BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 8 2007, pages 942–949
doi:10.1093/bioinformatics/btm061

Sequence analysis

POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties

Chun-Wei Tung[1] and Shinn-Ying Ho[1,2,*]

[1]Institute of Bioinformatics and [2]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan

Received on October 28, 2006; revised and accepted on February 14, 2007

Advance Access publication March 24, 2007

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Both modeling of antigen-processing pathway including major histocompatibility complex (MHC) binding and immunogenicity prediction of those MHC-binding peptides are essential to develop a computer-aided system of peptide-based vaccine design that is one goal of immunoinformatics. Numerous studies have dealt with modeling the immunogenic pathway but not the intractable problem of immunogenicity prediction due to complex effects of many intrinsic and extrinsic factors. Moderate affinity of the MHC–peptide complex is essential to induce immune responses, but the relationship between the affinity and peptide immunogenicity is too weak to use for predicting immunogenicity. This study focuses on mining informative physicochemical properties from known experimental immunogenicity data to understand immune responses and predict immunogenicity of MHC-binding peptides accurately.

Results: This study proposes a computational method to mine a feature set of informative physicochemical properties from MHC class I binding peptides to design a support vector machine (SVM) based system (named POPI) for the prediction of peptide immunogenicity. High performance of POPI arises mainly from an inheritable bi-objective genetic algorithm, which aims to automatically determine the best number *m* out of 531 physicochemical properties, identify these *m* properties and tune SVM parameters simultaneously. The dataset consisting of 428 human MHC class I binding peptides belonging to four classes of immunogenicity was established from MHCPEP, a database of MHC-binding peptides (Brusic *et al*., 1998). POPI, utilizing the $m = 23$ selected properties, performs well with the accuracy of 64.72% using leave-one-out cross-validation, compared with two sequence alignment-based prediction methods ALIGN (54.91%) and PSI-BLAST (53.23%). POPI is the first computational system for prediction of peptide immunogenicity based on physicochemical properties.

Availability: A web server for prediction of peptide immunogenicity (POPI) and the used dataset of MHC class I binding peptides (PEPMHCI) are available at http://iclab.life.nctu.edu.tw/POPI

Contact: syho@mail.nctu.edu.tw

1 INTRODUCTION

Developing a computer-aided system to design peptide vaccines is one goal of immunoinformatics. The major work of previous studies for peptide vaccine designs is to identify cytotoxic T lymphocyte (CTL) epitopes and investigate their corresponding immunogenicity. The CTL cells play a critical role in protective immunity by recognizing and eliminating self-altered cells, which recognize short peptides derived from intracellular degradation of foreign proteins in combination with major histocompatibility complex (MHC) class I molecules (Hämmerling *et al*., 1999). The immunogenicity of MHC class I binding peptides is their ability to induce CTL responses. Accurate predictions of the CTL epitopes and their corresponding immunogenicity are critical in developing a computer-aided system for vaccine designs.

Direct approach to predicting the CTL epitopes has been studied initially but its accuracy is fairly low (Deavin *et al*., 1996). Instead, indirect approach to predicting the MHC-binding peptides is useful because peptides must be processed prior to inducing cellular immune responses. The recent studies of bioinformatics utilized the information about antigen-processing pathway to predict the CTL epitopes. At first, the peptides are cleaved by proteasomal cleavage. Several studies elucidating the specificity of proteasome have been presented. To predict proteasomal cleavage sites, NetChop used a neural network method (Keşmir *et al*., 2002) and Pcleavage is based on a support vector machine (SVM) learning model (Bhasin and Raghava, 2005).

After cleavage, peptide fragments are transported into endoplasmic reticulum by TAP, which is the transporter associated with antigen processing. Some studies of investigating the TAP transport efficiency were presented, such as the affinity prediction of TAP-binding peptides using the cascade SVM (Bhasin and Raghava, 2004) and the prediction of TAP transport efficiency of epitope precursors using a simple scoring matrix (Peters *et al*., 2003). Finally, the peptide fragments that bound to MHC class I molecules are subsequently translocated to the cell surface, where these complexes may active CTL. Some methods have been developed to predict MHC class I binding affinity, such as the SVM-based SVMHC (Dönnes and Elofsson, 2002) and Gibbs sampling method (Nielsen *et al*., 2004). Moreover, the hybrid approaches integrated the above-mentioned methods like the prediction of proteasomal

*To whom correspondence should be addressed.

cleavage, TAP transport efficiency and MHC binding to advance the prediction performance (Dönnes and Kohlbacher, 2005; Larsen *et al.*, 2005).

After the prediction of CTL epitopes, defining peptide immunogenicity is desirable to accurately predict immunogenicity of epitopes for the vaccine design. The peptide immunogenicity is influenced by many factors, including intrinsic physicochemical properties and extrinsic factors such as host immunoglobulin repertoire (Kanduc, 2005; Van Regenmortel, 2001). Several studies aimed to clarify the relationship between the peptide-binding affinity to the MHC molecule and its immunogenicity (Feltkamp *et al.*, 1994; Ochoa-Garay *et al.*, 1997). These studies revealed that moderate binding affinity of peptide-MHC molecules is essential to induce immune responses, but the ability of peptides to induce CTL responses does not strongly correlate with their affinity for the MHC molecule.

Physicochemical properties of amino acids were extensively and successfully used in sequence-based prediction methods (Blythe and Flower, 2005; Cao *et al.*, 2006; Idicula-Thomas *et al.*, 2006; Liu *et al.*, 2006; Nanni and Lumini, 2006). Because of the weak correlation between peptide immunogenicity and peptide-MHC binding affinity, mining informative physicochemical properties is a potentially good approach to designing a classifier for predicting immunogenicity. Because the number of available physicochemical properties is as large as more than 500, the properties used in previous studies are usually selected according to domain knowledge (Liu *et al.*, 2006) or the rank-based method (Sarda *et al.*, 2005). Therefore, these methods cannot be effectively applied to the investigated intractable problems because of limited knowledge or neglect of correlated effects among multiple properties (Blythe and Flower, 2005). This study aims to design an accurate predictor by efficiently selecting a small set of informative physicochemical properties considering the correlated effects.

It is well recognized that feature selection and classifier design should be optimized simultaneously to maximize prediction accuracy (Ho *et al.*, 2006). The SVM-based learning methods are shown effective for various prediction methods from protein sequences (Bhasin and Raghava, 2005; Dönnes and Elofsson, 2002). However, internal detection of relevant-feature correlation is not offered by conventional SVMs; meanwhile, appropriate setting of their control parameters is often treated as another independent problem (Chang and Lin, 2001). Let there be n candidates of physicochemical properties of amino acids. To maximize accuracy of the investigated prediction problem by selecting a small number m out of n properties while cooperating with SVM simultaneously, it is equivalent to solve the binary combinatorial optimization problem having a huge search space of $C(n,m) = n!/(m!(n-m)!)$.

This study proposes an efficient method to mine a feature set of informative physicochemical properties from MHC class I binding peptides to design an SVM-based system (named POPI) for prediction of peptide immunogenicity. High performance of POPI arises mainly from an inheritable bi-objective genetic algorithm (Ho *et al.*, 2004a), which aims to automatically determine the best number m out of $n = 531$ physicochemical properties, identify these m properties and tune SVM parameters simultaneously by maximizing the prediction accuracy of 10-fold cross-validation (10-CV). In this study, the used dataset consisting of 428 human MHC class I binding peptides belonging to four classes of immunogenicity was established from MHCPEP, a database of MHC-binding peptides (Brusic *et al.*, 1998). POPI, utilizing the $m = 23$ selected properties, performs well with accuracy of 64.72% using leave-one-out cross-validation, compared with two sequence alignment-based prediction methods ALIGN (54.91%) and PSI-BLAST (53.23%).

In contrast to the existing affinity-based methods of predicting immunogenicity by way of predicting MHC-binding peptides, POPI is the first computational system based on physicochemical properties to predict peptide immunogenicity using epitopes associated with human MHC class I molecules, which has been implemented as a web server (http://iclab.life.nctu.edu.tw/POPI).

2 METHODS

2.1 Dataset and physicochemical properties

Table 1 shows the used dataset PEPMHCI of peptides associated with human MHC class I molecules extracted from MHCPEP. The keywords used to construct the dataset are ‘HLA’ and ‘CLASS-1’ in the ‘MHCMolecule’ field. The immunogenicity of a peptide is determined by measuring the concentration of peptides giving 50% of maximum specific lysis by CTLs of target cells displaying the peptide, and is given a descriptive value. The initial numbers of peptides extracted belonging to the six classes, None, Little, Moderate, High, Immunogenic-not-quantified and Unknown, are 147, 95, 125, 132, 867 and 3251, respectively. The peptides of the classes Immunogenic-not-quantified and Unknown were not considered. After removing 19 duplicate records and 52 inconsistent records, PEPMHCI with no artificial peptide contains 428 peptides, as shown in Table 1. The shortest, averaged and longest lengths of the 428 peptides are 7, 10.26 and 25, respectively.

There are 544 physicochemical properties of amino acids extracted from amino acid index database version 9.0 (AAindex), which is a collection of published amino acid indices representing different physicochemical and biological properties of amino acids (Kawashima and Kanehisa, 2000). Each physicochemical property consists of a set of 20 numerical values for amino acids. The property having the value ‘NA’ in a value set of amino acid index was discarded. Finally, 531 properties were used for the following mining method.

Table 1. The dataset PEPMHCI of peptides associated with human MHC class I molecules extracted from MHCPEP, a database of MHC-binding peptides (Brusic *et al.*, 1998)

Immunogenicity class	Number of peptides
None	144
Little	83
Moderate	100
High	101
Total	428

2.2 Support vector machine

Support vector machine (SVM) is a learning model dealing with binary classification problems. SVM constructs a binary classifier by finding a hyperplane to separate two classes with a maximal distance between margins of two classes consisting of support vectors. In order to make linear separation of samples easier, SVM uses one of various kernel functions to transform the samples into a high-dimensional search space. In this work, the commonly used radial basis function is applied to non linearly transform the feature space, defined as follows:

$$K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|\right), \quad \gamma > 0. \tag{1}$$

The kernel parameter γ determines how the samples are transformed into a high-dimensional search space. The cost parameter $C > 0$ of SVM adjusts the penalty of total error. These two parameters C and γ must be tuned to get the best prediction performance.

For multi-class classification problems, 'one-against-one' strategy is applied to transform the multi-class problem into several binary classification problems. Given h classes, there are $h(h-1)/2$ classifiers constructed and each one trains the samples from two classes. A voting strategy is applied to give a final prediction for test samples. In this study, $h = 4$ and the used SVM is obtained from LIBSVM package version 2.81 (Chang and Lin, 2001).

2.3 Orthogonal experimental design

Statistic design of experiments is a process of planning experiments. Orthogonal experimental design with orthogonal array and factor analysis is an efficient method to analyze the effect of several factors simultaneously (Dey, 1985; Wu, 1978). The factors are the parameters, which affect response variables, and a discriminative value of a factor is regarded as a level of the factor. A 'complete factorial' experiment would make measurements at each of all possible level combinations. However, the number of level combinations is often so large that this is impractical, and a subset of level combinations must be judiciously selected to be used, resulting in a 'fractional factorial' experiment. Orthogonal experimental design utilizes properties of fractional factorial experiments to efficiently determine the best combination of factor levels to use in design problems.

Orthogonal array is a fractional factorial array, which assures a balanced comparison of levels of any factor. Orthogonal array can reduce the number of level combinations for factor analysis. Each row of an orthogonal array represents the levels of factors in each combination, and each column represents a specific factor that can be changed from each combination. The term 'main effect' of one factor designates the effect on response variables that one can trace to a design parameter, which does not bother the estimation of the main effect of another factor. After proper tabulation of experimental results, the summarized data are analyzed using factor analysis to determine the relative-level effects of factors.

Factor analysis can evaluate the effects of individual factors on the evaluation function, rank the most effective factors, and determine the best level for each factor such that the evaluation function is optimized. Table 2 shows an illustrative example of orthogonal experimental design using a two-level orthogonal array $L_M(2^{M-1})$ with M rows and $M-1$ columns. In this example of $M = 8$, there are seven factors where each corresponds to a physicochemical property and its two levels correspond to exclusion and inclusion of the feature in the proposed feature selection. Let f_t denote a function value (prediction accuracy of 10-CV in this study) of the combination t. Define the main effect of factor j with level k as S_{jk} where $j = 1, \ldots, M-1$ and $k = 1, 2$:

$$S_{jk} = \sum f_t \cdot F_t, \quad t = 1, \ldots, M, \tag{2}$$

where $F_t = 1$ if the level of factor j of combination t is k; otherwise, $F_t = 0$. Since the objective function is to be maximized, the level 1 of factor j makes a better contribution to the function than level 2 of factor j does when $S_{j1} > S_{j2}$. The main effect reveals the individual effect of a factor. After the better one of two levels of each factor is determined, a good combination consisting of all factors with the better levels can be easily reasoned (Ho *et al.*, 2004b).

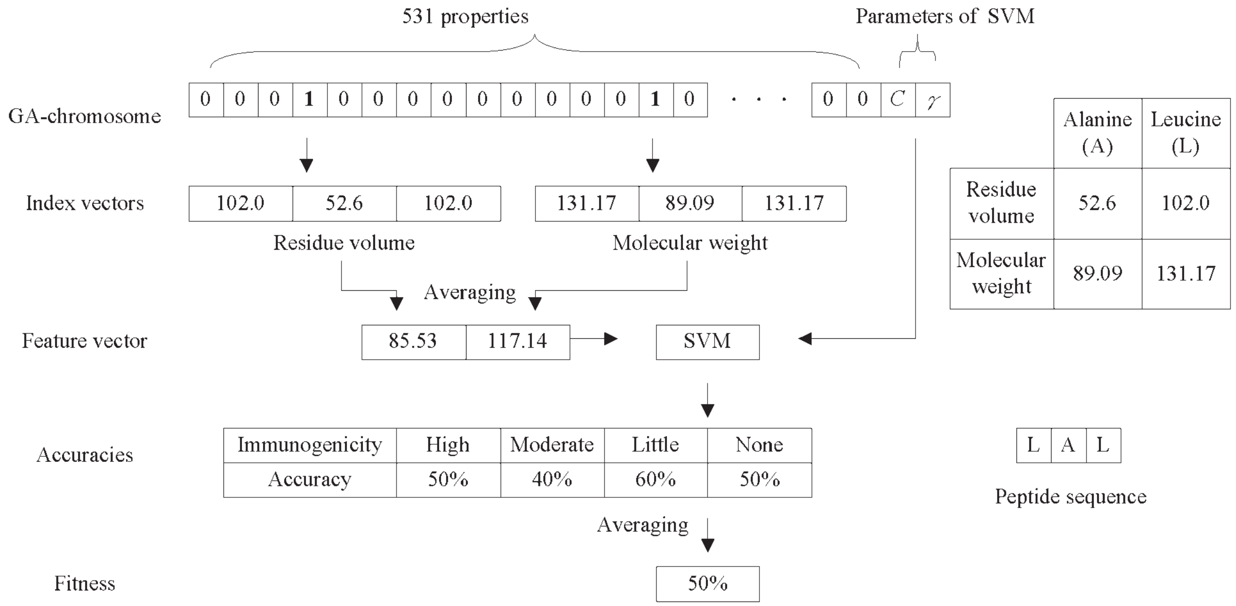
The rank in Table 2 shows the rank of the combination t in all 128 ($=2^7$) possible combinations. In this example, the reasoned combination gets the best accuracy with rank 1. Notably, the reasoned combination is not guaranteed to be the best one in general cases. The most effective factor j has the largest main effect difference MED $= |S_{j1} - S_{j2}|$. The 6th factor having the largest MED 36.3 is the most effective factor.

2.4 Inheritable bi-objective genetic algorithm

Selecting a minimal number of informative features while maximizing prediction accuracy is a bi-objective 0/1 combinatorial optimization problem. An efficient inheritable bi-objective genetic algorithm

Table 2. An illustration example of orthogonal array $L_8(2^7)$ and factor analysis

t	Factors							Accuracy (%) f_t	Rank
	1	2	3	4	5	6	7		
1	1	1	1	1	1	1	1	28.8	33/128
2	1	1	1	2	2	2	2	18.8	97/128
3	1	2	2	1	1	2	2	28.8	33/128
4	1	2	2	2	2	1	1	17.5	100/128
5	2	1	2	1	2	1	2	20.0	88/128
6	2	1	2	2	1	2	1	41.3	4/128
7	2	2	1	1	2	2	1	33.8	14/128
8	2	2	1	2	1	1	2	20.0	88/128
S_{j1}	93.8	108.8	101.3	111.3	118.8	86.3	121.3		
S_{j2}	115.0	100.0	107.5	97.5	90.0	122.5	87.5		
MED	21.3	8.8	6.3	13.8	28.8	36.3	33.8		
Rank	4	6	7	5	3	1	2		
Better level	2	1	2	1	1	2	1	42.5	1/128



	Alanine (A)	Leucine (L)
Residue volume	52.6	102.0
Molecular weight	89.09	131.17

Immunogenicity	High	Moderate	Little	None
Accuracy	50%	40%	60%	50%

Fig. 1. An illustration example of fitness function evaluation from decoding a GA-chromosome.

(IBCGA, Ho *et al.*, 2004a) is utilized to solve this optimization problem. IBCGA consists of an intelligent genetic algorithm (Ho *et al.*, 2004b) with an inheritable mechanism. The intelligent genetic algorithm uses a divide-and-conquer strategy and an orthogonal array crossover to efficiently solve large-scale parameter optimization problems. In this study, the intelligent genetic algorithm can efficiently explore and exploit the search space of $C(n, r)$. IBCGA can efficiently search the space of $C(n, r \pm 1)$ by inheriting a good solution in the space of $C(n, r)$ (Ho *et al.*, 2004a). Therefore, IBCGA can economically obtain a complete set of high-quality solutions in a single run where r is specified in an interesting range such as [5, 45].

The proposed chromosome encoding scheme of IBCGA consists of both binary genes for feature selection and parametric genes for tuning SVM parameters, where the gene and chromosome are commonly used terms of genetic algorithm (GA), named GA-gene and GA-chromosome for discrimination in this article. The GA-chromosome consists of $n = 531$ binary GA-genes b_i for selecting informative properties and two 4-bit GA-genes for tuning the parameters C and γ of SVM. If $b_i = 0$, the ith property is excluded from the SVM classifier; otherwise, the ith property is included. This encoding method maps the 16 values of γ and C into $\{2^{-7}, 2^{-6}, \ldots, 2^{8}\}$. Figure 1 shows the encoding scheme of GA-chromosome and process of constructing feature vectors for fitness function evaluation using a concise example.

The feature vector for training the SVM classifier is obtained from decoding a GA-chromosome using the following steps. Consider a given peptide sequence, e.g. lysosomal acid lipase (LAL). At first, the index vectors for all selected physicochemical properties (residue volume and molecular weight in this example) are constructed from AAindex for each amino acid. Feature vector of a peptide consists of the selected features whose values are obtained by averaging the values in their corresponding index vectors. Finally, all values of the feature vectors are normalized into $[-1, 1]$ for applying SVM.

Fitness function is the only guide for IBCGA to obtain desirable solutions. To avoid from the prediction bias for some immunogenic levels, the averaged accuracies (AA) of four immunogenic levels, defined in (6), is adopted as the fitness function. The performance of selected properties associated with the parameter values of SVM is measured by 10-CV. Therefore, the fitness value of a GA-chromosome is obtained by computing the mean accuracy of 10 runs.

IBCGA with the fitness function $f(X)$ can simultaneously obtain a set of solutions, X_r, where $r = r_{\text{start}}, r_{\text{start}} + 1, \ldots, r_{\text{end}}$ in a single run. The algorithm of IBCGA with the given values r_{start} and r_{end} is described as follows:

Step 1. (Initiation) Randomly generate an initial population of N_{pop} individuals. All the n binary GA-genes have r 1s and $n - r$ 0s where $r = r_{\text{start}}$.

Step 2. (Evaluation) Evaluate the fitness values of all individuals using $f(X)$.

Step 3. (Selection) Use the traditional tournament selection that selects the winner from two randomly selected individuals to form a mating pool.

Step 4. (Crossover) Select $P_{\text{c}} \cdot N_{\text{pop}}$ parents from the mating pool to perform orthogonal array crossover on the selected pairs of parents, where P_{c} is the crossover probability.

Step 5. (Mutation) Apply the swap mutation operator to the randomly selected $P_{\text{m}} \cdot N_{\text{pop}}$ individuals in the new population, where P_{m} is the mutation probability. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual.

Step 6. (Termination test) If the stopping condition for obtaining the solution X_r is satisfied, output the best individual as X_r. Otherwise, go to Step 2.

Step 7. (Inheritance) If $r < r_{\text{end}}$, randomly change one bit in the binary GA-genes for each individual from 0 to 1; increase the number r by one, and go to Step 2. Otherwise, stop the algorithm.

2.5 Evaluation of POPI

The selected m physicochemical properties and the associated parameter setting of SVM by IBCGA are used to implement the computational system POPI for prediction of peptide immunogenicity. Four measurements were used to evaluate POPI using leave-one-out cross-validation (LOOCV) on the dataset PEPMHCI, namely percentage accuracy (ACC_i) and Matthew's correlation coefficient (MCC_i)

for the ith immunogenicity class, $i=1,\ldots,4$, and overall accuracy (OA) and averaged accuracies (AA) for all classes:

$$\mathrm{ACC}_i=\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FN}_i}\times 100\%, \tag{3}$$

$$\mathrm{MCC}_i=\frac{\mathrm{TP}_i\times\mathrm{TN}_i-\mathrm{FP}_i\times\mathrm{FN}_i}{\sqrt{(\mathrm{TP}_i+\mathrm{FN}_i)\times(\mathrm{TP}_i+\mathrm{FP}_i)\times(\mathrm{TN}_i+\mathrm{FP}_i)\times(\mathrm{TN}_i+\mathrm{FN}_i)}}, \tag{4}$$

$$\mathrm{OA}=\sum\frac{\mathrm{TP}_i}{N}, \tag{5}$$

$$\mathrm{AA}=\sum\frac{\mathrm{ACC}_i}{h} \tag{6}$$

where TP_i, TN_i, FP_i and FN_i are the number of true positive, true negative, false positive and false negative, respectively. N (=428) is the total number of sequences and h (=4) is the number of immunogenicity classes.

3 RESULTS

3.1 Mining informative physicochemical properties

IBCGA is performed to mine informative physicochemical properties using the whole dataset PEPMHCI. In this study, the parameters of IBCGA are set as $N_{\mathrm{pop}}=50$, $P_{\mathrm{c}}=0.8$, $P_{\mathrm{m}}=0.05$, $r_{\mathrm{start}}=5$ and $r_{\mathrm{end}}=45$. For each feature set with size r, IBCGA selected a small set of physicochemical properties and parameter values of SVM. Figure 2 shows a potentially good result in terms of averaged accuracy (AA), and the number of used features obtained from a single run of IBCGA using 10-CV. The result reveals that the best number of selected features is $m=23$, where the SVM classifier with $C=2$ and $\gamma=2$ has the best-averaged accuracy AA = 63.67% and overall accuracy OA = 66.12%. The SDs of AA and OA among the 10 cross-validation results are 10.87 and 9.73%, respectively.

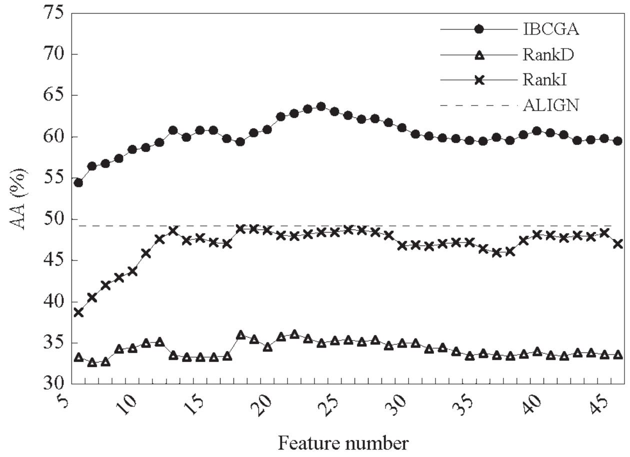


Fig. 2. Averaged accuracies (AAs) of 10-CV for IBCGA, rank-based methods (RankD and RankI) and the alignment-based method (ALIGN).

To further evaluate the feature selection of IBCGA, a traditional rank-based method for evaluating performance of a single feature is also implemented for comparison. The feature selection of the rank-based method is performed using the following steps. (1) For each physicochemical property, the prediction accuracy AA of 10-CV using SVM and the single feature was computed. (2) All physicochemical properties were ranked according to their AA accuracies. (3) The r properties with the highest ranks and SVM were used to predict peptide immunogenicity. Figure 2 shows the AA accuracies of various feature sets with size r, where $r=5,\ldots,45$.

The rank-based method suffers from the incapability of finding appropriate values of C and γ to train SVM classifiers. In order to achieve high performance, two parameter settings of SVM were tested. The first rank-based method named RankD using the default values of SVM parameters that $C=1$ and $\gamma=1/r$. The best performance of RankD is AA = 36.08% with 21 features. The second rank-based method named RankI using the same values of $C=2$ and $\gamma=2$ obtained from IBCGA. The best performance of RankI is AA = 48.87% with 18 features. Figure 2 shows the performance of RankI is better than that of RankD, revealing that the parameter setting of SVM parameters derived from IBCGA is effective. Furthermore, the performance of feature selection of IBCGA is much better than that of the rank-based method. This result is well recognized that the feature selection by additionally considering the correlated effects among physicochemical properties can advance prediction performance. Table 3 lists the AAindex identities of the 23 physicochemical properties selected by IBCGA.

3.2 Analyzing individual effects of properties

Estimating the individual effects of selected properties is important for immunologists to understand peptide immunogenicity comprehensively. Orthogonal experimental design used in IBCGA is capable of estimating individual effects of factors according to the value of MED. The property with the largest

Table 3. The AAindex identities of the 23 physicochemical properties selected by IBCGA, which are ranked according to their effectiveness of prediction

Rank by MED	ID of AAindex	Rank by RankI
1	GEIM800103	257
2	OOBM770104	347
3	PALJ810115	99
4	QIAN880132	462
5	OOBM850102	347
6	NADH010106	47
7	RADA880106	281
8	QIAN880112	347
9	WEBA780101	347
10	QIAN880125	347
11	JOND750101	48
12	QIAN880124	337
13	MUNV940101	347
14	HUTJ700102	530
15	MITS020101	481
16	KARP850103	312
17	FAUJ880113	347
18	ISOY800106	197
19	RACS820113	347
20	GEOR030105	308
21	QIAN880114	336
22	DIGM050101	347
23	MIYS850101	94

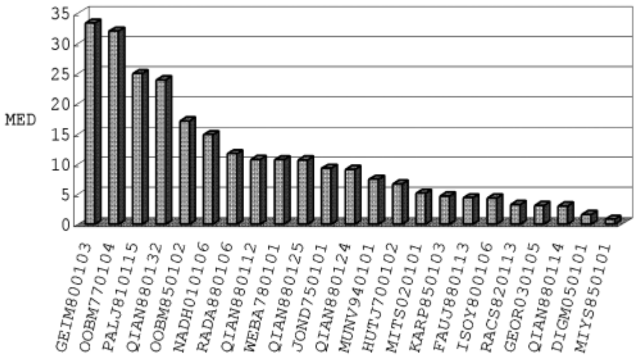


Fig. 3. Individual effects of 23 selected properties sorted by MED.

value of MED is the most effective property. Figure 3 shows the value of MED for each selected property. The property of AAindex identity GEIM800103 is the most effective property with MED = 33.29, which corresponds to 'Alpha-helix indices for beta-proteins' (Geisow and Roberts, 1980). The least effective property is MIYS850101 with MED = 0.80, which corresponds to 'Effective partition energy' (Miyazawa and Jernigan, 1985).

Since all the properties were selected at the same time based on the prediction performance, the feature set obtained by IBCGA would be not the same always for each run of IBCGA due to the reasons: (1) IBCGA is a non-deterministic algorithm; (2) the selected kernel function and parameter setting of SVM would slightly affect the prediction performance and (3) the feature selection is a machine-learning approach and its result depends on the distribution of samples in the dataset. A larger training dataset would make the selected feature set more stable.

In the computer experiment of mining informative features, there are 72 independent runs performed by IBCGA. The largest, mean and smallest numbers m of selected features are 45, 29.10 and 8, respectively. The highest, mean and lowest AA accuracies in the training phase are 63.78, 61.11 and 58.56%, respectively. The statistic result reveals that a small set of effective properties is more stable in each run of IBCGA. For example, the three properties QIAN880112, MITS020101 and KARP850103 with ranks 8, 15 and 16 shown in Figure 3 have the highest ranks 1, 6 and 6, respectively, according to the selection frequency in the 72 runs.

Table 3 also shows the ranks of the selected properties based on the prediction accuracy of RankI. The best one of selected properties is NADH010106 in terms of the rank by RankI, which has the accuracy of AA = 32.98% and rank 47. On the other hand, the most effective property GEIM800103 has the rank 257 by RankI. Table 3 reveals that the ranks by RankI for the 23 selected properties are uniformly distributed. This scenario indicates that a set of properties should be considered simultaneously rather than single property at a time because of strong correlation among physicochemical properties.

3.3 Prediction system POPI

The prediction system POPI is implemented by adopting the 23 selected informative properties (shown in Table 3) and the established SVM-based classifier in the training phase. To evaluate the ability of POPI in predicting novel peptides, the LOOCV performance is applied on the whole dataset PEPMHCI.

Table 4. Performance comparisons of ALIGN, PSI-BLAST and POPI using LOOCV on the whole dataset PEPMHCI

Immunogenicity class	ALIGN		PSI-BLAST		POPI	
	ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC
None	69.44	0.61	82.14	0.59	83.33	0.63
Little	39.76	0.32	45.59	0.40	50.60	0.44
Moderate	39.00	0.22	34.67	0.12	55.00	0.47
High	62.38	0.37	46.99	0.37	59.41	0.49
OA	54.91		53.23		64.72	
AA	52.64		52.35		62.09	

Table 4 shows the performance of POPI in terms of ACC and MCC for the four immunogenicity classes, and the prediction accuracies of OA and AA. The ACC accuracies of the four classes None, Little, Moderate and High are 83.33, 50.60, 55.00 and 59.41%, respectively. The mean of MCC performance is 0.51.

The test performance of POPI (OA = 64.72 and AA = 62.09%) is slightly worse than the training performance (OA = 66.12 and AA = 63.67%). This result indicates that the overfitting problem is not obviously occurred in selecting informative features.

3.4 Alignment-based prediction

Sequence alignment may be an efficient approach to predicting peptide immunogenicity because similar sequences may have similar peptide immunogenicity. In order to compare the alignment-based prediction methods with POPI, two methods including global sequence alignment tool ALIGN (Myers and Miller, 1988) and advanced sequence comparison method PSI-BLAST that is capable of detecting remote homologs (Altschul *et al.*, 1997) were applied to search for similar sequences. For each tested peptide, ALIGN and PSI-BLAST using three iterations were applied separately to search for its homologs.

For comparison, LOOCV was used to evaluate their prediction performances on the same dataset. The immunogenicity class with the highest similarity score was assigned to the test peptide. If there are multiple peptides with the same score, voting strategy is applied. Otherwise, if two or more immunogenicity classes have equal votes, the candidate immunogenicity classes will be ranked by sample size in the dataset and the immunogenicity class with highest rank was assigned to the test peptide.

Table 4 shows the results of ALIGN (OA = 54.91 and AA = 52.64%) and PSI-BLAST (OA = 53.23 and AA = 52.35%). Notably, the accuracy of PSI-BLAST shown in Table 4 is measured by considering only the peptides whose homologs can be obtained. When considering the 118 of 428 peptides with no homolog found, the accuracy of PSI-BLAST

Table 5. Performance comparisons between AFFIPRE and POPI

Immunogenicity class	Number of peptides	AFFIPRE		POPI	
		ACC (%)	MCC	ACC (%)	MCC
None and Little	87	35.63	0.17	80.46	0.39
Moderate	31	32.26	0.01	25.81	0.23
High	42	52.38	0.15	45.24	0.27
OA		39.38		60.63	
AA		40.09		50.50	

would be decreased. The lines shown in Figure 2 represents the performance (AA = 49.23 and OA = 51.17%) of ALIGN using 10-CV for comparison. The results reveal that POPI performs well compared with the alignment-based methods ALIGN and PSI-BLAST, when the size of the reference dataset is not sufficiently large. Notably, the performance of all methods using 10-CV is not significantly different from that using LOOCV.

3.5 Affinity-driven prediction

In the past, affinity was considered as an important index to predict peptide immunogenicity. To evaluate the affinity-driven prediction method, an additional dataset was established by extracting MHC class I binding peptides with known activity levels in both fields of 'BINDING' and 'IMMUNOGENICITY' from the MHCPEP database. However, there are four levels in the field of 'IMMUNOGENICITY', but the field of 'BINDING' has only three levels without the level 'none'. To fairly evaluate the prediction performance of the affinity-driven prediction, the immunogenic class None was combined with the class Little. The dataset contains 160 peptides belonging to three classes.

To evaluate the affinity-driven prediction method, a prediction system named AFFIPRE to predict peptide immunogenicity was implemented using the following criterion. If the immunogenic level and the affinity level of a peptide are identical, this test is regarded as a successful prediction. Otherwise, this prediction is fail. The four measurements were used to evaluate AFFIPRE, which are the same with those for IBCGA. Table 5 shows the results of AFFIPRE (OA = 39.38% and AA = 40.09%) and POPI (OA = 60.63% and AA = 50.50%). The poor performance of AFFIPRE reveals that the affinity only cannot be directly used to predict peptide immunogenicity and this result is consistent with previous studies that the affinity of peptide-MHC molecules is not the main factor for predicting peptide immunogenicity (Feltkamp *et al.*, 1994; Ochoa-Garay *et al.*, 1997).

4 DISCUSSIONS

The effectiveness of vaccination depends on peptide immunogenicity in designing peptide-based vaccines. Accurate prediction of peptide immunogenicity will decrease many experimental efforts. This study investigates the prediction problem of peptide immunogenicity and proposes an efficient prediction system POPI to predict immunogenicity of peptides with variable lengths. POPI is an SVM-based classifier with a set of informative features selected by the proposed IBCGA.

In this study, a dataset PEPMHCI of peptides associated with human MHC class I molecules extracted from MHCPEP was established. Considering the correlated effects among physicochemical properties and the cooperation with the SVM classifier, both feature selection and parameter tuning are simultaneously optimized using IBCGA. A feature set consisting of 23 physicochemical properties was selected to implement the prediction system POPI.

To our knowledge, POPI is the first computational system for prediction of peptide immunogenicity based on physicochemical properties. To evaluate POPI comprehensively, the feature selection method was compared with a rank-based selection method and the selected properties were analyzed using the factor analysis of orthogonal experimental design. Simulation results show that IBCGA can select a small set of informative properties considering the correlated effects, compared with the rank-based method.

In order to further evaluate POPI, three prediction methods were tested for comparison, namely the alignment-based methods ALIGN and PSI-BLAST, and the affinity-driven prediction method AFFIPRE. Because the reference dataset is not sufficiently large, ALIGN and PSI-BLAST cannot work well. This poor performance of AFFIPRE shows that affinity is not suitable to predict peptide immunogenicity directly. This result is consistent with previous studies that the peptide immunogenicity does not strongly correlate with its affinity for the MHC molecule (Feltkamp *et al.*, 1994; Ochoa-Garay *et al.*, 1997).

To cope with the small size of the training dataset in mining informative physicochemical properties, the proposed method can provide each selected property with the effectiveness according to its main effect difference in discriminating immunogenic levels and the robustness in terms of selection frequency. The valuable information is helpful in determining a best set of features to implement an accurate prediction system as well as to further understand immune responses from the informative physicochemical properties. The future work is to collect more immunogenicity data by combining biological knowledge and related sources, such as Immune Epitope Database and Analysis Resource (IEDB, Peters *et al.*, 2005), to advance the prediction performance of IBCGA.

In fact, the feature selection method of IBCGA has been shown effective in solving large-scale binary combinatorial optimization problems (Ho *et al.*, 2004a, b, 2006). On the other hand, the SVM-based learning methods are shown effective for protein sequence-based predictions. As a result, IBCGA with SVM can be easily used to design an SVM-based classifier for solving sequence-based prediction problems by mining informative features of physicochemical properties from an experimental dataset.

ACKNOWLEDGEMENT

The authors would like to thank the National Science Council of Taiwan for financially supporting this research under the contract numbers NSC 95-2627-B-009-002 and NSC 95-2221-E-009-116.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped. BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bhasin,M. and Raghava,G.P. (2004) Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci*., **13**, 596–607.

Bhasin,M. and Raghava,G.P. (2005) Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. *Nucleic Acids Res*., **33**, W202–W207.

Blythe,M.J. and Flower,D.R. (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci*., **14**, 246–248.

Brusic,V. *et al*. (1998) MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res*., **26**, 368–371.

Cao,Y. *et al*. (2006) Prediction of protein structural class with rough sets. *BMC Bioinformatics*, **7**, 20.

Chang,C.C. and Lin,C.J. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Deavin,A.J. *et al*. (1996) Statistical comparison of established T-cell epitope predictors against a large database of human and murine antigens. *Mol. Immunol*., **33**, 145–155.

Dey,A. (1985) *Orthogonal Fractional Factorial Designs*. Wiley, New York.

Dönnes,P. and Elofsson,A. (2002) Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*, **3**, 25.

Dönnes,P. and Kohlbacher,O. (2005) Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci*., **14**, 2132–2140.

Feltkamp,M.C. *et al*. (1994) Efficient MHC class I-peptide binding is required but does not ensure MHC class I-restricted immunogenicity. *Mol. Immunol*., **31**, 1391–1401.

Geisow,M.J. and Roberts,R.D.B. (1980) Amino acid preferences for secondary structure vary with protein class. *Int. J. Biol. Macromol*., **2**, 387–389.

Hämmerling,G.J. *et al*. (1999) Antigen processing and presentation – towards the millennium. *Immunol. Rev*., **172**, 5–9.

Ho,S.-Y. *et al*. (2004a) Inheritable genetic algorithm for bi-objective 0/1 combinatorial optimization problems and its applications. *IEEE Trans. Syst. Man Cybern. B Cybern*., **34**, 609–620.

Ho,S.-Y. *et al*. (2004b) Intelligent evolutionary algorithms for large parameter optimization problems. *IEEE Trans. Evol. Comput*., **8**, 522–541.

Ho,S.-Y. *et al*. (2006) Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. *Biosystems*, **85**, 165–176.

Idicula-Thomas,S. *et al*. (2006) A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. *Bioinformatics*, **22**, 278–284.

Kanduc,D. (2005) Peptimmunology: immunogenic peptides and sequence redundancy. *Curr. Drug Discov. Technol*., **2**, 239–244.

Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res*., **28**, 374.

Keşmir,C. *et al*. (2002) Prediction of proteasome cleavage motifs by neural networks. *Protein Eng*., **15**, 287–296.

Larsen,M.V. *et al*. (2005) An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol*., **35**, 2295–2303.

Liu,W. *et al*. (2006) Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics*, **7**, 182.

Miyazawa,S. and Jernigan,R.L. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

Myers,E.W. and Miller,W. (1988) Optimal alignments in linear space. *Comput. Appl. Biosci*., **4**, 11–17.

Nanni,L. and Lumini,A. (2006) An ensemble of K-local hyperplanes for predicting protein-protein interactions. *Bioinformatics*, **22**, 1207–1210.

Nielsen,M. *et al*. (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, **20**, 1388–1397.

Ochoa-Garay,J. *et al*. (1997) The ability of peptides to induce cytotoxic T cells in vitro does not strongly correlate with their affinity for the H-2Ld molecule: implications for vaccine design and immunotherapy. *Mol. Immunol*., **34**, 273–281.

Peters,B. *et al*. (2003) Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol*., **171**, 1741–1749.

Peters,B. *et al*. (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol*., **3**, e91.

Sarda,D. *et al*. (2005) pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinformatics*, **6**, 152.

Van Regenmortel,M.H. (2001) Antigenicity and immunogenicity of synthetic peptides. *Biologicals*, **29**, 209–213.

Wu,Q. (1978) On the optimality of orthogonal experimental design. *Acta Math. Appl. Sin*., **1**, 283–299.