusic V. MULTIPRED2: A computational system for large-scale identification of peptides predicted to bind to HLA supertypes and alleles. J Immunol. Methods. 2011a

Zhang GL, Khan AM, Srinivasan KN, Heiny A, Lee K, Kwoh CK, August JT, Brusic V. Hotspot Hunter: a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes. BMC Bioinformatics. 2008; 9(Suppl 1):S19. 2008. [PubMed: 18315850]

Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V. MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. Nucleic Acids Res. 2005; 33:W172–9. Web Server issue. [PubMed: 15980449]

Zhang GL, et al. Machine learning in immunology competition: prediction of HLA binding peptides. J Immunol Methods. 2011b in press, this issue.

Zhang, YQ.; Rajapakse, JC. Machine learning in bioinformatics. Wiley; 2009.

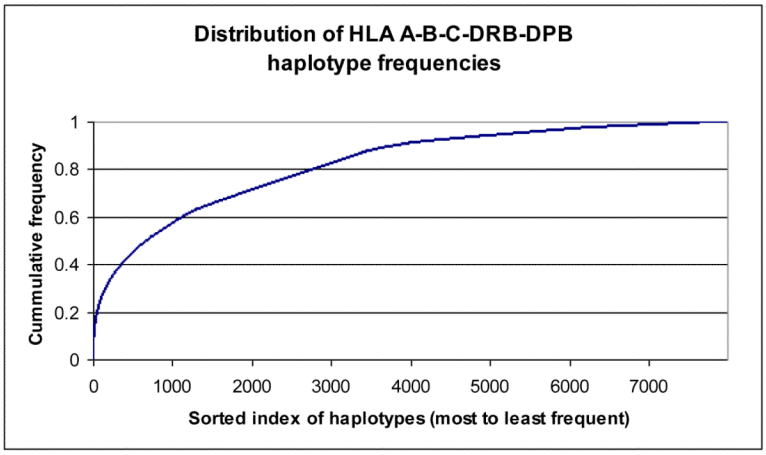

Figure 1.

HLA haplotype frequencies in the North American population. The 1st percentile of most common HLA haplotypes are present in 23.3% of population, the 10th percentile are present in 52.8% population, and 50th percentile are present in 91% of population. The two most common North American HLA haplotypes A*0101/B*0801/C*0701/DRB1*0301/DQB1*0201 and A*0301/B*0702/C*0702/DRB1*1501/DQB1*0602 are present in 2.7% and 1.4% of the population respectively.

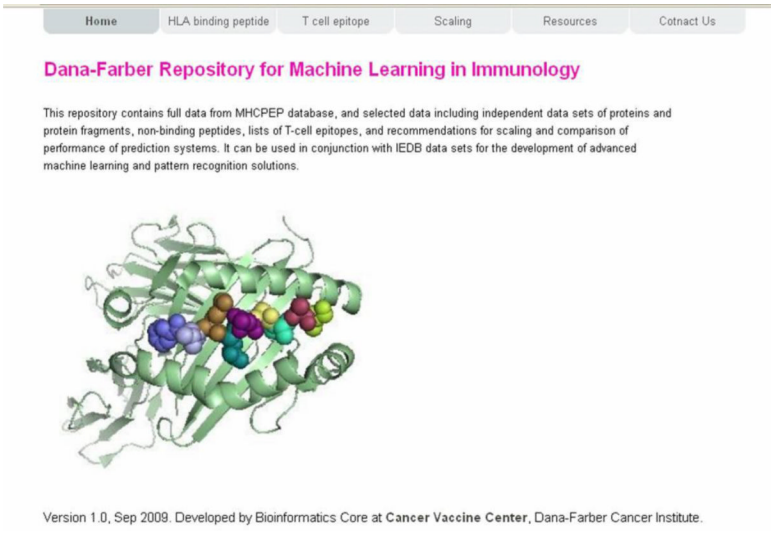

Figure 2.
The web page of DFRMLI

Table 1

A representative set of publically available data repositories.

Domain	Description	URL (http://)
General	Databases, domain theories and data generators for ML	mlearn.ics.uci.edu/MLRepository.html
General	Time series data sets	robjhyndman.com/TSDL
General	Datasets for statistics	lib.stat.cmu.edu/datasets
General	Datasets for statistics and ML	trec.nist.gov/data.html
General	Datasets for evaluating ML systems	www.cs.utoronto.ca/~delve
General	Annual KDD competition	www.sigkdd.org/kddcup/index.php
Biology	Human genomic splice sites dataset	www.sci.unisannio.it/docenti/rampone
Biology	Microarray data sets	smd.stanford.edu
Biology	Gene expression ML repository	gemler.fzv.uni-mb.si
Linguistics	Annotated linguistic structures in naturally occurring texts	www.cis.upenn.edu/~treebank
Pattern recognition	Database of handwritten digits	yann.lecun.com/exdb/mnist
Pattern recognition	Face recognition data	www.cs.cmu.edu/afs/cs.cmu.edu/user/avrim/www/ML94/face_homework.html
Space Science	Astrophysical data	nssdc.gsfc.nasa.gov
World Wide Web	knowledge base mirroring the content of the WWW	www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/

Table 2

Description of MHC binding datasets in the DFRMLI repository, as of March 2011.

Datasets	T1	T2	T3	T4	V1	V2	V3	V4	V5	CV
A*0101	1157	56	447	-	135	42	207	265	177	649
A*0201	3087	1316	444	2646	136	42	207	265	177	649
A*0301	2092	1082	331	99	136	42	207	-	-	384
A*1101	1983	1093	219	263	136	42	-	-	-	178
A*2402	195	78	367	-	136	42	-	-	-	178
B*0702	1261	205	232	1574	136	42	207	87	61	472
B*0801	708	-	119	-	136	42	-	-	-	178
B*1501	976	-	114	-	136	42	-	-	-	178
DRB1*0101	3882	-	-	-	103	-	-	-	-	-
DRB1*0301	502	-	-	-	103	-	-	-	-	-
DRB1*0401	512	-	-	-	103	-	-	-	-	-
DRB1*0701	505	-	-	-	103	-	-	-	-	-
DRB1*1101	520	-	-	-	103	-	-	-	-	-
DRB1*1301	30	-	-	-	103	-	-	-	-	-
DRB1*1501	520	-	-	-	103	-	-	-	-	-

T9-1 (Peters *et al.*, 2006), T9-2 (Nielsen *et al*., 2003), and T9-3 (Zhang *et al*., 2005) stand for training sets for 9-mer peptides, V9 (Bachinsky *et al*., 2005; Shingler *et al*., 2008; Lin *et al.*, 2008a; Crowe *et al*., 2011, this issue) for complete validation set for 9-mer peptides, and C9 for validation set for 9-mers used in the MLI competition (Crowe *et al*., 2011, this issue). T10 (Peters *et al*., 2006) stands for training set for 10-mers while, V10 for complete validation set for 10-mers, while C10 stands for validation set for 10-mers used in the MLI competition (Crowe *et al*., 2011, this issue). Abbreviation: T1: IEDB 9-mer training data set; T2: IEDB 10-mer training data set; T3: CBS 9-mer training data set; V1: validation data set 1; CV: combined validation data set.