OXFORD

Briefings in Bioinformatics, 2023, **24(1)**, 1–9

https://doi.org/10.1093/bib/bbac525

Problem Solving Protocol

DMPPred: a tool for identification of antigenic regions responsible for inducing type 1 diabetes mellitus

Nishant Kumar[†], Sumeet Patiyal[†], Shubham Choudhury, Ritu Tomer, Anjali Dhall and Gajendra P. S. Raghava

Corresponding author. Gajendra P. S. Raghava, Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, Okhla Industrial Estate, Phase III, (Near Govind Puri Metro Station), New Delhi 110020, India. Tel.: +1-11-26907444; E-mail: raghava@iiitd.ac.in

[†]These authors contributed equally to this work.

Abstract

There are a number of antigens that induce autoimmune response against β-cells, leading to type 1 diabetes mellitus (T1DM). Recently, several antigen-specific immunotherapies have been developed to treat T1DM. Thus, identification of T1DM associated peptides with antigenic regions or epitopes is important for peptide based-therapeutics (e.g. immunotherapeutic). In this study, for the first time, an attempt has been made to develop a method for predicting, designing, and scanning of T1DM associated peptides with high precision. We analysed 815 T1DM associated peptides and observed that these peptides are not associated with a specific class of HLA alleles. Thus, HLA binder prediction methods are not suitable for predicting T1DM associated peptides. First, we developed a similarity/alignment based method using Basic Local Alignment Search Tool and achieved a high probability of correct hits with poor coverage. Second, we developed an alignment-free method using machine learning techniques and got a maximum AUROC of 0.89 using dipeptide composition. Finally, we developed a hybrid method that combines the strength of both alignment free and alignment-based methods and achieves maximum area under the receiver operating characteristic of 0.95 with Matthew's correlation coefficient of 0.81 on an independent dataset. We developed a web server 'DMPPred' and stand-alone server for predicting, designing and scanning T1DM associated peptides (https://webs.iiitd.edu.in/raghava/dmppred/).

Keywords: diabetes mellitus, type 1 diabetes, β-cells, BLAST, machine learning, web server

Introduction

Diabetes mellitus (DM) is a chronic metabolic disorder majorly occurring due to the abnormality in the blood sugar or glucose level, which further leads to serious damages in heart, eyes, nerves, kidney and blood vessels [1, 2]. According to the WHO (World Health Organization), approximately 422 million people in the world are affected with DM and 1.5 million deaths have been reported every year. DM is majorly categorized into two types: T1DM (type 1 diabetes mellitus) or insulin dependent diabetes and T2DM (type 2 diabetes mellitus) or adult onset diabetes [1, 3]. T2DM primarily affects middle-aged and older persons who suffer from persistent hyperglycemia and affects around 6.28% of the world's population and is also known as a lifestyle disorder [4]. On the other hand, T1DM is an immune-mediated serious lifelong incurable autoimmune disorder that affects around 5–10% of all cases of diabetes and mainly identified in children or adolescents. It is a serious condition in which insulin producing β-cells in the pancreas are destroyed [5, 6]; insulin is a peptide hormone that regulates the blood glucose levels [7, 8]. Yoon *et al.* [9] reported that factors which are involved in the pathogenesis of autoimmune diabetes include macrophages, T-lymphocytes, B-lymphocytes, β-cell autoantigens and dendritic cells.

The pathogenesis of T1DM includes both genetic and environmental factors [10] (Supplementary Figure S1). The genetic factors involved in disease severity are mainly human leukocyte antigen (HLA) class-I/II [11]. Primavera *et al.* [12] reported that presence of specific allele haplotypes like DR4-DQ8 increases the pathogenesis of T1DM. The risk is not only limited to the DR4-DQ8 haplotype but also to some other haplotypes like DR3-DQ2 [12], autoantibodies, autoreactive and anti-islet antigen-specific T-cells, which have been identified in most T1DM patients [13–17]. The HLA alleles encoding molecules involved in presenting antigens to T cells account for 50%–60% of the genetic risk for T1DM [18]. Autoantibodies act as robust predictive and diagnostic biomarkers for the detection of T1DM. For example, insulin autoantibodies and GAD (glutamate decarboxylase)

Nishant Kumar is currently a PhD student in computational biology at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Sumeet Patiyal is currently a PhD student in computational biology at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Shubham Choudhury is currently a PhD student in computational biology at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Ritu Tomer is currently a PhD student in computational biology at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Anjali Dhall is currently a PhD student in computational biology at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Gajendra P. S. Raghava is currently working as a professor and the head of department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Received: July 22, 2022. **Revised:** October 27, 2022. **Accepted:** November 4, 2022

autoantibodies are biomarkers for the early detection of beta cell autoimmunity [13]. Moreover, autoreactive T cells also play a significant role in the destruction of β cells [18]. As shown in Supplementary Figure S1, several diseases are associated with the progression of T1DM such as celiac disease, coronary heart disease, retinopathy, rheumatoid arthritis, autoimmune thyroid diseases, type A gastritis, vitiligo, systemic lupus erythematosus, Addison disease, etc. [19–22]. Identification of T1DM associated peptides is a crucial step in understanding the fate of disease. These peptides mainly bind to major histocompatibility complex (MHC) class II alleles and induce immune responses responsible for destroying β-cells. In the previous studies, limited attempts have been made to predict T1DM associated peptides using *in silico* techniques. Cai *et al*. [23] developed a method GPS-MBA for the prediction of T1DM associated MHC class II binders. The major limitation of this study is that it is developed on a small dataset consisting of only HLA-DQ8 binders [23]. Thus, it is the need of the hour to develop a better and accurate tool for the prediction of T1DM associated peptides using the latest available information. In this study, we have collected experimentally validated T1DM associated peptides from the immune epitope database (IEDB) [24]. Due to the limited availability of negative data, we generated random peptides from Swiss-Prot [25]. We analysed these peptides to understand their characteristics.

Material and methods

Dataset collection

In order to construct a successful prediction model, data acquisition is the primary and the most cumbersome step [26, 27]. The IEDB is a major resource in the field of immunology, as it maintains experimentally validated MHC binders and epitopes [27]. We extracted T1DM associated peptides from IEDB and all redundant peptides were removed. It was observed that most of the peptides have 8 to 30 amino acids, so we removed peptides having length more than 30 amino acids or less than 8 amino acids. Supplementary Figure S2 shows the frequency of length of peptides in our dataset. Finally, we got 815 unique T1DM associated peptides which we labelled as positive data. Due to the limited availability of experimentally validated non-T1DM associated peptides, we generated random peptides from proteins in Swiss-Prot [28, 29]. Finally, we got 815 T1DM associated peptides and 815 random (non-T1DM associated) peptides. In order to evaluate models without any biases, we used 80% of data for building models called training dataset. Remaining 20% data is used for testing our models, this is called independent/validation dataset. Our independent dataset is not used for any training or tuning hyperparameters, but only used for evaluating the final model. On the other hand, our training dataset is used for training and testing to optimize variables of models.

Composition analysis

Amino acid composition (AAC) is the compositional representation of peptide sequences, which represents the percent occurrence frequency of 20 amino acids in the protein/peptide sequence. It generates a 20-dimensional feature vector that specifies the number of each type of amino acid normalized with the total number of amino acids in the length of the provided peptide sequence, and can be calculated by using the following equation: [26, 27]

$$\mathrm{AAC_i} = \frac{R_i}{L} \times 100 \qquad (1)$$

where $\mathrm{AAC_i}$ is the percent composition of an amino acid i; $\mathrm{R_i}$ is the number of residues of type i and L is the total number of residues in the peptide [29].

Sequence logo

For the generation of sequence logos, we have used the software named ‘Weblogo’. It provides the graphical representation in the form of a stack of amino acids. The sequence conservation is measured in bits. The logo generation requires FASTA or CLUSTAL formats. The overall height of each stack indicates the sequence conservation at that position, whereas the height of symbols within the stack reflects the relative frequency of the corresponding amino acid at that position [30].

Feature generation

In order to build an alignment free model, we need to generate features or descriptors corresponding to peptides. As the length of peptides varies from 8 to 30 amino acids, we need to compute composition-based features [26, 31]. In this study, we have computed a wide range of compositional features like AAC, dipeptide composition (DPC) and atomic composition using Pfeature [32].

Feature selection

Another crucial step in classification is feature selection, as most of the features are not significant. It is commonly used to identify the most relevant features. Pfeature generates a large pool of features, and [33] mostly redundant and irrelevant features exist in the original feature set, which can create over-fitting [31]. The selection of the best features can minimize the risk of overfitting and increase efficiency by reducing the model’s complexity [31, 34]. Selecting the relevant set of features from the enormous dimension of features is one of the primary concerns in this study. There are numerous approaches for feature selection [28]; we adopted the mRMR (minimum redundancy maximum relevance) algorithm [34]. We have applied machine learning (ML) algorithms on the selected top 50, 100 and 150 features using the mRMR approach and compute the performance.

ML techniques

We used a number of ML approaches to construct prediction models [29] including decision tree (DT), logistic regression (LR), Gaussian Naive Bayes (GNB), Random Forest (RF), k-nearest neighbor (KNNs), Extra-Trees (ET), XGBoost (XGB) and support vector classification (SVC). The scikit-learn python library was used to implement these classification approaches [28].

Five-fold cross-validation

In order to develop general models that are not biased/overfitted for a dataset, we have implemented a 5-fold cross-validation algorithm [28, 33]. According to standard protocols of 5-fold cross-validation, in each cycle the complete training dataset was divided into five equal sets where the four sets were designated as the training dataset, and the fifth set was assigned as the testing dataset. In the results section, we presented the average scores of five cycles after repeating this approach five times.

Similarity search

One of the commonly used techniques for annotation of proteins/peptides is based on similarity search. In this technique, the query peptide is aligned with all peptides whose function is known. The query peptide is annotated based on its alignment score with known peptides. Basic Local Alignment Search

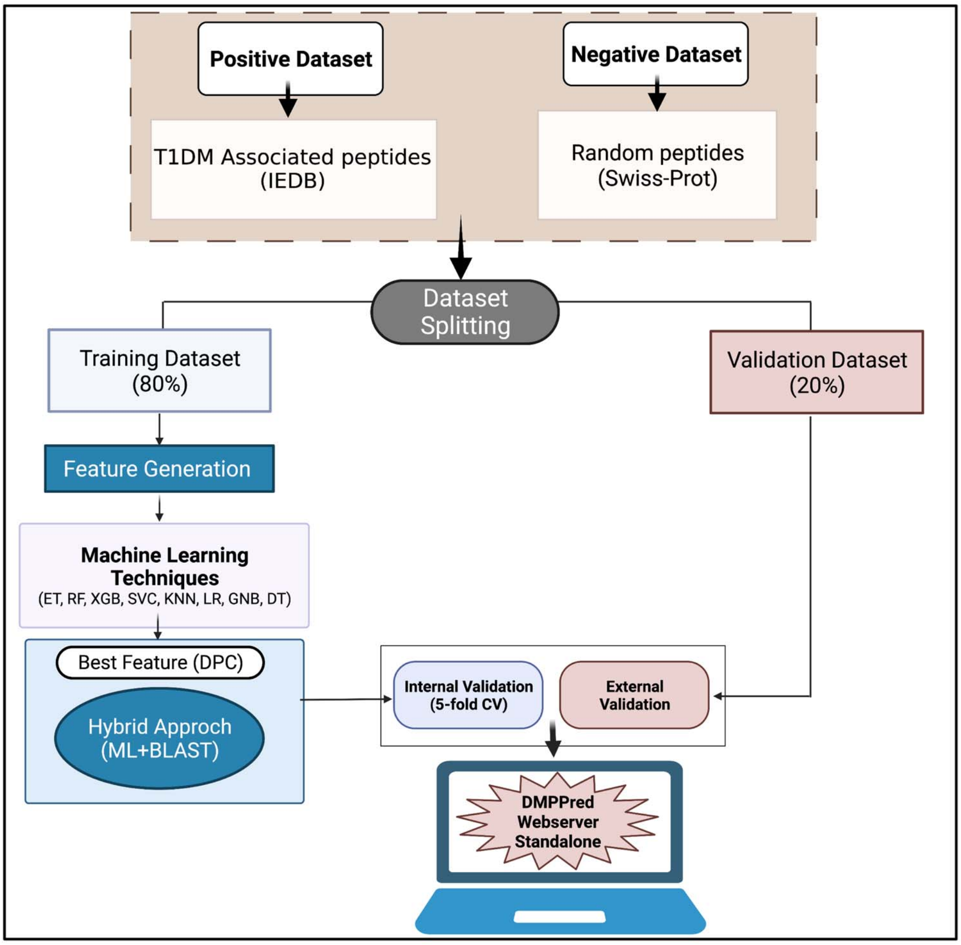

Positive Dataset
Negative Dataset
T1DM Associated peptides (IEDB)
Random peptides (Swiss-Prot)
Dataset Splitting
Training Dataset (80%)
Validation Dataset (20%)
Feature Generation
Machine Learning Techniques (ET, RF, XGB, SVC, KNN, LR, GNB, DT)
Best Feature (DPC)
Hybrid Approch (ML+BLAST)
Internal Validation (5-fold CV)
External Validation
DMPPred Webserver Standalone

Figure 1. Overall architecture of DMPPred including creation of dataset, training, internal and external validation.

Tool (BLAST) is a very popular method for similarity search [35–38]. Currently, we have implemented BLAST-based search for identifying similarity of peptides/epitopes with T1DM associated and non-T1DM associated peptides. We have implemented the blastp-short task of the blastp (BLAST+ 2.7.1) suite with the default parameters (scoring matrix = PAM30; word_size = 2; window_size = 15; gapopen = 9; gapextend = 1; and threshold = 16) [39]. This returns the most similar sequence from the database for the query sequences. Initially, a database was created using the sequences in the training dataset, and query sequences from the independent dataset were queried against it at various e-values ranging from 1e-6 to 1e+4. We have considered the top hits only and assigned the class based on the same, such that, if the top hit is T1DM associated then the query sequence was assigned as T1DM associated and vice versa.

Evaluation parameters

Our study includes the well-established evaluation parameters for the evaluation of the ML models. We include both the parameters, i.e. threshold-dependent parameters and threshold-independent parameters. Sensitivity, specificity, accuracy and Matthew's correlation coefficient (MCC) are threshold dependent parameters, while area under the receiver operating characteristic (AUROC) curve is a threshold-independent parameter [28, 29, 33]. These measurements are calculated using Equations (2)–(6):

$$Sensitivity = \frac{T_P}{T_P + F_N} \tag{2}$$

$$Specificity = \frac{T_N}{T_N + F_P} \tag{3}$$

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{4}$$

$$F1 - Score = \frac{2T_P}{2T_P + F_P + F_N} \tag{5}$$

$$MCC = \frac{(T_P * T_N) - (F_P * F_N)}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}} \tag{6}$$

where T_P, T_N, F_P and F_N stands for true positive, true negative, false positive and false negative, respectively.

Results

In this study, we incorporate 815 T1DM associated peptides and annotate them as a positive dataset. The negative dataset includes random 815 non-T1DM associated peptides generated from proteins in Swiss-Prot. The prediction and analysis was performed on these peptides. Figure 1 represents the overall workflow of the study.

Positional analysis

In this study, we have developed a sequence logo to check the frequency of a particular residue at a specific position in T1DM associated peptides. As depicted in Figure 2, the hydrophobic

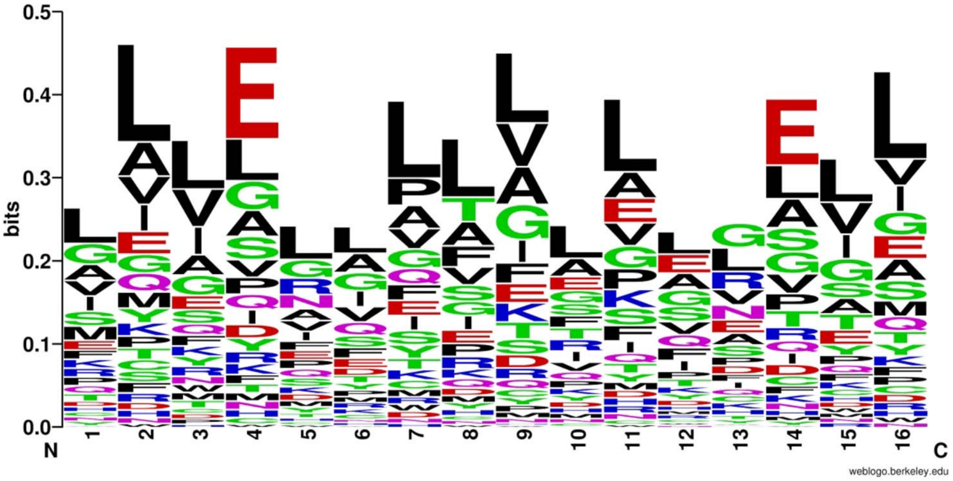
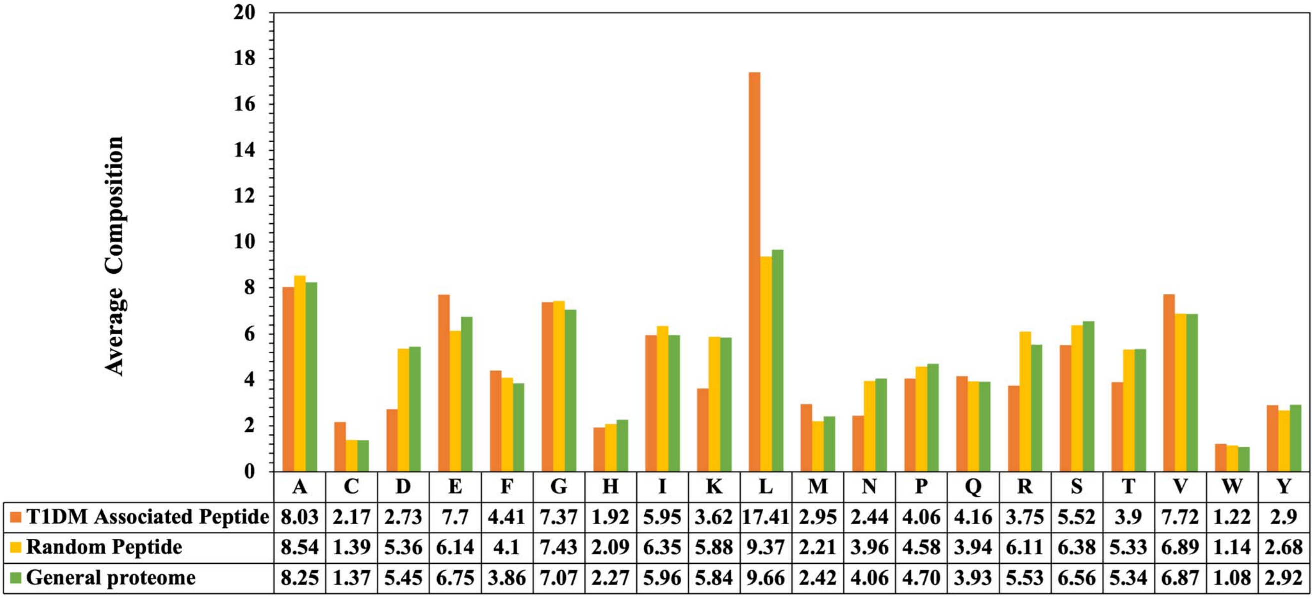


Figure 2. Sequence logo of T1DM associated peptides, leucine residue is dominant in most of the positions.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
■ T1DM Associated Peptide	8.03	2.17	2.73	7.7	4.41	7.37	1.92	5.95	3.62	17.41	2.95	2.44	4.06	4.16	3.75	5.52	3.9	7.72	1.22	2.9
■ Random Peptide	8.54	1.39	5.36	6.14	4.1	7.43	2.09	6.35	5.88	9.37	2.21	3.96	4.58	3.94	6.11	6.38	5.33	6.89	1.14	2.68
■ General proteome	8.25	1.37	5.45	6.75	3.86	7.07	2.27	5.96	5.84	9.66	2.42	4.06	4.70	3.93	5.53	6.56	5.34	6.87	1.08	2.92

Figure 3. Percent average amino acid composition of T1DM associated peptides, random peptides and general proteome.

residue leucine (L) is a highly abundant and conserved residue, whereas alanine (A) is more preserved at 2nd, 9th, 11th and 14th position; however, valine (V) dominates the 2nd, 3rd, 9th and 16th, position. On the other hand, hydrophilic residue like glutamic acid (E) is predominant at 2nd, 4th, 11th and 14th position.

Composition-based analysis

Here, we compute the AAC for T1DM associated and non-T1DM associated peptides, by calculating the average composition of T1DM associated peptides, random peptides and general proteome as shown in Figure 3. The residues leucine (L), methionine (M), valine (V) and glutamic acid (E) are most abundant amino acids in the positive dataset, whereas residues aspartic acid (D), lysine (K) and arginine (R) are highly conserved in the negative dataset.

Frequency of HLA alleles

In the past, several studies report that class II HLA alleles play a key role in the development of T1DM [14]. In order to check the binding efficacy of T1DM associated peptides, we have computed the frequency of HLA alleles with the T1DM associated peptides. In Table 1, we represented the top 10 class I and II HLA binders in the T1DM associated dataset.

Table 1. Frequency distribution of HLA-binders in type 1 diabetes mellitus associated peptides

HLA allele	T1DM associated peptides
HLA-A*02:01	345
HLA-DRB1*04:01	185
HLA-DQA1*03:01/DQB1*03:02	103
HLA-DR4	59
HLA-A2	57
HLA-DQA1*05:01/DQB1*03:02	45
HLA-DR	45
HLA-DR3	39
HLA-DR53	21
HLA-DRB4*01:01	20

Table 2. The performance of machine-learning-based models developed on 12 different composition-based features using ET algorithm

Name and description of descriptor	Training		Validation	
	AUROC	MCC	AUROC	MCC
AAC (amino acid composition)	0.869	0.572	0.870	0.571
DPC (dipeptide composition)	0.893	0.621	0.891	0.620
APAAC (amphiphilic pseudo amino acid composition)	0.869	0.551	0.875	0.571
ATC (atomic composition)	0.710	0.302	0.749	0.356
CETD (composition-enhanced transition distribution)	0.849	0.531	0.864	0.546
DDR (distance distribution of residue)	0.837	0.500	0.797	0.417
PAAC (pseudo amino acid composition)	0.868	0.569	0.869	0.546
PCP (physico-chemical properties composition)	0.840	0.503	0.847	0.497
QSO(quasi-sequence order)	0.858	0.552	0.861	0.565
RRI (residue repeat Information)	0.854	0.54	0.81	0.423
SPC (Shannon entropy of physico-chemical properties)	0.791	0.442	0.812	0.462
CTD (conjoint triad descriptors)	0.856	0.537	0.835	0.473

AUROC, area under receiver operating curve; MCC, Matthew's correlation coefficient.

ML and performance evaluation

Performance of composition-based features

We have used ML algorithms such as DT, RF, LR, GNB, ET, XGB, KNN and SVC to develop prediction models. Initially, we developed prediction models based on 12 different types of features computed using Pfeature. We observe that the ET-based classifier performs best among all other ML models. In Table 2, we incorporate the prediction performance of all 12 descriptors in terms of the AUROC and MCC. As shown in Table 2, DPC-based features outperform other models with an AUROC of 0.893 on training and 0.891 the on validation dataset. Models developed using AAC-based features also perform quite well on both training and validation datasets with an AUROC of 0.869 and 0.870, respectively. Comprehensive results of other classifiers are reported in Supplementary Table S1.

Performance of selected features

As we have obtained maximum performance on DPC features using an ET-based model, hence we used these features in order to improve the classification performance. Here, we have used the mRMR feature selection algorithm to obtain the best set of features which can classify T1DM associated peptides and non-T1DM associated peptides. In terms of AUROC, the top-50, 100 and 150 selected features have reasonable discriminatory power. Figure 4 shows the performance (in terms of AUROC) of different classifiers using different feature sets for training and validation dataset. The detailed results are provided in Supplementary Table S2.

As shown in Table 3, we have provided the performance of the top-150 features selected using the mRMR approach. We observed that ET achieves maximum performance with an AUROC 0.893 and 0.871 on training and validation datasets, respectively, with balanced sensitivity and specificity. RF-based models also achieve quite similar performance with a slight dip in the accuracy and AUROC on the training dataset.

Furthermore, we ranked the selected 150 features on the basis of their dipeptide composition difference between positive dataset and negative dataset. In order to check the significance of the difference, we have applied a two-sample t-test for each dipeptide composition in the positive and negative dataset, and the difference was found to be statistically significant (see Supplementary Table S3). Bases on the analysis, the top-10 dipeptides were found out to be LL, VE, LV, SL, AL, LQ, LE, IL, ER and LA, which were considerably more frequent in the T1DM associated peptides as compared to non-T1DM associated peptides.

Similarity search approach

To further improve our model accuracy, we have designed a similarity search module based on the BLAST similarity search score, as BLAST has been previously used for the annotation and assignment of the function to a protein on the basis of similarity search [35, 40]. We also used the same approach (blastp) for assigning the query peptides as T1DM associated or non-T1DM associated peptides. We created a local database using the training dataset against which the query sequence (test set sequences) were searched at different e-values ranging from $1e^{-6}$ to $1e^{+4}$. We have used the top hit to assign the class to each query sequence, such as, if the query sequence has the top hit against a T1DM-associated peptide then the query peptide is assigned as a T1DM-associated peptide, non-T1DM associated peptide elsewise. As shown in Table 4, probability of correct prediction (chits) ranges from 17.18% to 87.73% on T1DM associated dataset and 0.00% to 61.96% on non-T1DM associated dataset.

Performance of hybrid method

In the current study, we have applied a hybrid/ensemble approach by combining ML prediction and BLAST similarity search score. At first, a given peptide is classified using BLAST at different e-values, in which query peptides in the independent dataset are searched against the local database of sequences in the training dataset. Based on the top hit, we assigned the score of '+0.5' for a correct positive prediction (T1DM associated peptides), '–0.5' for a correct negative prediction (non-T1DM associated peptides) and '0' if hit is not found. This similar approach has been heavily used in previous studies [40, 41]. Secondly, the prediction score was computed using ML-based models. Finally, the BLAST score and ML score were added for each query sequence to get the overall score. Further, the overall score was used to find the optimal threshold at which the difference between the sensitivity and specificity measure is minimum. As shown in Tables 2 and 3, ET-based classifier performs the best using DPC and DPC-150 features. Hence, we generated hybrid models using DPC-150 features (see Table 5). We have calculated performance at different e-values and found that at e-value '0.1' we obtained maximum AUROC of 0.94 and 0.95, with accuracy 88.57 and 90.46 on both training and validation datasets, respectively.

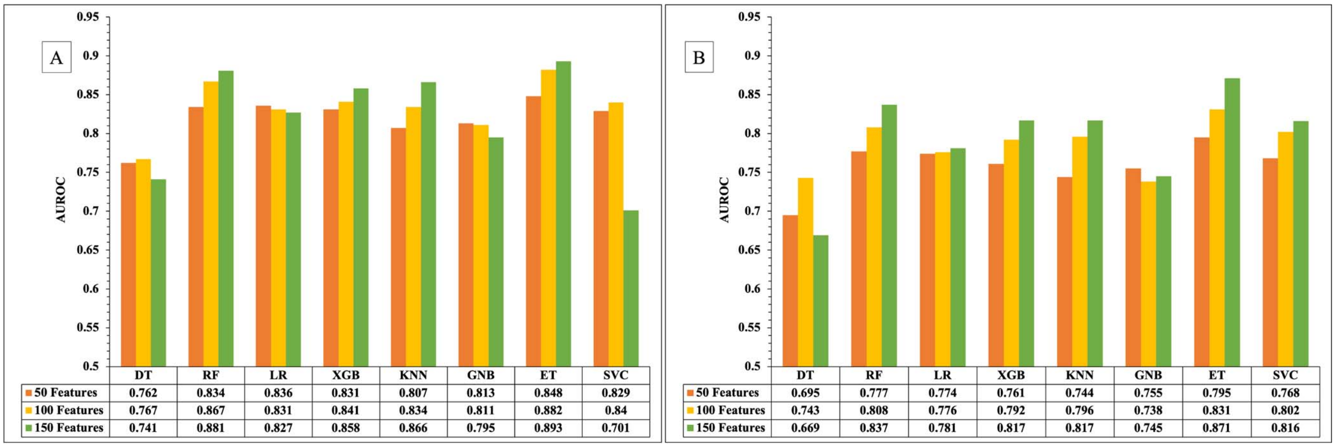


Figure 4. Performance of different classifiers on selected 50, 100 and 150 features: (**A**) training dataset, (**B**) validation dataset.

Table 3. The performance of machine learning-based models developed using DPC-150 selected features on training and validation datasets

Classifier	Training					Validation				
	Sens	Spec	Acc	AUROC	MCC	Sens	Spec	Acc	AUROC	MCC
DT	69.479	69.939	69.709	0.741	0.394	68.098	64.417	66.258	0.669	0.325
RF	79.908	79.755	79.831	0.881	0.597	76.074	74.233	75.153	0.837	0.503
LR	75.460	75.613	75.537	0.827	0.511	71.779	69.325	70.552	0.781	0.411
XGB	77.607	77.914	77.761	0.858	0.555	74.847	74.233	74.540	0.817	0.491
KNN	77.914	78.374	78.144	0.866	0.563	73.006	76.074	74.540	0.817	0.491
GNB	72.853	74.693	73.773	0.795	0.476	68.712	69.325	69.018	0.745	0.380
ET	81.902	81.135	81.518	0.893	0.630	82.209	77.301	79.755	0.871	0.596
SVC	78.374	77.761	78.067	0.865	0.561	73.620	74.233	73.926	0.831	0.479

Sens, sensitivity; Spec, specificity; Acc, accuracy; AUROC, area under receiver operating curve; MCC, Matthew's correlation coefficient; DT, Decision Tree; RF, Random Forest; LR, logistic regression; XGB, XGBoost; KNNs, k-nearest neighbours; GNB, Gaussian Naive Bayes; ET, Extra-Trees Classification; SVC, support vector classification.

Table 4. The performance of BLAST-based search on validation dataset

E-Value	T1DM associated			Non-T1DM associated		
	Chits	Whits	No-hits	Chits	Whits	No hits
1e-6	28 (17.18%)	0 (0.00%)	135 (82.82%)	0 (0.00%)	0 (0.00%)	163 (100.00%)
1e-5	41 (25.15%)	0 (0.00%)	122 (74.85%)	0 (0.00%)	0 (0.00%)	163 (100.00%)
1e-4	54 (33.13%)	0 (0.00%)	109 (66.87%)	0 (0.00%)	0 (0.00%)	163 (100.00%)
1e-3	62 (38.04%)	0 (0.00%)	101 (61.96%)	0 (0.00%)	0 (0.00%)	163 (100.00%)
1e-2	81 (49.69%)	0 (0.00%)	82 (50.31%)	1 (0.61%)	0 (0.00%)	162 (99.39%)
1e-1	91 (55.83%)	0 (0.00%)	72 (44.17%)	1 (0.61%)	0 (0.00%)	162 (99.39%)
1e+0	111 (68.10%)	2 (1.23%)	50 (30.67%)	2 (1.23%)	0 (0.00%)	161 (98.77%)
1e+1	124 (76.07%)	5 (3.07%)	34 (20.86%)	21 (12.88%)	11 (6.75%)	131 (80.37%)
1e+2	143 (87.73%)	19 (11.66%)	1 (0.61%)	90 (55.21%)	46 (28.22%)	27 (16.56%)
2e+2	143 (87.73%)	19 (11.66%)	1 (0.61%)	97 (59.51%)	53 (32.52%)	13 (7.98%)
1e+3	143 (87.73%)	19 (11.66%)	1 (0.61%)	100 (61.35%)	55 (33.74%)	8 (4.91%)
1e+4	143 (87.73%)	19 (11.66%)	1 (0.61%)	101 (61.96%)	56 (34.36%)	6 (3.68%)

Chits: correct hits; Whits: wrong hits.

Web server interface

To better serve the scientific community, we have developed a user-friendly prediction web interface named 'DMPPred (https://webs.iiitd.edu.in/raghava/dmppred/) and executed our best models to predict the T1DM associated peptides. The modules available on the web server include "Predict", "Design", "Protein Scan" and "Blast Scan". The module 'Predict' allows users to classify the submitted sequence as T1DM associated peptides or random peptides. The 'Protein Scan' module allows users to scan or identify T1DM associated regions in the submitted amino-acid sequence. The module 'Design' allows users to create all possible analogs of T1DM associated peptides of the input sequence. The module 'Blast Scan' allows users to search the query sequence against the database of known T1DM associated peptides. A query sequence is predicted as T1DM associated, or random peptide depending upon the match or hit in the database. If found matched or hit in the database predicted as T1DM associated peptide; otherwise it is classified as a random peptide. We also allow users to download the positive and negative dataset that we have used in this study, available in FASTA file format.

Table 5. Model performance developed using hybrid method (BLAST and DPC-150) on training and validation dataset

E-value	Training					Validation				
	Sens	Spec	Acc	AUROC	MCC	Sens	Spec	Acc	AUROC	MCC
1e-6	83.55	83.72	83.64	0.91	0.67	81.58	90.13	85.86	0.93	0.72
1e-5	84.05	84.87	84.46	0.91	0.69	81.58	91.45	86.51	0.93	0.73
1e-4	84.87	84.87	84.87	0.92	0.70	82.24	91.45	86.84	0.93	0.74
1e-3	85.53	85.53	85.53	0.92	0.71	84.87	91.45	88.16	0.94	0.76
1e-2	86.84	86.51	86.68	0.93	0.73	85.53	92.11	88.82	0.94	0.77
1e-1	88.98	88.16	88.57	0.94	0.77	87.50	93.42	90.46	0.95	0.81
1e+0	89.80	89.15	89.47	0.94	0.79	86.84	90.13	88.49	0.94	0.77

Sens, sensitivity; Spec, specificity; Acc, accuracy; AUROC, area under receiver operating characteristics curve; MCC, Matthews correlation coefficient; ET, Extra-Trees Classification.

Table 6. Potential T1DM inducing antigenic peptides predicted by our tool in CBV4 virus polyprotein, selected based on hybrid score

Start	End	Sequence	Hybrid score
1128	1142	EWLKVKILPEVREKH	0.97
1129	1143	WLKVKILPEVREKHE	0.97
1130	1144	LKVKILPEVREKHEF	0.97
1122	1136	KIQKFIEWLKVKILP	0.96
1127	1141	IEWLKVKILPEVREK	0.92
1126	1140	FIEWLKVKILPEVRE	0.89
1125	1139	KFIEWLKVKILPEVR	0.86
1131	1145	KVKILPEVREKHEFL	0.86
1123	1137	IQKFIEWLKVKILPE	0.85
1124	1138	QKFIEWLKVKILPEV	0.85

Case study

Previous studies have shown that there are a number of factors responsible for the progression of diabetes. These studies have reported viruses that promote T1DM including enteroviruses like coxsackievirus B (CVB) [42], cytomegalovirus [43], mumps virus [44] and rotavirus [14, 45–48]. The most commonly found enteroviral strain in diabetic and pre-diabetic patients is CBV4 [49]. We predicted T1DM associated peptides in two proteins of CBV4, polyprotein and VP1 using the 'Protein Scan' module of DMPPred with peptide length 15 residues. In order to minimize false positive prediction, we used a high cut-off score (threshold) 0.70 instead of default cut-off 0.58. Our method predicted 46 T1DM associated peptides in polyprotein which have a score 0.70 or more. It was found that none of the peptides had a score of more than 0.65 in VP1 proteins. It means that the polyprotein is dominated by T1DM associated peptides according to DMPPred. These results are in alignment with existing studies where it has been shown that polyprotein of CBV4 is involved in T1DM [45]. In Table 6, we have shown the top 10 highly potential T1DM associated peptides in the polyprotein. We have provided the complete result in Supplementary Tables S4 and S5.

Discussion and conclusion

The treatment for T1DM is still a long way off [3]. Currently, only insulin treatment is available that helps in the management of T1DM [50]. Another therapy used for treatment purposes is glucagon therapy, which is provided at the time when the concentration of blood glucose decreases [51]. A few non-insulin drugs are also available for type 1 diabetes, such as dipeptidyl peptidase-4 inhibitors, glucagon-like peptide-1 receptor agonists, metformin and sodium-glucose co-transporter-2 (SGLT2) inhibitors [52]. Recently developed antigen-specific immunotherapy for T1DM has higher efficacy and has proven to be safe in clinical outcomes. These antigens are easy to synthesize and can be delivered as proteins, peptides, DNA plasmids or nanoparticles, etc. [53–55]. Of note, it is very crucial to identify T1DM-specific epitopes/peptides for the experimental and biomedical designing. In this study, we have proposed a new method for the prediction of T1DM associated and non-T1DM associated peptides. Here we have selected experimentally validated 815 T1DM associated peptides from IEDB. For the negative dataset, we have generated 815 random peptides of the same length using the Swiss-Prot database. The compositional and positional analysis reveals that amino acid 'L', 'M' and 'E' are highly conserved in the T1DM associated peptides in comparison with the negative dataset. We have used Pfeature to compute composition-based features using a sequence dataset. Here we have computed a total of 1153 features, and they were further reduced using mRMR feature selection technique. We have developed various prediction models using different ML techniques such as DT, RF, LR, XGB, KNN, GNB, ET and SVC. We observed dipeptide composition-based features outperform other models with an AUROC of 0.893 and 0.891 on the training and validation dataset. After that, we develop prediction models on selected features, i.e. DPC-50, 100 and 150 using the mRMR method. DPC-150 features perform best with an AUROC of 0.893 and 0.871 on the training and validation dataset using ET classifier. Finally, the hybrid model was developed by combining BLAST and DPC-150; it achieved the highest AUROC of 0.945 and 951 on both the training and validation dataset. We compared the performance of our method with the existing method GPS-MBA [23]. GPS-MBA takes sequences with the length of 9 amino acids as input; therefore, we selected the peptides of length 9 from our independent dataset (i.e. 50 T1DM associated and 50 non-T1DM associated sequences). We have provided these sequences to GPS-MBA and DMPPred server. We computed the sensitivity, specificity and accuracy, and it was observed that DMPPred outperformed the existing method resulting in the sensitivity of 92%, specificity of 100% and accuracy of 96.0%, whereas GPS-MBA attained sensitivity of 6%, specificity of 90% and accuracy of 48.0%. In addition, we have identified the antigenic regions of CVB4 virus using DMPPred, most importantly EWLKVKILPEVREKH, WLKVKILPEVREKHE, LKVKILPEVREKHEF and KIQKFIEWLKVKILP, that induce TIDM with maximum prediction score. We anticipate that our method has several applications in the field of immunotherapy and vaccine development. As shown in previous studies, antigen-based immunotherapy is designed for specifically targeting T-cell population which can drive disease; therefore, developing antigen specific T-cell tolerance is required to design for therapeutic use [56, 57]. Our method will provide the facility for identification of potential disease causing antigens which can cause β-cell destruction in T1DM. In addition, it is also reported that few environmental factors, such as food, virus and toxins, also favour T1DM progression [45, 49, 58–60]. The peptide regions that are associated with the progression of T1DM can be identified using our tool. In addition, several therapeutic peptides have failed during clinical trials due to the presence of allergic and toxin regions. Our method can predict the antigenic regions in therapeutic proteins/peptides that are associated with T1DM. We anticipate that this method will serve the scientific community working in this era. DMPPred is a highly accurate method for the classification of T1DM associated peptides, as this method utilizes the strengths of two techniques, i.e. alignment-based (BLAST) and alignment-free (ML) to improve the performance. We have provided a user-friendly web-server DMPPred (https://webs.iiitd.edu.in/raghava/dmppred/) for the prediction, scanning and designing of T1DM associated peptides.

BioRxiv doi

https://doi.org/10.1101/2022.07.20.500753.

Key Points

- Prediction of peptides responsible for inducing immune system against β-cells
- Compilation and analysis of type 1 diabetes associated HLA binders
- BLAST-based similarity search against type 1diabetes associated peptides
- Alignment free method using machine learning techniques and composition
- A hybrid method using alignment free and alignment-based approach

Authors' contributions

N.K., S.P. and G.P.S.R. collected and processed the datasets. S.P., N.K. and G.P.S.R. implemented the algorithms and developed the prediction models. A.D., S.P., N.K. and G.P.S.R. analysed the results. S.P. and S.C. created the back-end and front-end user interface of the web server. A.D., N.K., R.T. and G.P.S.R. penned the manuscript. G.P.S.R. conceived and coordinated the project. All authors have read and approved the final manuscript.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

Authors are thankful to the University Grants Commission (UGC), Department of Biotechnology (DBT) and Department of Science and Technology (DST-INSPIRE) for fellowships and financial support, and the Department of Computational Biology, IIITD New Delhi for infrastructure and facilities.

Data availability

All the datasets used in this study are available at the 'DMPPred' web server, https://webs.iiitd.edu.in/raghava/dmppred/dataset.php.

Funding

Department of Biotechnology funding agency Grant Number: BT/PR40158/BTIS/137/24/2021.

References

1. Sapra A, Bhandari P. *Diabetes Mellitus*. Treasure Island (FL): StatPearls, 2022.
2. American DA. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2009;**32**(Suppl 1):S62–7.
3. DiMeglio LA, Evans-Molina C, Oram RA. Type 1 diabetes. *Lancet* 2018;**391**:2449–62.
4. Khan MAB, Hashim MJ, King JK, *et al.* Epidemiology of type 2 diabetes - global burden of disease and forecasted trends. *J Epidemiol Glob Health* 2020;**10**:107–11.
5. Mobasseri M, Shirmohammadi M, Amiri T, *et al.* Prevalence and incidence of type 1 diabetes in the world: a systematic review and meta-analysis. *Health Promot Perspect* 2020;**10**:98–115.
6. Lucier J, Weinstock RS. *Diabetes Mellitus Type 1*. Treasure Island (FL): StatPearls, 2022.
7. Wilcox G. Insulin and insulin resistance. *Clin Biochem Rev* 2005;**26**:19–39.
8. Fu Z, Gilbert ER, Liu D. Regulation of insulin synthesis and secretion and pancreatic Beta-cell dysfunction in diabetes. *Curr Diabetes Rev* 2013;**9**:25–53.
9. Yoon JW, Jun HS. Autoimmune destruction of pancreatic beta cells. *Am J Ther* 2005;**12**:580–91.
10. Forbes JM, Cooper ME. Mechanisms of diabetic complications. *Physiol Rev* 2013;**93**:137–88.
11. Noble JA, Erlich HA. Genetics of type 1 diabetes. *Cold Spring Harb Perspect Med* 2012;**2**:a007732.
12. Primavera M, Giannini C, Chiarelli F. Prediction and prevention of type 1 diabetes. *Front Endocrinol (Lausanne)* 2020;**11**:248.
13. Regnell SE, Lernmark A. Early prediction of autoimmune (type 1) diabetes. *Diabetologia* 2017;**60**:1370–81.
14. Principi N, Berioli MG, Bianchini S, *et al.* Type 1 diabetes and viral infections: What is the relationship? *J Clin Virol* 2017;**96**:26–31.
15. Todd JA. Etiology of type 1 diabetes. *Immunity* 2010;**32**:457–67.
16. Lee KH, Wucherpfennig KW, Wiley DC. Structure of a human insulin peptide-HLA-DQ8 complex and susceptibility to type 1 diabetes. *Nat Immunol* 2001;**2**:501–7.
17. Hu X, Deutsch AJ, Lenz TL, *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat Genet* 2015;**47**:898–905.
18. Pugliese A. Autoreactive T cells in type 1 diabetes. *J Clin Invest* 2017;**127**:2881–91.
19. Kahaly GJ, Hansen MP. Type 1 diabetes associated autoimmunity. *Autoimmun Rev* 2016;**15**:644–8.
20. Kahaly GJ, Frommer L, Schuppan D. Celiac Disease and Glandular Autoimmunity. *Nutrients* 2018;**10**:814.

21. Perros P, McCrimmon RJ, Shaw G, *et al.* Frequency of thyroid dysfunction in diabetic patients: value of annual screening. *Diabet Med* 1995;**12**:622–7.
22. Kordonouri O, Dieterich W, Schuppan D, *et al.* Autoantibodies to tissue transglutaminase are sensitive serological parameters for detecting silent coeliac disease in patients with type 1 diabetes mellitus. *Diabet Med* 2000;**17**:441–4.
23. Cai R, Liu Z, Ren J, *et al.* GPS-MBA: computational analysis of MHC class II epitopes in type 1 diabetes. *PLoS One* 2012;**7**:e33884.
24. Vita R, Mahajan S, Overton JA, *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**:D339–43.
25. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 2000;**28**:45–8.
26. Qian L, Wen Y, Han G. Identification of cancerlectins using support vector machines with fusion of G-gap dipeptide. *Front Genet* 2020;**11**:275.
27. Liang X, Li F, Chen J, *et al.* Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. *Brief Bioinform* 2021;**22**:bbaa312.
28. Dhall A, Patiyal S, Sharma N, *et al.* Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Brief Bioinform* 2021;**22**:936–45.
29. Gupta S, Kapoor P, Chaudhary K, *et al.* In silico approach for predicting toxicity of peptides and proteins. *PLoS One* 2013;**8**:e73957.
30. Crooks GE, Hon G, Chandonia JM, *et al.* WebLogo: a sequence logo generator. *Genome Res* 2004;**14**:1188–90.
31. Yang R, Zhang C, Gao R, *et al.* A novel feature extraction method with feature selection to identify Golgi-resident protein types from imbalanced data. *Int J Mol Sci* 2016;**17**:218.
32. Pande A, Patiyal S, Lathwal A, *et al.* Computing wide range of protein/peptide features from their sequence and structure. *Journal of Computational Biology*. 2022.
33. Jain S, Dhall A, Patiyal S, *et al.* IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides. *Comput Biol Med* 2022;**143**:105297.
34. Radovic M, Ghalwash M, Filipovic N, *et al.* Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics* 2017;**18**:9.
35. Kaur D, Arora C, Raghava GPS. A hybrid model for predicting pattern recognition receptors using evolutionary information. *Front Immunol* 2020;**11**:71.
36. Boratyn GM, Schaffer AA, Agarwala R, *et al.* Domain enhanced lookup time accelerated BLAST. *Biol Direct* 2012;**7**:12.
37. Kumar M, Gromiha MM, Raghava GP. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J Mol Recognit* 2011;**24**:303–13.
38. Singh H, Raghava GP. BLAST-based structural annotation of protein residues using Protein Data Bank. *Biol Direct* 2016;**11**:4.
39. Pearson WR. Selecting the right similarity-scoring matrix. *Curr Protoc Bioinformatics* 2013;**43**: 3.5.1–3.5.9.
40. Sharma N, Patiyal S, Dhall A, *et al.* AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Brief Bioinform* 2021;**22**:bbaa294.
41. Saha S, Raghava GP. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res* 2006;**34**:W202–9.
42. Hyoty H, Taylor KW. The role of viruses in human diabetes. *Diabetologia* 2002;**45**:1353–61.
43. Pak CY, Eun HM, McArthur RG, *et al.* Association of cytomegalovirus infection with autoimmune type 1 diabetes. *Lancet* 1988;**2**:1–4.
44. Hyoty H, Leinikki P, Reunanen A, *et al.* Mumps infections in the etiology of type 1 (insulin-dependent) diabetes. *Diabetes Res* 1988;**9**:111–6.
45. Esposito S, Toni G, Tascini G, *et al.* Environmental factors associated with type 1 diabetes. *Front Endocrinol (Lausanne)* 2019;**10**: 592.
46. Honeyman MC, Coulson BS, Stone NL, *et al.* Association between rotavirus infection and pancreatic islet autoimmunity in children at risk of developing type 1 diabetes. *Diabetes* 2000;**49**: 1319–24.
47. Honeyman MC, Stone NL, Harrison LC. T-cell epitopes in type 1 diabetes autoantigen tyrosine phosphatase IA-2: potential for mimicry with rotavirus and other environmental agents. *Mol Med* 1998;**4**:231–9.
48. Hu YF, Du J, Zhao R, *et al.* Complete genome sequence of a recombinant coxsackievirus B4 from a patient with a fatal case of hand, foot, and mouth disease in Guangxi, China. *J Virol* 2012;**86**:10901–2.
49. Filippi CM, von Herrath MG. Viral trigger for type 1 diabetes: pros and cons. *Diabetes* 2008;**57**:2863–71.
50. Subramanian S, Baidal D. The management of type 1 diabetes. In: Feingold KR, Anawalt B, Boyce A, et al. (eds). *Endotext*. South Dartmouth (MA): MDText.com, Inc., 2000.
51. Haymond MW, Liu J, Bispham J, *et al.* Use of glucagon in patients with type 1 diabetes. *Clin Diabetes* 2019;**37**:162–6.
52. Lyons SK, Hermann JM, Miller KM, *et al.* Use of adjuvant pharmacotherapy in type 1 diabetes: International comparison of 49,996 individuals in the prospective diabetes follow-up and T1D Exchange Registries. *Diabetes Care* 2017;**40**: e139–40.
53. Mallone R, You S. The SAgA of antigen-specific immunotherapy for type 1 diabetes. *Diabetes* 2021;**70**:1247–9.
54. Peakman M, von Herrath M. Antigen-specific immunotherapy for type 1 diabetes: maximizing the potential. *Diabetes* 2010;**59**: 2087–93.
55. Kreiner FF, von Scholten BJ, Coppieters K, *et al.* Current state of antigen-specific immunotherapy for type 1 diabetes. *Curr Opin Endocrinol Diabetes Obes* 2021;**28**:411–8.
56. MacLeod MK, Anderton SM. Antigen-based immunotherapy (AIT) for autoimmune and allergic disease. *Curr Opin Pharmacol* 2015;**23**:11–6.
57. Harrison LC, Wentworth JM, Zhang Y, *et al.* Antigen-based vaccination and prevention of type 1 diabetes. *Curr Diab Rep* 2013;**13**: 616–23.
58. Virtanen SM. Dietary factors in the development of type 1 diabetes. *Pediatr Diabetes* 2016;**17**(Suppl 22):49–55.
59. Myers MA, Mackay IR, Rowley MJ, *et al.* Dietary microbial toxins and type 1 diabetes–a new meaning for seed and soil. *Diabetologia* 2001;**44**:1199–200.
60. Serena G, Camhi S, Sturgeon C, *et al.* The role of gluten in celiac disease and type 1 diabetes. *Nutrients* 2015;**7**:7143–62.