GPS-MBA: Computational Analysis of MHC Class II Epitopes in Type 1 Diabetes

Ruikun Cai1*, Zexian Liu2*, Jian Ren3, Chuang Ma4, Tianshun Gao1, Yanhong Zhou1, Qing Yang1*, Yu Xue1*

1 Huabei Bioinformatics and Molecular Imaging Key Laboratory, Department of Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China, 2 Hefei National Laboratory for Physical Sciences, Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, China, 3 State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong, China, 4 Saban Research Institute of Children's Hospital Los Angeles, Department of Pediatrics, University of Southern California, Los Angeles, California, United States of America

Abstract

As a severe chronic metabolic disease and autoimmune disorder, type 1 diabetes (T1D) affects millions of people worldwide. Recent advances in antigen-based immunotherapy have provided a great opportunity for further treating T1D with a high degree of selectivity. It is reported that MHC class II I-A\(^{\text{b}}\) in the non-obese diabetic (NOD) mouse and human HLA-DQ8 are strongly linked to susceptibility to T1D. Thus, the identification of new I-A\(^{\text{b}}\) and HLA-DQ8 epitopes would be of great help to further experimental and biomedical manipulation efforts. In this study, a novel GPS-MBA (MHC Binding Analyzer) software package was developed for the prediction of I-A\(^{\text{b}}\) and HLA-DQ8 epitopes. Using experimentally identified epitopes as the training data sets, a previously developed GPS (Group-based Prediction System) algorithm was adopted and improved. By extensive evaluation and comparison, the GPS-MBA performance was found to be much better than other tools of this type. With this powerful tool, we predicted a number of potentially new I-A\(^{\text{b}}\) and HLA-DQ8 epitopes. Furthermore, we designed a T1D epitope database (TEDB) for all of the experimentally identified and predicted T1D-associated epitopes. Taken together, this computational prediction result and analysis provides a starting point for further experimental considerations, and GPS-MBA is demonstrated to be a useful tool for generating starting information for experimentalists. The GPS-MBA is freely accessible for academic researchers at: http://mba.biocuckoo.org.

Introduction

Type 1 diabetes (Diabetes mellitus type 1, T1D or T1DM) is a severe chronic autoimmune disease with a relapsing-remitting course that is characterized by the insidious loss of self-tolerance and progressive destruction of insulin-producing pancreatic \(\beta\)-cells in the islets of Langerhans, with the presence of overt hyperglycemia at the time of clinical diagnosis [1–7]. The incidence and prevalence of T1D is believed to be controlled by both genetic and environmental factors [1–5,8]. The cumulative analysis has revealed that a variety of immune cell types, including CD4\(^{+}\), CD8\(^{+}\) T cells, macrophages and dendritic cells (DCs) are involved in \(\beta\)-cell death, and CD4\(^{+}\) T cells play the predominant role in the overall T1D pathology [1,2,5]. Thus, the development of immunoregulatory therapeutic approaches has come to be an urgent demand for preventing, treating or even curing T1D [1–7].

Besides immunosuppressive drugs and antibody-based immunotherapies, antigen-based “tolerogenic” immunotherapy has attracted considerable attention as a third-generation approach, particularly for its highly selective targeting of aberrant T cells [1–6]. It was demonstrated that the MHC class II haplotype, I-A\(^{\text{b}}\), is strongly linked to susceptibility to T1D in the non-obese diabetic (NOD) mouse [9–11]. Similar linkage to the human HLA-DQ8 molecule, I-A\(^{\text{b}}\) is expressed by DCs to present \(\beta\)-cell epitopes from certain well-defined autoantigens, including insulin, glutamic acid decarboxylase (GAD) and insulinoma antigen 2 (IA-2) [1–6,8]. These epitopes are usually composed of 10 to 30 amino acids, with a 9-amino acid core sequence for I-A\(^{\text{b}}\)/HLA-DQ8 and T-cell receptor (TCR) binding [9–11]. In this regard, identification of I-A\(^{\text{b}}\)/HLA-DQ8 epitopes is fundamental for an understanding of the molecular mechanisms of T1D and the improved design of immunotherapeutic peptides. In 2009, the first-in-human beings Phase I clinical study reported that proinsulin peptide injection is both well tolerated and safe [12]. Recently, a C-peptide deduced from the GAD 65 isoform has generated promising results in Phase II trials, and three Phase III trials are still ongoing [1,2,5].

As a complement to labor-intensive and time-consuming experimental assays, the in silico prediction of MHC-binding
epitopes has emerged as an efficient approach to generate useful information for the purposes of biomedical design [13,14] (see also http://mba.biocuckoo.org/links.php). For example, the prediction results of SYFPEITHI [13] and BIMAS [16] were successfully used for the experimental identification of novel MHC class I epitopes derived from type 1 diabetes autoantigens [17–19]. Since I-A\textsuperscript{\textit{b}} is the only expressed MHC class II molecule in the NOD mouse [9,10], additional efforts have subsequently been expended on the prediction of I-A\textsuperscript{\textit{b}} or HLA-DQ\textsuperscript{\textit{b}} epitopes [20–23]. In 2006, Rajapakse \textit{et al.} developed the first online server of PRED\textsuperscript{\textit{aD}} for the prediction of I-A\textsuperscript{\textit{b}}, and the two MHC class I molecules K\textsuperscript{\textit{d}} and D\textsuperscript{\textit{b}} binding peptides in the NOD mouse [20]. They subsequently refined the predictor using multi-objective evolutionary algorithms (MOEA) [23]. Chang \textit{et al.} used an expectation-maximization alignment algorithm to design computational programs for the prediction of I-A\textsuperscript{\textit{b}} [21] and HLA-DQ\textsuperscript{\textit{b}} [22] epitopes, respectively. Furthermore, the two integrative tools of MHC2Pred [24,25] and RANKPEP [26] also include predictors for I-A\textsuperscript{\textit{b}} and HLA-DQ\textsuperscript{\textit{b}}, although they were developed for the comprehensive prediction of a variety of MHC class I and/ or II binding peptides. Currently, although a number of computational studies have been performed, only MHC2Pred [24,25] and RANKPEP [26] are accessible over the internet.

In this work, we developed a novel GPS-MBA software package for the prediction of I-A\textsuperscript{\textit{b}} and HLA-DQ\textsuperscript{\textit{b}}. The experimentally identified epitopes were obtained from the scientific literature, and a modified Gibbs sampling approach was adopted to determine the core nonamers in these epitopes. For training and prediction, a refined GPS algorithm [27,28] was used. By extensive evaluation and comparison, the prediction performance of GPS-MBA was shown to be highly promising and much better than the other tools currently in use. Moreover, by cross-evaluation using the HLA-DQ\textsuperscript{\textit{b}} predictor in GPS-MBA to predict the I-A\textsuperscript{\textit{b}} epitopes and vice versa, the results show that I-A\textsuperscript{\textit{b}} and HLA-DQ\textsuperscript{\textit{b}} recognize highly similar peptide profiles. With this powerful tool, we predicted potentially novel I-A\textsuperscript{\textit{b}} and HLA-DQ\textsuperscript{\textit{b}} binding peptides from T1D-associated epitopes, which bind to other types of MHC molecules. All of the experimentally identified T1D antigens together with their epitopes were absorbed into TEDB 1.0. The \textit{ab initio} predicted epitopes were also provided. Taken together, the prediction and analysis results are helpful for further experimental investigation, and the GPS-MBA can serve as a practically useful adjunct program for experimentalists. The online service and local packages of GPS-MBA 1.0 were implemented in JAVA and freely accessible for academic research purposes at: http://mba.biocuckoo.org.

### Methods

#### Data preparation

A search of the scientific literature from PubMed (before Sept. 20th, 2011) with the keywords “I-A\textsuperscript{\textit{b}} peptide”, “HLA-DQ\textsuperscript{\textit{b}} peptide”, or “Type 1 diabetes epitope”, we collected 318 experimentally verified and naturally processed mouse I-A\textsuperscript{\textit{b}} binding peptides in 177 proteins, and 134 human HLA-DQ\textsuperscript{\textit{b}} epitopes from 84 proteins (Table 1). Additional keywords were tried, but the data set was not changed. The protein sequences were retrieved from the UniProt database (http://www.uniprot.org/uniprot/).

In our data set, the length of most of the epitopes varies from 9–30aa. Thus, we adopted a refined Gibbs sampling approach [29,30] to determine the 9aa core peptides, and obtained 301 unique nonamer epitopes for the I-A\textsuperscript{\textit{b}} and 127 HLA-DQ\textsuperscript{\textit{b}} core peptides for training. We also prepared positive (\textdagger) and negative (\textdaggerdbl) data sets for testing. The 318 I-A\textsuperscript{\textit{b}} and 134 HLA-DQ\textsuperscript{\textit{b}} epitopes of known length were regarded as the (\textdagger) set. If at least one predicted nonamer is fully located in the epitope region, the epitope is predicted as a positive hit. To avoid any bias, all of the 9-mer lengths of the same proteins which were either not covered or not fully covered by the original epitopes were taken to be the (\textdaggerdbl) set.

By further literature review, we collected 203 T1D-associated epitopes in 25 proteins which have the potential to be recognized by other types of MHC or unknown molecules (Table 1).

#### Performance evaluation

As previously described [27,28], we used the four measurements of sensitivity (\textit{Sn}), specificity (\textit{Sp}), accuracy (\textit{Ac}) and Mathew’s Correlation Coefficient (\textit{MCC}). The measurements were defined as below:

\[
\text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad \text{Ac} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP}}
\]

and

\[
\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FN} \times \text{FP})}{\sqrt{(\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FN})}}.
\]

To evaluate the prediction performance and robustness, the leave-one-out (LOO) validation and 4-, 6-, 8- and 10-fold (\textit{n}-fold) cross-validations were performed. In the LOO validation, each core nonamer in the data set was picked out in turn as an independent test sample, and all the remaining core nonamers were regarded as training data. This process was repeated until each nonamer was used as test data one time. In the \textit{n}-fold cross-validation, all the (\textdagger) core nonamers and (\textdaggerdbl) nonamers were mixed and then divided equally into \textit{n} parts, keeping the same distribution of (\textdagger) and (\textdaggerdbl) nonamers in each part. Then \textit{n}-1 parts were merged into a training data set while the remnant part was taken as a testing data set. This process was repeated 20 times and the average performance of \textit{n}-fold cross-validation was computed. Furthermore, the Receiver Operating Characteristic (ROC) curves were drawn, and AROC (area under ROC) values were calculated.

#### The algorithm

Previously, we developed a series of GPS algorithms for the prediction of post-translational modification (PTM) sites in proteins [27,28]. For the prediction of the I-A\textsuperscript{\textit{b}} and HLA-DQ\textsuperscript{\textit{b}} binding peptides, we used the original method to develop a new

| Table 1. The statistical data on the experimentally validated epitopes collected in this study. |
|-------------------------------------------------|----------------|----------------|----------------|
| **Experimental data set**                       | **Protein** | **Epitope** | **Core 9-mer** |
| Mouse I-A\textsuperscript{\textit{b}}           | 177         | 318          | 301            |
| Human HLA-DQ\textsuperscript{\textit{b}}       | 84          | 134          | 127            |
| Other T1D-associated                            | 25          | 203          |                |
| **Total**                                       | 245         | 623          |                |

\textsuperscript{\textdagger}Epitope, the number of original epitopes; \textsuperscript{\textdaggerdbl}Core 9-mer, the number of nonamer core peptides derived from the adapted Gibbs sampling procedure.

\textbf{doi:10.1371/journal.pone.0033884.t001}
algorithm containing two computational parts, a scoring strategy and performance improvement.

The basic hypothesis behind the scoring strategy is that similarly short peptides would exhibit similar 3D structures and biochemical properties [27,28]. Thus, we can directly use an amino acid substitution matrix, e.g., BLOSUM62, to calculate the similarity between two 9-mer peptides A and B as:

\[ S(A, B) = \sum_{1 \leq i \leq 9} \text{Score}(A[i], B[i]) \]

\( \text{Score}(A[i], B[i]) \) denotes the substitution score of the two amino acids of \( A[i] \) and \( B[i] \) in the BLOSUM62 at the position \( i \). If \( S_A, B_i < 0 \), we redefined it as \( S_A, B_i = 0 \). A given nonapeptide is then compared with each of the 9aa core peptides from the training data in a pairwise manner to calculate the similarity score. The average value of the substitution scores is regarded as the final score.

The performance improvement procedure is comprised of two steps, weight training (WT) and matrix mutation (MaM). To evaluate the degree of performance improvement, we calculated the scores for all sites of the training data set in each time. By gradually increasing the threshold, we computed the \( S_n \) at 90% and compared the \( S_n \) values of the LOO validation.

1) Weight training (WT). In this step, the substitution score between the two 9-mer peptides A and B was updated as follows:

\[ S'(A, B) = \sum_{1 \leq i \leq 9} w_i \text{Score}(A[i], B[i]) \]

Initially, the weight of each position was defined as 1. The \( w_i \) value is the weight of the position \( i \). Again, if \( S'(A, B) \) is < 0, we redefined it as \( S'(A, B) = 0 \). Then we randomly picked out the weight of any position for +1 or −1 and re-computed the LOO result. The manipulation was adopted if the \( S_n \) value was increased. This process was continued until the \( S_n \) value did not increase any further.

2) Matrix mutation (MaM). The aim of this step is to generate an optimal or near-optimal scoring matrix from an initial matrix. We re-calculated the LOO result to improve the \( S_n \) value by randomly picking out an element of the BLOSUM62 matrix for +1 or −1. The process was repeated until convergence was reached. Selecting a different initial matrix, e.g., BLOSUM45 will generate a convergent result if the training time is sufficient (Data not shown).

Implementation of the online service and local packages

The online service and local packages of GPS-MBA 1.0 were implemented in JAVA. For the online service, we tested GPS-MBA 1.0 on a variety of internet browsers, including Internet Explorer 6.0, 8.0 and 9.0, Mozilla Firefox 8.0 and Google Chrome under the Windows XP and Windows 7 Operating System (OS), Mozilla Firefox 8.0 and Google Chrome under Fedora Core 15 and Ubuntu 10.04 LTS (Linux), and Safari 5.1.1 under Apple Mac OS ×10.5 (Leopard) and 10.7 (Lion). For the Windows and Linux systems, the latest version of Java Runtime Environment (JRE) package (Java SE 5.0 or later versions) should be pre-installed. However, for Mac OS, GPS-MBA 1.0 can be directly used without any additional packages. For convenience, we also developed local packages of GPS-MBA 1.0 which support the three major Operating Systems Windows, Linux and Mac OS X.

Results

Determination of the core nonamers from the I-A\(^{\beta7}\) and HLA-DQ8 epitopes

From the scientific literature, we collected 318 naturally processed I-A\(^{\beta7}\) epitopes and 134 HLA-DQ8 binding peptides of various lengths from 8–30aa (Table 1). The prerequisite for the usage of GPS algorithm is that the length of peptides must be fixed and identical in the training data set [27,28]. Previously, experimental analyses had suggested that the I-A\(^{\beta7}\) and HLA-DQ8 epitopes contain the 9aa core sequences needed for recognition and binding [9–11]. Since there is only one epitope with 8aa, we added one residue upstream and one residue downstream for the 9aa peptide so as to form a decapptide. Then we used an adapted Gibbs sampling approach to determine the core nonamers of the I-A\(^{\beta7}\) (Table S1) and HLA-DQ8 epitopes (Table S2) [29–31].

Given a set of \( N \) epitopes \( S_1, ..., S_N \), we sought to identify the most probable nonapeptides that were mutually present in both epitopes (Figure 1). First, one 9-mer length per epitope as \( P_1, ..., P_N \) was randomly selected, while we randomly singled out one epitope \( S_i \) together with its nonapeptide \( P_i \) (Figure 1). Then we calculated the similarity scores for all of the 9-mer peptides sequentially in \( S_i \) as described below:

\[ \text{Score}(P) = \sum_{1 \leq i \leq N, j \neq i} S(P_i, P_j) / N - 1 \]

The \( \text{Score}(P) \) is the average final similarity score compared with the other nonamers, while \( S(P_i, P_j) \) is the similarity score between \( P_i \) and \( P_j \). The 9-mer with the maximal \( \text{Score}(P) \) is sampled and then updated to the new nonapeptide \( P_i \) (Figure 1). Such a sampling procedure was iteratively repeated until convergence was attained (Figure 1). Ultimately, the redundant 9-mer core peptides were made clearly evident (Table 1, Table S1 and S2).

Development of GPS-MBA for the prediction of I-A\(^{\beta7}\) and HLA-DQ8 binding peptides

The series of GPS algorithms contain two computational procedures of a scoring strategy and performance improvement [27,28]. The scoring step has remained the same in all the versions of the GPS algorithms, while the latter process is still in the process.
of being improved for better performance [27,28]. In the latest GPS 3.0 release, the performance improvement procedure is comprised of the four sequential steps of k-means clustering, motif length selection (MLS), WT and MaM [28]. Originally, we proposed that such a training order could not be changed [28], whereas our recent analysis instead suggests that such an order can be shuffled if the training time is sufficient (Data not shown).

When the training data set is large, the k-means clustering approach can be used to classify positive data into multiple groups [28]. However, due to limited data available, this method was not used in this work. The MLS approach was designed for determining the motif length for optimal performance [28]. Since experimental studies suggest that the nonamer core peptides are essential for I-A\textsuperscript{\textgamma} and HLA-DQ\textgreek{a} recognition [9–11], this strategy was also not adopted.

Taken together, the GPS algorithm has been improved, and the performance improvement process only required the two steps of WT and MaM. The training order was shuffled for better performance, and the online service and software packages of GPS-MBA 1.0 were implemented in JAVA. As an example, the prediction results for the mouse Ig\textkappa chain C region (UniProt ID: P01837) are shown (Figure 2). Previously, a peptide in the mouse Ig\textkappa L chain (\textsuperscript{174}ERQNGV LNSWTDQDS\textsuperscript{188}, identical to 46–60 in the Ig\textkappa chain C region) was sequenced as an I-A\textsuperscript{\textgamma} epitope by MS/MS analysis [32]. In our results, four potential I-A\textsuperscript{\textgamma} epitopes of APTVSIFPP\textsuperscript{4-12}, NGVLSWTVD\textsuperscript{49-57}, SSTTLTLTKD\textsuperscript{66-76}, and SSTTLTLTKE\textsuperscript{69-77} were predicted (Figure 2). The NGVLSWTVD\textsuperscript{49-57} nonamer gives complete coverage of the experimental epitope (46–60), while the other three predicted hits are available for further experimental investigation.

Performance evaluation and comparison

To evaluate the prediction performance and robustness of GPS-MBA, LOO validation and 4-, 6-, 8-, 10-fold cross-validations were performed (Figure 3A, 3B). ROC curves were drawn, while the AROC values were 0.909 (LOO), 0.909 (4-fold), 0.904 (6-fold), 0.916 (8-fold) and 0.919 (10-fold) for mouse I-A\textsuperscript{\textgamma} (Figure 3A), and 0.921 (LOO), 0.933 (4-fold), 0.928 (6-fold), 0.937 (8-fold) and 0.931 (10-fold) for human HLA-DQ\textgreek{a} (Figure 3B). Since the results of the 4-, 6-, 8- and 10-fold cross-validations were closely similar to the LOO validation, the GPS-MBA 1.0 is evidently a stable and robust predictor. The performance of the LOO validation was also used for the cut-off setting and further comparison, and the three thresholds of high, medium and low were selected with the Sp values of 97%, 95% and 90%, respectively (Table 2). In addition, given the highest MCC values, the medium thresholds were chosen.
as the default thresholds of I-Ag7 (0.1541) and HLA-DQ8 (0.1534), respectively (Table 2).

To clearly demonstrate the superiority of GPS-MBA, we also used the same data sets to evaluate the performances of MHC2Pred [24,25] and RANKPEP [26]. We fixed the $Sp$ values of MHC2Pred and RANKPEP so as to be similar to GPS-MBA and compared the $Sn$ values (Table 2). For I-Ag7, when the $Sp$ value was ~97%, the $Sn$ values of GPS-MBA, MHC2Pred and RANKPEP were 62.89%, 23.58%, and 33.65%, respectively (Table 2). Also, when the $Sp$ value was ~95%, the $Sn$ values of GPS-MBA, MHC2Pred and RANKPEP were 71.70%, 32.70%, and 45.91%, respectively (Table 2). Furthermore, when the $Sp$ value was ~95% for HLA-DQ8, the $Sn$ values of GPS-MBA, MHC2Pred and RANKPEP were 62.89%, 23.58%, and 33.65%, respectively (Table 2).

Figure 3. The performance evaluation of GPS-MBA 1.0 and a comparison with other approaches. The LOO validation and 4-, 6-, 8- and 10-fold cross-validations were performed for (A) mouse I-Ag7 and (B) human HLA-DQ8, respectively. We compared the performance of GPS-MBA 1.0 with MHC2Pred and RANKPEP using the LOO validation for (C) I-Ag7 and (D) HLA-DQ8, respectively. We also performed a cross-evaluation by using the HLA-DQ8 predictor in GPS-MBA to predict (C) I-Ag7 epitopes and (D) vice versa. 

figure caption:
- Figure 3. The performance evaluation of GPS-MBA 1.0 and a comparison with other approaches. The LOO validation and 4-, 6-, 8- and 10-fold cross-validations were performed for (A) mouse I-Ag7 and (B) human HLA-DQ8, respectively. We compared the performance of GPS-MBA 1.0 with MHC2Pred and RANKPEP using the LOO validation for (C) I-Ag7 and (D) HLA-DQ8, respectively. We also performed a cross-evaluation by using the HLA-DQ8 predictor in GPS-MBA to predict (C) I-Ag7 epitopes and (D) vice versa.
value was ~90%, the Sn of GPS-MBA (84.91%) was still much better than MHC2Pred (47.80%) and RANKPEP (64.15%) (Table 2). Once again, for HLA-DQ8, the GPS-MBA performance is still much better than the other two predictors. In addition, ROC curves were drawn, showing that the AROC values of the GPS-MBA were generally better than the other approaches to I-A\textsuperscript{\gamma} (Figure 3C) and HLA-DQ8 (Figure 3D).

Previous experimental studies suggested that the mouse I-A\textsuperscript{\gamma} haplotype is equivalent to the human HLA-DQ8 linkage, and exhibits a similar specificity for peptide recognition and binding [1–6,8–11]. To investigate this viewpoint, we performed a cross-evaluation using the HLA-DQ8 predictor in GPS-MBA to predict I-A\textsuperscript{\gamma} epitopes (Figure 3C) and vice versa (Figure 3D). In our results, the cross-evaluation performance is closely similar to the LOO validations (Figure 3C, 3D, and Table 2). Thus, we propose that the binding patterns of I-A\textsuperscript{\gamma} and HLA-DQ8 are highly similar and conserved.

**Prediction of potentially new I-A\textsuperscript{\gamma} and HLA-DQ8 epitopes in T1D**

Besides I-A\textsuperscript{\gamma} and HLA-DQ8, certain other MHC class I and II molecules are also implicated in T1D [1–7,17–19]. We collected 203 epitopes in 25 proteins from the scientific literature, with 70 MHC class I epitopes, 84 MHC class II binding peptides and 49 epitopes for which the MHC molecules are still undetermined (Table 3, Table S3). Although at present there is a lack of experimental verification, we propose that a number of these epitopes will also come to be recognized by I-A\textsuperscript{\gamma} and/or HLA-DQ8.

### Table 2. Comparison of GPS-MBA 1.0 with other approaches.

| Method    | Threshold | Mouse I-A\textsuperscript{\gamma} | Human HLA-DQ8 |
|-----------|-----------|-----------------------------------|---------------|
|           |           | Ac (%)   | Sn (%)   | Sp (%)   | MCC   | Ac (%)   | Sn (%)   | Sp (%)   | MCC   |
| GPS-MBA   | High      | 96.88    | 62.89    | 97.01    | 0.2066 | 97.03    | 51.49    | 97.17    | 0.1575 |
|           | Medium    | 94.96    | 71.70    | 95.05    | 0.1829 | 95.29    | 64.93    | 95.38    | 0.1533 |
|           | Low       | 90.32    | 84.91    | 90.34    | 0.1530 | 90.11    | 81.34    | 90.13    | 0.1307 |
| Cross\textsuperscript{a} | Low       | 96.83    | 48.11    | 97.01    | 0.1569 | 97.04    | 50.00    | 97.18    | 0.1532 |
|           | Low       | 94.90    | 65.09    | 95.01    | 0.1646 | 95.21    | 60.45    | 95.32    | 0.1430 |
|           | Low       | 90.02    | 82.70    | 90.05    | 0.1461 | 90.09    | 80.60    | 90.12    | 0.1293 |
| MHC2Pred  | High      | 96.79    | 23.58    | 97.07    | 0.0736 | 96.90    | 17.16    | 97.14    | 0.0469 |
|           | Medium    | 94.79    | 32.70    | 95.02    | 0.0768 | 94.98    | 26.87    | 95.19    | 0.0564 |
|           | Low       | 89.86    | 47.80    | 90.02    | 0.0763 | 90.19    | 44.03    | 90.33    | 0.0637 |
| Rankpep   | High      | 96.77    | 33.65    | 97.01    | 0.1075 | 96.84    | 30.60    | 97.04    | 0.0886 |
|           | Medium    | 94.85    | 45.91    | 95.03    | 0.1130 | 95.13    | 40.30    | 95.30    | 0.0916 |
|           | Low       | 89.96    | 64.15    | 90.06    | 0.1093 | 90.02    | 64.18    | 90.10    | 0.0993 |

For the construction of the GPS-MBA 1.0 software, the three thresholds of high, medium and low were chosen. For comparison, we fixed the Sp values of GPS-MBA 1.0 to be identical or similar to other methods and compared the Sn values.

\textsuperscript{a}Cross, cross-evaluation by using the HLA-DQ8 predictor to predict I-A\textsuperscript{\gamma} epitopes and vice versa.

**Table 3. The statistical data for the prediction of potentially new I-A\textsuperscript{\gamma} and HLA-DQ8 epitopes in T1D.**

| Tool    | Data set | Protein | Epitope | I-A\textsuperscript{\gamma} | HLA-DQ8 | Either |
|---------|----------|---------|---------|-----------------------------|---------|--------|
|         |          |         |         | Num.\textsuperscript{a} | Per.\textsuperscript{b} | Num. | Per. | Num. | Per. |
| GPS-ARM | Class I  | 15      | 70      | 1 | 1.43% | 3 | 4.29% | 4 | 5.71% |
|         | Class II | 11      | 84      | 31 | 36.90% | 28 | 33.33% | 36 | 42.86% |
|         | Unknown  | 5       | 49      | 23 | 46.94% | 19 | 38.78% | 28 | 57.14% |
|         | Total    | 25      | 203     | 55 | 27.09% | 50 | 24.63% | 68 | 33.50% |
| MHC2Pred| Class I  | 15      | 70      | 5 | 7.14% | 2 | 2.86% | 7 | 10.00% |
|         | Class II | 11      | 84      | 25 | 29.76% | 24 | 28.57% | 42 | 50.00% |
|         | Unknown  | 5       | 49      | 22 | 44.90% | 14 | 28.57% | 29 | 59.18% |
|         | Total    | 31      | 203     | 52 | 25.62% | 40 | 19.70% | 78 | 38.42% |
| RANKPEP | Class I  | 15      | 70      | 14 | 20.00% | 3 | 4.29% | 15 | 21.43% |
|         | Class II | 11      | 84      | 42 | 50.00% | 24 | 28.57% | 48 | 57.14% |
|         | Unknown  | 5       | 49      | 28 | 57.14% | 18 | 36.73% | 33 | 67.35% |
|         | Total    | 31      | 203     | 84 | 41.38% | 45 | 22.17% | 96 | 47.29% |

\textsuperscript{a}Num., the number of manually collected epitopes predicted with core 9-mers.

\textsuperscript{b}Per., percentiles.
To predict potential I-A^*7 and HLA-DQ^8 epitopes, we used GPS-MBA 1.0 with the default thresholds. In total, 68 epitopes were predicted to interact with either I-A^*7 or HLA-DQ^8 (33.5%) (Table 3). Also, the prediction results of MHC2Pred [24,25] and RANKPEP [26] were present with default thresholds (Table 3). In particular, the prediction performance of GPS-MBA 1.0 is much poorer in terms of the MHC class I epitopes, such that only one and three epitopes were predicted as I-A^*7 and HLA-DQ^8 binding peptides, respectively (Table 3). However, GPS-MBA 1.0 displayed a considerably effective performance for MHC class II epitopes by predicting 31 I-A^*7 and 28 HLA-DQ^8 binding peptides (Table 3). Also, the distributions of results from MHC2Pred and RANKPEP are similar (Table 3). In this regard, the sequence profiles of MHC class I and II are quite different, whereas GPS-MBA, MHC2Pred and RANKPEP have the

### Table 4. The statistical results of the TEDB 1.0 database.

| Organism     | Protein Known epitopes | Predicted epitopes | Total | I-A^*7 | HLA-DQ^8 | Either |
|--------------|------------------------|--------------------|-------|--------|----------|--------|
| H. Sapiens   | 20                     | 553                | 773   | 944    |          |        |
| M. musculus  | 180                    | 6,427              | 6,587 | 6,485  |          |        |
| R. norvegicus| 13                     | 331                | 505   | 503    |          |        |
| Others       | 32                     | 801                | 1,123 | 1,105  |          |        |
| Total        | 245                    | 6,312              | 8,988 | 9,037  |          |        |

doi:10.1371/journal.pone.0033884.t004

Figure 4. The search options for the TEDB 1.0 database. (A) Users are able to simply input ‘RAN’ and select “Gene Name” for querying. (B) The results are shown in a tabular format. Users can then click on the TEDB ID (TEDB-HS-00015) to visualize the detailed information. (C) The detailed information on human RAN. The experimentally identified and predicted epitopes are presented.
doi:10.1371/journal.pone.0033884.g004
The development and usage of TEDB 1.0

To provide an integrative platform for computational analysis of the T1D epitopes, all of the experimentally identified T1D-associated antigens together with their epitopes were collected for the development of TEDB (Table 4). Also, we used GPS-MBA 1.0 at the default threshold to predict potential I-A\(^{\beta}\) or HLA-DQ\(^{\beta}\) binding peptides in 245 antigens (Table 4). In addition, the prediction results of MHC2Pred and RANKPEP were also included in the TEDB 1.0.

TEDB 1.0 was developed in a user-friendly manner. The search option (http://mba.biocuckoo.org/database.php) provides an interface for querying the TEDB database with one or several keywords such as TEDB ID, MHC Type, or UniProt Accession, etc (Figure 4A). We also provided three advance options, including advance search, browse and BLAST search (Figure 4A). For example, if the keyword ‘RAN’ is inputted and submitted (Figure 4A), the result is shown in a tabular format, with the features of TEDB ID, UniProt accession, and protein/gene names/aliases (Figure 4B). By clicking on the TEDB ID (TEDB-HS-00015), the detailed information on human RAN is shown (Figure 4C). The experimentally identified epitopes and predicted I-A\(^{\beta}\) and HLA-DQ\(^{\beta}\) binding peptides are provided, while the protein sequence, Gene Ontology annotation, domain organization, molecular weight and computed/theoretical \(I_p\) are also provided (Figure 4C).

Discussion

T1D, also termed Insulin-dependent diabetes mellitus (IDDM) or juvenile diabetes, is a chronic metabolic disorder and autoimmune disease that takes a heavy toll world-wide [1–7,33]. Currently, it is estimated that approximately 70,000 children develop T1D per year, ~30,000 of which cases are in the United States, and an annual global increase of ~3% [4,5,33]. Although a number of immunotherapeutic and regenerative approaches have been developed, the effective prevention and treatment of T1D are both still a great challenge [1–7,33]. Recent advances in the development of antigen-based “tolerogenic” immunotherapy have provided a great opportunity for treating T1D with a high degree of selectivity, while two investigations are still ongoing as clinical trials, and the outcome is uncertain [1–7,12,33]. In this regard, the Identification of T1D-specific epitopes is needed for further experimental and biomedical design.

Although a number of computational analyses were carried out for the prediction of the I-A\(^{\beta}\) or HLA-DQ\(^{\beta}\) epitopes, the online services have been not available over the internet [20–23]. Again, although the two integrative tools of MHC2Pred and RANKPEP do contain predictors of I-A\(^{\beta}\) and HLA-DQ\(^{\beta}\), they were actually designed for the general prediction of MHC class I and/or II epitopes [24–26]. Thus, they might exhibit a sensitive performance for other MHC molecules, and yet not I-A\(^{\beta}\) and/or HLA-DQ\(^{\beta}\).

In this study, we focused on the prediction of I-A\(^{\beta}\) and HLA-DQ\(^{\beta}\) by constructing the GPS-MBA software package. By comparison, the performance of GPS-MBA is much better than MHC2Pred and RANKPEP (Table 2, Figure 3C, 3D).

Originally, the series of GPS algorithms were developed for the prediction of PTM sites in proteins [27,28], and this is the first use of the algorithm for MHC epitopes. The major difference between PTM sites and MHC epitopes is that the PTM site positions are anchored by the middle residues, while the lengths and positions of the MHC epitopes are promiscuous and difficult to fix. In this regard, the MHC epitope core sequences of defined length must be determined prior to training. Originally, the Gibbs sampling method was designed for detecting short conserved motifs from multiple DNA or protein sequences [29–31]. By utilizing the position-specific scoring matrices (PSSMs), the amino acids frequencies were counted in the foreground and background data sets, respectively. In this procedure, the ratio of the pattern probability to the background probability was calculated and improved by sampling, until convergence was attained [29–31]. However, the average similarity score, but not the frequency ratio, was calculated for further sampling in our analysis. The 9-mer core peptides were determined by this approach for I-A\(^{\beta}\) and HLA-DQ\(^{\beta}\), respectively.

Here, we used the WebLogo server [34] to analyze the sequence profiles of the core nonamers for I-A\(^{\beta}\) (Figure 5A) and HLA-DQ\(^{\beta}\) (Figure 5B), respectively. Previously, experimental studies based on limited epitopes had suggested that P4, P6 and P9 are three conserved positions, while P1 is a degenerate position [9–11]. However, our results suggest that P9 is the most informative position, along with a less the comparatively weakly conserved position of P6 (Figure 5A, 5B). P4 is weakly conserved in I-A\(^{\beta}\) core nonamers (Figure 5A) and not conserved in HLA-DQ\(^{\beta}\) (Figure 5B). Although the sequence logo of I-A\(^{\beta}\) is not evidently similar to HLA-DQ\(^{\beta}\), our cross-validation results indicate that I-A\(^{\beta}\) and HLA-DQ\(^{\beta}\) are able to recognize similar peptide profiles (Table 2, Figure 3C, 3D). We speculated that whether most of the HLA-DQ\(^{\beta}\) epitopes would also be I-A\(^{\beta}\) binding peptides in the training data set, since the NOD mouse is the best model for T1D. However, there are only 24 epitopes in the original data set which were found to bind with both I-A\(^{\beta}\) and HLA-DQ\(^{\beta}\). Based on these
results, we propose that the sequence profile of I-A\textsuperscript{G7} is intrinsically similar to that of HLA-DQ\textsubscript{8}.

In addition to the predictions of GPS-MBA, all of the experimentally validated T1D-associated epitopes were collected in the online validated T1D database of TEDB 1.0. The affinity predictions with GPS-MBA were also integrated into TEDB. Thus, such an integrative platform should prove to be useful for experimentalists. We believe that computational analysis, together with subsequent experimental identification, will help advance the study of T1D into a new and highly productive phase.

Supporting Information

Table S1 The core nonamers of the mouse I-A\textsuperscript{G7} epitopes. a. Epitope, the original epitopes; b. Position, the positions of core nonamers in the protein sequences; c. Core nonamer, the finally aligned core nonamers were marked in red; d. The protein sequence was retrieved from the UniParc (UniProt Archive) Database (http://www.uniprot.org/help/uniparc).

Table S2 The core nonamers of the human HLA-DQ\textsubscript{8} epitopes.

Table S3 From the scientific literature, we collected 203 epitopes in 25 proteins, with 70 MHC class I epitopes, 84 MHC class II binding peptides, and 49 epitopes for which the MHC molecules are still unknown. The detailed prediction results of GPS-ARM were provided. a. UniProt, the UniProt accession numbers of T1D antigens; b. Pos., the position of the original known epitopes; c. Peptide, the experimentally identified epitopes; d. MHC Type, the experimentally identified MHC molecules that recognize the epitopes; e. Pre., the predicted position of the binding peptides; f. Pre. Peptide, the predicted core nonamers; g. Pre. Type, the predicted MHC molecules of I-A\textsuperscript{G7} and HLA-DQ\textsubscript{8} that potentially recognize the core 9-mers.

Acknowledgments

The authors are thankful for two anonymous reviewers, whose suggestions have greatly improved the presentation of this manuscript. Pacific Edit reviewed the manuscript prior to submission.

Author Contributions

Conceived and designed the experiments: YX QY RC ZL. Performed the experiments: RC ZL QY YX. Analyzed the data: RC ZL QY YX. Contributed reagents/materials/analysis tools: JR CM TG YZ. Wrote the paper: YX QY RC ZL.

References

1. Fierabracci A (2011) Peptide immunotherapies in Type 1 diabetes: lessons from animal models.Curr Med Chem 18: 577–586.
2. Luo X, Herold KC, Miller SD (2010) Immunotherapy of type 1 diabetes: where are we and where should we be going? Immunol 32: 488–499.
3. Bluestone JA, Herold K, Eisenbarth G (2010) Genetics, pathogenesis and clinical interventions in type 1 diabetes. Nature 464: 1293–1300.
4. Sherr J, Sosenson J, Myler JS, Herold KC (2008) Prevention of type 1 diabetes: the time has come. Nat Clin Pract Endocrinol Metab 4: 334–343.
5. Waldron-Lynch F, Herold KC (2011) Immunomodulatory therapy to preserve pancreatic beta-cell function in type 1 diabetes. Nat Rev Drug Discov 10: 439–452.
6. Panagiotopoulos C, Trudeau JD, Tan R (2004) T-cell epitopes in type 1 diabetes. Curr Diab Rep 4: 67–74.
7. von Herrath M, Sanda S, Herold K (2007) Type 1 diabetes as a relapsing-remitting disease? Nat Rev Immunol 7: 908–914.
8. Lefebvre A, Diana J, Zacccone P, Cooke A (2010) Immunee cell crosstalk in type 1 diabetes. Nat Rev Immunol 10: 501–513.
9. Rezinski B, Eisenstein M, Bockova J, Kenen-Waisman S, Mor F, et al. (1997) Molecular characterization of the diabetes-associated mouse MHC class II protein, I-A\textsuperscript{G7}. International Immunology 9: 43–51.
10. Harrison LG, Honeyman MC, Trembleau S, Gregori S, Gallazzi F, et al. (1997) A peptide-binding motif for I-A\textsuperscript{G7}, the class II major histocompatibility complex (MHC) molecule of NOD and Biozzi AB/H mice. J Exp Med 185: 1013–1021.
11. Suri A, Walters JJ, Gross ML, Unanue ER (2005) Natural peptides selected by the diabetogenic I-A\textsuperscript{G7} molecule are presented by the autologous APCs of I-A\textsuperscript{G7}-transgenic mice. J Immunol 174: 3161–3167.
12. Tong JC, Tan Diab Rep 4: 67–94.
13. von Herrath M, Sanda S, Herold K (2007) Type 1 diabetes as a relapsing-remitting disease? Nat Rev Immunol 7: 908–914.
14. Rezinski B, Eisenstein M, Bockova J, Kenen-Waisman S, Mor F, et al. (1997) Molecular characterization of the diabetes-associated mouse MHC class II protein, I-A\textsuperscript{G7}. International Immunology 9: 43–51.
15. Harrison LG, Honeyman MC, Trembleau S, Gregori S, Gallazzi F, et al. (1997) A peptide-binding motif for I-A\textsuperscript{G7}, the class II major histocompatibility complex (MHC) molecule of NOD and Biozzi AB/H mice. J Exp Med 185: 1013–1021.
16. Suri A, Walters JJ, Gross ML, Unanue ER (2005) Natural peptides selected by the diabetogenic I-A\textsuperscript{G7} molecule are presented by the autologous APCs of I-A\textsuperscript{G7}-transgenic mice. J Immunol 174: 3161–3167.
17. Tong JC, Tan Diab Rep 4: 67–94.
18. von Herrath M, Sanda S, Herold K (2007) Type 1 diabetes as a relapsing-remitting disease? Nat Rev Immunol 7: 908–914.
19. Culina S, Boitard C, Mallone R (2011) Antigen-based immune therapeutics for type 1 diabetes: magic bullets or ordinary blanks? Clin Dev Immunol 2011: 296249.
20. Rajapakse M, Zhang GL, Sinivasan KN, Schmidt B, Petrosky N, et al. (2006) PREDNOD, a prediction server for peptide binding to the H-2d7 haplotype of the non-obese diabetic mouse. Autoimmunity 39: 645–650.
21. Chang KY, Suri A, Unanue ER (2007) Predicting peptides bound to I-A\textsuperscript{G7} class II histocompatibility molecule using a novel expectation-maximization alignment algorithm. Proteomics 7: 367–377.
22. Chang KY, Unanue ER (2009) Prediction of HLA-DQ\beta1beta cell peptidome using a computational program and its relationship to autoreactive T cells. Int Immunol 21: 705–713.
23. Rajapakse M, Schmidt B, Feng L, Bucic V (2007) Predicting peptides binding to MHC class II molecules using multi-objective evolutionary algorithms. BMC Bioinformatics 8: 439.
24. Bhasin M, Raghava GPS (2004) SVM based method for predicting HLA-DRB1*1501 binding peptides in an antigen sequence. Bioinformatics 20: 421–423.
25. Bhasin M, Raghava GPS (2004) SVM based method for predicting HLA-DRB1*1501 binding peptides in an antigen sequence. Bioinformatics 20: 421–423.
26. Bhasin M, Raghava GPS (2004) SVM based method for predicting HLA-DRB1*1501 binding peptides in an antigen sequence. Bioinformatics 20: 421–423.
27. Bhasin M, Raghava GPS (2004) SVM based method for predicting HLA-DRB1*1501 binding peptides in an antigen sequence. Bioinformatics 20: 421–423.
28. Xue Y, Liu Z, Gao X, Ma Q, Ren J, et al. (2011) GPS-CCDT: a novel computational program for the prediction of calpain cleavage sites. PLoS One 6: e19001.
29. Xue Y, Liu Z, Gao X, Jin C, Wen L, et al. (2010) GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm. PLoS One 5: e11290.
30. Lawrence CE, Alsahl SF, Boguski MS, Liu JS, Neunald AF, et al. (1993) Detecting subtle sequence signals: a Gibbss sampling strategy for multiple alignment. Science 262: 208–214.
31. Neunald AF, Liu JS, Lawrence CE (1995) Gibbss motif sampling: detection of bacterial outer membrane protein repeats. Proc Natl Acad Sci USA 92: 3296–3300.
32. Lawrence CE, Alsahl SF, Boguski MS, Liu JS, Neunald AF, et al. (1993) Detecting subtle sequence signals: a Gibbss sampling strategy for multiple alignment. Science 262: 208–214.
33. Neunald AF, Liu JS, Lawrence CE (1995) Gibbss motif sampling: detection of bacterial outer membrane protein repeats. Proc Natl Acad Sci USA 92: 3296–3300.
34. Neunald AF, Liu JS, Lawrence CE (1995) Gibbss motif sampling: detection of bacterial outer membrane protein repeats. Proc Natl Acad Sci USA 92: 3296–3300.
35. Lawrence CE, Alsahl SF, Boguski MS, Liu JS, Neunald AF, et al. (1993) Detecting subtle sequence signals: a Gibbss sampling strategy for multiple alignment. Science 262: 208–214.
36. Lawrence CE, Alsahl SF, Boguski MS, Liu JS, Neunald AF, et al. (1993) Detecting subtle sequence signals: a Gibbss sampling strategy for multiple alignment. Science 262: 208–214.