Prediction of celiac disease associated epitopes and motifs in a protein

Ritu Tomer[#], Sumeet Patiyal[#], Anjali Dhall[#], Gajendra P. S. Raghava[*]

Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Phase 3, New Delhi-110020, India.

Mailing Address of Authors

Ritu Tomer (RT) : ritut@iiitd.ac.in ORCID ID: https://orcid.org/0000-0002-6171-8660

Sumeet Patiyal (SP): sumeetp@iiitd.ac.in ORCID ID: https://orcid.org/0000-0003-1358-292X

Anjali Dhall (AD): anjalid@iiitd.ac.in ORCID ID: https://orcid.org/0000-0002-0400-2084

Gajendra P. S. Raghava (GPSR): raghava@iiitd.ac.in ORCID ID: https://orcid.org/0000-0002-8902-2876

 # Equal Contribution

***Corresponding Author**

Prof. Gajendra P. S. Raghava

Head and Professor

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

Okhla Industrial Estate, Phase III, (Near Govind Puri Metro Station)

New Delhi, India – 110020 Office: A-302 (R&D Block)

Phone: 011-26907444

Email: raghava@iiitd.ac.in

Website: http://webs.iiitd.edu.in/raghava/

Abstract

Celiac disease (CD) is an autoimmune gastrointestinal disorder which causes immune-mediated enteropathy against gluten. The gluten immunogenic peptides have the potential to trigger immune responses which leads to damage the small intestine. HLA-DQ2 and HLA-DQ8 are major alleles that bind to epitope/antigenic region of gluten and induce celiac disease. There is a need to identify CD associated epitopes in protein-based foods and therapeutics. In addition, prediction of CD associated epitope/peptide is also required for developing antigen-based immunotherapy against celiac disease. In this study, computational tools have been developed to predict CD associated epitopes and motifs. Dataset used in this study for training, testing and evaluation contain experimentally validated CD associated and non-CD associate peptides. Our analysis support existing hypothesis that proline (P) and glutamine (Q) are highly abundant in CD associated peptides. A model based on density of P&Q in peptides has been developed for predicting CD associated which achieve maximum AUROC 0.98. We discovered CD associated motifs (e.g., QPF, QPQ, PYP) which occurs specifically in CD associated peptides. We also developed machine learning based models using peptide composition and achieved maximum AUROC 0.99. Finally, we developed ensemble method that combines motif-based approach and machine learning based models. The ensemble model-predict CD associated motifs with 100% accuracy on an independent dataset, not used for training. Finally, the best models and motifs has been integrated in a web server and standalone software package “CDpred”. We hope this server anticipate the scientific community for the prediction, designing and scanning of CD associated peptides as well as CD associated motifs in a protein/peptide sequence (https://webs.iiitd.edu.in/raghava/cdpred/).

Keywords

Celiac disease, Gluten immunogenic peptides, HLA-DQ2/DQ8, Ensemble method, Motif

Key Points

- Celiac disease is one of the prominent autoimmune diseases
- Gluten immunogenic peptides are responsible for celiac disease
- Mapping of celiac disease associated epitopes and motifs on a proteins
- Identification of proline and glutamine rich regions

- A web server and software package for predicting CD associate peptides

Author's Biography

1. Ritu Tomer is currently working as Ph.D. in Computational Biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
2. Sumeet Patiyal is currently working as Ph.D. in Computational biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
3. Anjali Dhall is currently working as Ph.D. in Computational Biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
4. Gajendra P. S. Raghava is currently working as Professor and Head of Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

Abbreviations

- CD – Celiac Disease
- HLA – Human leukocyte antigens
- CXCR3 – Chemokine receptor 3
- tTG – Tissue transglutaminase
- sIgA – Secretory Immunoglobulin A
- IEDB – Immune Epitope Database
- AUROC – Area under receiver operator curve
- DT – Decision Tree
- RF – Random Forest
- SVC – Support Vector Classifier
- XGB – XGBoost
- LR – Logistic Regression
- ET – Extra Tree classifier
- KNN – k-Nearest Neighbors
- GNB – Gaussian Naive Bayes

Introduction

Celiac disease (CD) is an auto-immunological disorder which mainly affects the small intestine of the infected person [1]. CD is a life-long disorder occurred due to the gluten associated foods which is found in various foods such as wheat, barley, spelt, kamut, and rye [2]. The prevalence rate of CD is around 1.4% worldwide and it may vary with genetic and environmental factors. The occurrence of disease is significantly higher in children in comparison to adults [3]. The presence of human leukocyte antigens (HLAs) play a crucial role in the induction and regulation of immunological responses [4]. The gluten immunogenic peptides bind with specific MHC class II binders i.e., HLA-DQ2/HLA-DQ8 in lamina propria region and activate both innate and adaptive immune system [5]. Studies shows that HLA-DQ2 found in almost 94.5% of CD cases while HLA-DQ8 present in 2.7% of the cases [6]. These binders are also associated with other autoimmunological disorders such as

HLA-DQ8 associated with Type I diabetes [7] while HLA-DQ6 and DQ8 are associated with Multiple sclerosis [8].

The entry of gluten inside the lamina propria region of small intestine follows majorly two pathways, transcellular pathway and paracellular pathway [9], depicted in Figure 1. In transcellular pathway, the entry of gluten is associated with the binding of secretory IgA (sIgA) in the apical region of intestine [10]. While in the paracellular pathway, the entry of gluten is associated with the binding of chemokine receptor 3 (CXCR3) present at enterocyte with the release of zonulin protein [11,12]. After entering inside the lamina propria region, a series of events occurs to an inflammatory cascade which leads to damaging the intestinal villi by the excessive release of antibodies (anti-tissue transglutaminase (tTG2), anti-IgA antibodies and anti-endomysial antibodies) and cytokine [13].

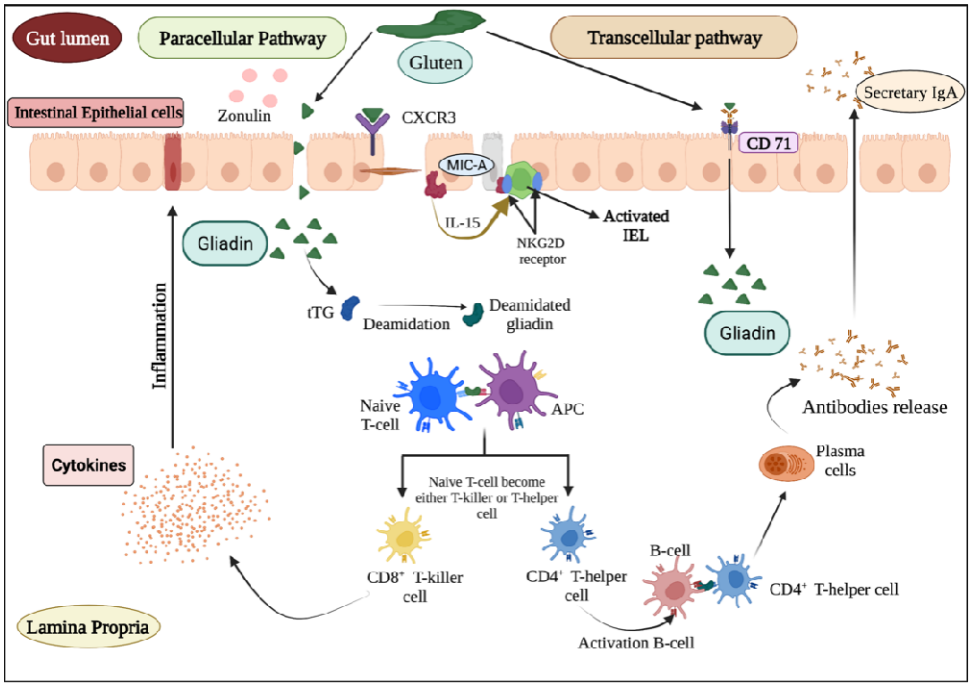

Figure 1: Schematic representation of celiac disease pathogenesis and immune response

Due to auto-inflammatory immune responses several gastrointestinal disorders like malabsorption, vomiting, bloating, diarrhoea, abdominal pain and distension occurred [14]. Recently, a number of biological and genetic tests (such as detection of antibodies, intestinal tissue biopsy, HLA-typing and gluten challenge test) are available for the disease detection [1]. It has been found in many studies that α-gliadin 33-mer peptide having the

property of resistant to gastrointestinal cleavage and makes it highly immunogenic peptide [15–18]. Despite tremendous understanding of CD, effective treatment for the disease is life-long gluten free diet. In order to manage severity of CD effectively, it is important to identify CD associated epitopes or immunogenic peptides responsible for CD. Identification of CD associated epitopes/peptides is not only important for identifying CD free food/therapeutic proteins, it is also important for designing antigen-based immunotherapy against CD.

In the pilot study, we have developed a computational approach for the prediction of CD associated peptides. We have extracted the experimentally validated gluten immunogenic peptides responsible for CD from the IEDB database. In order to create negative dataset, we have collected CD non-causing peptides and random peptides from IEDB and Swiss-Prot, respectively. We have identified highly conserved regions of disease-causing peptides using motif-based search. In addition, we have developed prediction models using composition-based features and machine learning algorithms. In order to facilitate the community, we have provided the webserver and standalone package for the prediction and scanning of CD causing protein/peptides using sequence information.

Material & Methods

The complete architecture of our study is illustrated in Figure 2. The detail of each step is described below.

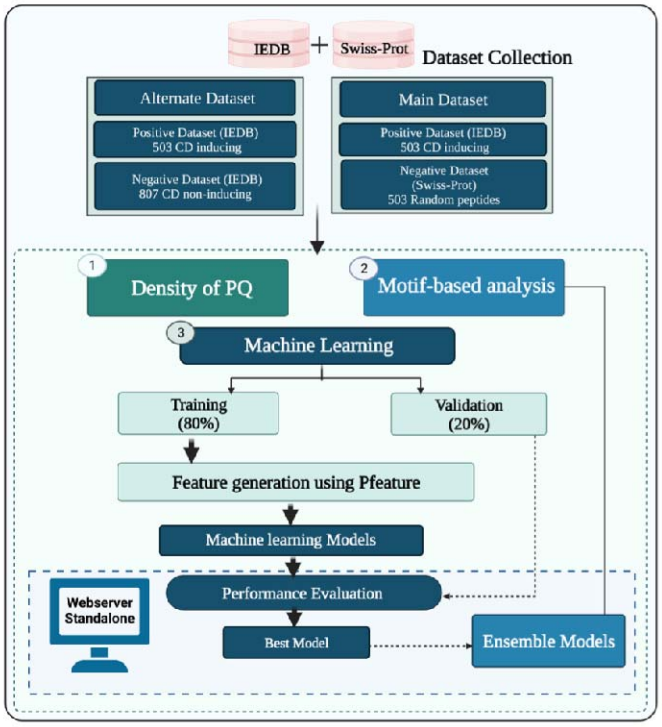

Figure 2: Overall architecture of the study

Dataset collection and pre-processing

In this study, we have collected experimentally validated peptides from the immune epitope database (IEDB) [19]. At first, we extracted a total of 521 unique CD causing/associated (gluten immunogenic peptides) from IEDB as a positive dataset. Further, we have selected unique peptides with a length of 9-20 amino-acid residues and got 503 CD associated peptides. Secondly, we extracted experimentally validated CD non-associated peptides from IEDB and random peptides from Swiss-Prot database [20]. The main dataset incorporates 503 CD associated called positive peptides and 503 random peptides called negative peptides. The alternate dataset consists of 503 CD associated and 807 non-associated peptides (which can cause autoimmune disorders other than celiac disease). Finally, we obtained two datasets, i.e., the main dataset comprises an equal number of positive and negative peptides and alternate dataset 503 positive and 807 negative peptides.

Sequence Logo

In order to understand the preference of amino-acid residues at a specific position, we have generated a one sample logo using WebLogo software [21]. This tool needs a fixed length

input sequence vector. Since, the minimum length of peptides in our datasets is 9 residues, so we have extracted 9-mers from N-terminal and 9-mers from C-terminal from each peptide. After that, we re-join both the regions in order to create a fixed length vector of 18 amino acids. The sequences of 18-residues were generated for all the peptides of both positive and negative datasets and used for the creation of one sample logo plots.

Amino-acid Composition

We have used Pfeature software [22] for the computation of composition-based features. In this current study, we have computed amino acid composition based (AAC) features. In the case of AAC, the composition of each residue is computed in the peptide sequences and a vector od 20 length is generated using the equation 1.

$$AAC_i = \frac{R_i}{L} \times 100 \qquad [1]$$

Where, AAC_i is amino-acid composition of residue type i, R_i is the number of residues in i, and L is the length of peptide sequence.

Machine Learning Models

We have employed a number of machine learning algorithms for the classification of CD-causing peptides. Currently, we have used scikit-learn [23] python library for the implementation of several classifiers including Decision Tree (DT), Random Forest (RF), XGBoost (XGB), Gaussian Naïve Bayes (GNB) Logistic Regression (LR), ExtraTree classifier (ET), and k-nearest neighbors (KNN).

Five-fold Cross Validation

In order to avoid overfitting we have train, test and validate the machine learning models by employing five-fold cross validation technique as implemented in previous studied [4,24–27]. At first, the complete dataset was divided into 80:20 ratio, where 80% dataset used for the training and 20% used for the external validation [28–30]. The five-fold cross-validation process is implemented on the 80% training dataset. In this process, the entire training dataset was divided into five equal sets, where each set is used for training and validation purpose. At first, four sets were used for training and fifth set was used for the testing, similarly the process is repeated five times so that each set can be used as testing dataset. Finally, we calculated the average performance of five sets which resulted after five iterations.

Model Evaluation

In this study, we have used standard parameters for the evaluation of prediction models. Here, we have calculated both threshold dependent as well as independent parameters as previously used in various studies [25]. In the case of threshold-dependent parameters we have computed, sensitivity (Sens), specificity (Spec), accuracy (Acc) and Matthews correlation coefficient (MCC) using the following equations (1-4). In addition, we have measured the performance of models with a well-established and threshold-independent parameter Area Under the Receiver Operating Characteristic (AUROC) curve.

$$Sensitivity = \frac{T_P}{T_P + F_N} \quad [2]$$

$$Specificity = \frac{T_N}{T_N + F_P} \quad [3]$$

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad [4]$$

$$F1 - Score = \frac{2T_P}{2T_P + F_P + F_N} \quad [5]$$

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

Where, T_P, T_N, F_P and F_N stand for true positive, true negative, false positive and false negative, respectively.

Ensemble method

The ensemble method is a hybrid approach in which both motifs based, and machine learning methods combined to achieve better performance. In this method, first motif-based approach is used to identify the disease-causing peptides and then we use machine learning methods to predict those peptides which are not covered by the motif-based approach. Finally, we generate an ensemble method which is a combination of both motif-based approach and machine learning method.

Web Implementation

We have developed a webserver named “CDpred” for the prediction of CD associated peptides. The webserver is implemented by HTML5, JAVA, CSS3 and PHP scripts and compatible on several devices such as iMac, desktop, tablet and mobile. The webserver provides five user-friendly modules such as predict, PQ density, motif scan, protein scan, and design.

Results

Positional Conservation Analysis

The specific position of a residue is important for specific role and structure arrangement of a particular peptide or protein. To identify the most significant position of an amino acid residue in the peptide, we perform the positional analysis of CD causing peptides and CD non-causing peptides by using WebLogo (See Figure3). It is worth noting that the first nine locations correspond to peptide N-terminal residues, whereas the latter nine positions correspond to peptide C-terminus. Here, we found that the proline (P) and glutamine (Q) residues are highly prominent at every position while the Phenylalanine (F) and glutamic acid (E) are also found at some positions.

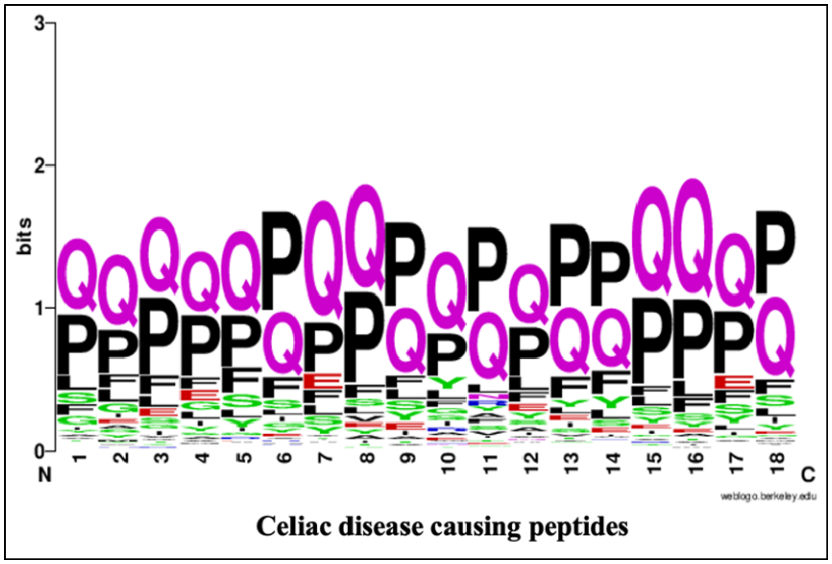

Figure 3: WebLogo of celiac disease-causing peptides

Composition Analysis

We compute the amino acid composition for main and alternate datasets. Figure 4 depicts the average composition of CD inducing and non-inducing peptides. In CD causing peptides, the average composition of Proline (P), Glutamine (Q) and Phenylalanine (F) is higher in comparison with disease non-causing peptides, negative random and general proteome.

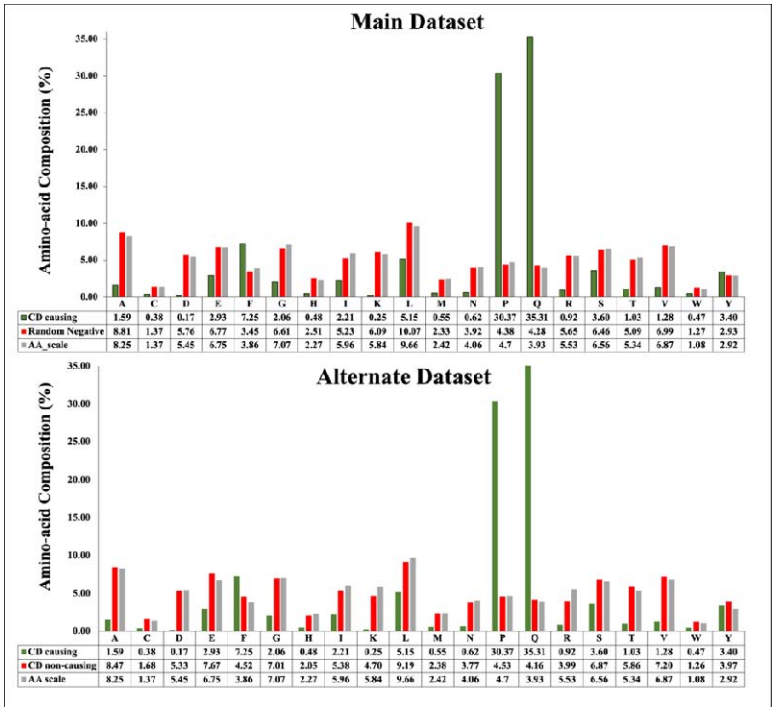

Figure 4: Average amino acid composition of peptides in main dataset and alternate dataset

Frequency of HLA alleles

In the past, a number of studies report that celiac disease occurred due to the presence of certain HLA molecules such as HLA-DQ2 and HLA-DQ8 [31]. We have also observed that the maximum gluten immunogenic peptide binders are associated with HLA-DQ2/DQ8 alleles as depicted in Table 1. The complete frequency distribution of HLA-alleles binders of CD causing and non-causing peptides are given in Supplementary Table S1.

Table 1: Distribution of HLA alleles in CD causing and non-causing peptides

HLA		CD causing (Positive)	CD non-causing (Negative)
HLA-class I	HLA-A	13	0
HLA-class II	HLA-DQ2	263	148
	HLA-DQ8	18	110
	HLA-DQ2/DQ8	24	9
	HLA-DR	3	402
	Other	182	138
	Total	503	807

Motif-based Analysis

Motifs are known as the specific regions of a protein sequence which helps to identify the amino acid arrangement shared by a family of protein. The motifs are identified in the CD causing peptide sequences by MERCI program. The MERCI program helps to identify the motif regions in a set of sequences. We utilized the MERCI tool to look for motifs seen only in CD-causing peptides and not in disease non-causing or random peptides. We also looked for motifs found only in disease non-causing and random peptides. Here, we found 50 motifs in CD causing peptides of different length in which P and Q residues are present in abundance in CD causing peptides. We also checked the common motifs found in disease causing, non-causing and random negative peptides. The list of motifs and their occurrence in all the three datasets are given in Table 2.

Table 2: Abundance of motifs in CD causing, non-causing and random negative peptides

Motifs	CD causing	CD non-causing	Random Negative
QPF	276	4	0
QQPF	170	1	0
PYP	120	3	0
PEQ	56	4	0
QPQ	350	0	1
PQPQ	189	0	1
QQPQ	131	0	1
PQL	84	0	1

PQ Density

On performing the compositional and motif analysis, it was found that (P) and (Q) are the most abundant residues in CD-causing peptides as compared to non-causing peptides. In order to classify the peptides based on the PQ density, we have first generated the overlapping patterns of window size ranging from 3 to 9 for each peptide, since 9 was the minimum length of the peptides, and calculated the composition of residues P and Q in each pattern. Each peptide in the dataset is represented by the maximum value of composition for the respective pattern size and found the optimal composition at which we can classify the peptides with balanced sensitivity and specificity. To find the optimal pattern size, we have

varied the size from 3 to 9, and found out that window size 5 and 6 performed best among the other sizes for main and alternate datasets, respectively as shown in Table 3.

Table 3: Performance of the PQ Density based method on main and alternate datasets

Main Dataset					
Window size	**Threshold**	**Sensitivity**	**Specificity**	**Accuracy**	**AUROC**
3	0.67	85.686	98.807	92.247	0.971
4	0.51	91.65	96.62	94.135	0.977
5	0.41	93.837	94.235	94.036	0.978
6	0.34	95.427	92.445	93.936	0.978
7	0.29	96.421	91.65	94.036	0.979
8	0.38	93.241	97.018	95.129	0.981
9	0.34	94.235	96.62	95.427	0.981
Alternate Dataset					
Window size	**Threshold**	**Sensitivity**	**Specificity**	**Accuracy**	**AUROC**
3	0.67	85.686	98.761	93.74	0.97
4	0.51	91.65	97.77	95.42	0.977
5	0.41	93.837	96.159	95.267	0.979
6	0.34	95.427	94.796	95.038	0.98
7	0.29	96.421	94.3	95.115	0.981
8	0.26	97.018	92.937	94.504	0.983
9	0.34	94.235	98.017	96.565	0.982

Machine learning based prediction

Various machine learning classifiers such as RF, DT, GNB, XGB, KNN, ETN, SVCN and LR are used to develop a prediction model. For this, we have computed the features of disease causing and disease non-causing peptides using composition-based module of Pfeature.

Performance of AAC based features

Firstly, we have computed features of amino acid composition, using which we applied different machine learning techniques. As shown in Table 4, ET achieves maximum

performance in comparison to other models with AUROC 0.991 and 0.995 and accuracy 96.02 and 97.03 on both training and validation dataset with a good balance of sensitivity and specificity in main data. Similarly, ET achieves maximum performance in comparison to other models with AUROC 0.995 and 0.999 and accuracy 97.519 and 98.092 on both training and validation dataset with a good balance of sensitivity and specificity in alternate data.

Table 4: The performance of machine learning classifiers on AAC based features for main and alternate datasets

	Main dataset							
	Training				**Validation**			
Classifier	**Sens**	**Spec**	**Acc**	**AUROC**	**Sens**	**Spec**	**Acc**	**AUROC**
DT	92.269	92.556	92.413	0.962	97.059	91	94.059	0.982
RF	95.262	95.533	95.398	0.989	98.039	97	97.525	0.994
LR	96.01	96.03	96.02	0.988	98.039	96	97.03	0.99
XGB	95.761	95.782	95.771	0.987	98.039	93	95.545	0.995
KNN	95.262	95.285	95.274	0.986	97.059	96	96.535	0.991
GNB	93.017	98.263	95.647	0.976	93.137	98	95.545	0.99
ET	96.01	96.03	96.02	0.991	98.039	96	97.03	0.995
SVC	95.761	95.782	95.771	0.987	97.059	96	96.535	0.991
	Alternate dataset							
	Training				**Validation**			
Classifier	**Sens**	**Spec**	**Acc**	**AUROC**	**Sens**	**Spec**	**Acc**	**AUROC**
DT	92.537	92.57	92.557	0.968	94.059	99.379	97.328	0.99
RF	97.015	97.368	97.233	0.995	98.02	97.516	97.71	0.998
LR	96.269	96.285	96.279	0.99	97.03	96.273	96.565	0.987
XGB	97.015	97.059	97.042	0.992	99.01	93.168	95.42	0.998
KNN	95.771	95.975	95.897	0.992	98.02	95.652	96.565	0.995
GNB	92.537	97.059	95.324	0.977	96.04	96.273	96.183	0.983
ET	97.512	97.523	97.519	0.995	98.02	98.137	98.092	0.999
SVC	97.015	96.904	96.947	0.993	98.02	96.894	97.328	0.996

DT: Decision tress; RF: Random Forest; LR: Logistic regression; XGB: XGBoost; KNN: k-nearest neighbour; GNB: Gaussian naïve base; ET: Extra tree classifier; SVC: support vector classifier; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy

Performance of Ensemble model

In ensemble method, first we used the motif-based approach by identifying the coverage of motifs in the given protein/peptide sequences. Our motif-based approach achieves 81.71% accuracy in the independent dataset. The rest sequences, which were not predicted using

motif-based approach, were covered by using the machine learning method. By combining both approaches we achieve the highest performance of 100% accuracy on independent dataset. Our ensemble method is the best approach for predicting the CD associated peptides.

Table 5: The table shows the occurrence of motif in positive sequences with percentage of correct prediction and cumulative coverage

Motif/ML	Occurrence	Percentage of correct prediction	Cumulative coverage
QPF	276	54.87	54.87
PQQP	41	8.15	63.02
PYP	33	6.56	69.58
QPQQ	28	5.57	75.15
PFP	14	2.78	77.93
PEQ	12	2.39	80.32
FPQP	4	0.8	81.11
FPQQ	2	0.4	81.51
PQLP	1	0.2	81.71
ML Prediction	92	18.29	100

Services to Scientific Community

We design a user-friendly prediction web server that incorporates several modules to determine CD-causing peptides in order to serve the scientific community. The prediction models which

used in the study are implemented in the web server. Based on the prediction models' score at a different threshold, users can predict whether a query peptide causes CD or not. The web server comprises five major modules 1) Prediction, 2) PQ Density, 3), Motif 4) Scan and 5) Design. The user can classify CD-causing peptides from disease non-causing peptides using the 'Predict' module. The "PQ Density" module used to calculate PQ content in a given query sequence based upon the window size. Users can map or scan CD-causing motifs in the query sequence using the "Motif" module. We used the MERCI software to extract themes from CD-causing peptides that had been empirically confirmed. The "Scan" module was used to scan the amino-acid sequence for CD-causing areas. The user can generate all potential analogs of the input sequence using the "Design" module. The positive and negative datasets utilized in this work are also available for download, and the peptide sequence are available

in FASTA format. HTML, JAVA, and PHP scripts were used to create the web server CDpred https://webs.iiitd.edu.in/raghava/cdpred/. The server is user-friendly and compatible with a variety of devices, including computers, Android phones, iPhones, and iPads. In addition, we provided a standalone package in the form of a Docker container.

Discussion & Conclusion

Celiac disease is a gluten-sensitive enteropathy caused due to the chronic, genetically predisposed, and autoimmune condition with a wide spectrum of clinical manifestations brought on by consuming gluten [32]. CD mainly effect the immune system of the patients and causes a number of diseases such as cirrhosis, autoimmune hepatitis, diabetes mellitus, gluten ataxia, peripheral neuropathies, etc [33,34]. During disease condition, enormous amount of CD4+ T cell response act against gluten peptides which are presented by HLA-DQ molecules [35]. In such situation, our immune system secretes excess amount of cytokines and gluten protein specific antibodies. The only effective life-long treatment of this disease is gluten-free diet. Due to increased number of cases in worldwide a number of gluten-free products are available for celiac susceptible people [17,36]. Thus, it is essential to identify or eliminate gluten immunogenic peptides from the food products which can induce the celiac disease and sensitive to celiac patients.

In this study, we have made a systematic attempt for the prediction of peptides responsible for causing the disease. We have collected the dataset from IEDB and Swiss-Prot databases. We have created two datasets for the analysis and prediction of CD causing peptides. Here, we observed that residues (P and Q) are highly abundant in CD causing peptides in comparison with negative and random peptides. The similar findings are supported by the previous studies where they have shown the abundance of P and Q amino acids in gluten proteins [37,38]. From the motif-based approach we identified motifs [QPQ, QPF, PQPQ, QQPF, QQPQ, PYP], which are highly conserved in CD causing peptides in comparison with negative dataset. So, we performed PQ based analysis where we calculate the abundance of PQ residues in the CD causing and non-causing peptides.

In addition, we have developed prediction models using amino-acid composition-based features. We achieved maximum performance with AUROC of 0.99 on the training and validation datasets, respectively. We have also developed an ensemble method by combining both motif-based approach and machine learning based models. This ensemble approach provides the 100% accuracy on independent dataset. In addition, we have developed a

webserver named CDpred (https://webs.iiitd.edu.in/raghava/cdpred/), standalone package (https://webs.iiitd.edu.in/raghava/cdpred/standalone.php) and GitLab (https://gitlab.com/raghavalab/cdpred) for the prediction of CD causing peptides.

Funding Source

The current work has received grant from the Department of Bio-Technology (DBT), Govt. of India, India.

Conflict of interest

The authors declare no competing financial and non-financial interests.

Authors' contributions

RT, AD and GPSR collected and processed the datasets. RT, SP and GPSR implemented the algorithms and developed the prediction models. RT, AD, SP and GPSR analysed the results. RT and SP created the back-end of the web server the front-end user interface. RT, AD, and GPSR penned the manuscript. GPSR conceived and coordinated the project. All authors have read and approved the final manuscript.

Acknowledgements

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

References

1. Lindfors K, Ciacci C, Kurppa K, et al. Coeliac disease. Nat. Rev. Dis. Prim. 2019; 5:3

2. Caio G, Volta U, Sapone A, et al. Celiac disease: a comprehensive current review. BMC Med. 2019; 17:142

3. Singh P, Arora A, Strand TA, et al. Global Prevalence of Celiac Disease: Systematic Review and Meta-analysis. Clin. Gastroenterol. Hepatol. 2018; 16:823-836.e2

4. Dhall A, Patiyal S, Kaur H, et al. Computing Skin Cutaneous Melanoma Outcome From the HLA-Alleles and Clinical Characteristics. Front. Genet. 2020; 11:221

5. Monsuur AJ, Wijmenga C. Understanding the molecular basis of celiac disease: what genetic studies reveal. Ann. Med. 2006; 38:578–591

6. Stankovic B, Radlovic N, Lekovic Z, et al. HLA genotyping in pediatric celiac disease patients. Bosn. J. basic Med. Sci. 2014; 14:171–176

7. Zhou Z, Reyes-Vargas E, Escobar H, et al. Type 1 diabetes associated HLA-DQ2 and DQ8 molecules are relatively resistant to HLA-DM mediated release of invariant chain-derived CLIP peptides. Eur. J. Immunol. 2016; 46:834–845

8. Luckey D, Bastakoty D, Mangalam AK. Role of HLA class II genes in susceptibility and resistance to multiple sclerosis: studies using HLA transgenic mice. J. Autoimmun. 2011; 37:122–128

9. Khaleghi S, Ju JM, Lamba A, et al. The potential utility of tight junction regulation in celiac disease: focus on larazotide acetate. Therap. Adv. Gastroenterol. 2016; 9:37–49

10. Heyman M, Menard S. Pathways of gliadin transport in celiac disease. Ann. N. Y. Acad. Sci. 2009; 1165:274–278

11. Drago S, El Asmar R, Di Pierro M, et al. Gliadin, zonulin and gut permeability: Effects on celiac and non-celiac intestinal mucosa and intestinal cell lines. Scand. J. Gastroenterol. 2006; 41:408–419

12. Sander GR, Cummins AG, Henshall T, et al. Rapid disruption of intestinal barrier function by gliadin involves altered expression of apical junctional proteins. FEBS Lett. 2005; 579:4851–4855

13. Gujral N, Freeman HJ, Thomson ABR. Celiac disease: prevalence, diagnosis, pathogenesis and treatment. World J. Gastroenterol. 2012; 18:6036–6059

14. Taylor AK, Lebwohl B, Snyder CL, et al. Celiac Disease. 1993;

15. Schalk K, Lang C, Wieser H, et al. Quantitation of the immunodominant 33-mer peptide from alpha-gliadin in wheat flours by liquid chromatography tandem mass spectrometry. Sci. Rep. 2017; 7:45092

16. Ciccocioppo R, Di Sabatino A, Corazza GR. The immune recognition of gluten in coeliac disease. Clin. Exp. Immunol. 2005; 140:408–416

17. Bascunan KA, Vespa MC, Araya M. Celiac disease: understanding the gluten-free diet. Eur. J. Nutr. 2017; 56:449–459

18. Shewry PR, Halford NG. Cereal seed storage proteins: structures, properties and role in grain utilization. J. Exp. Bot. 2002; 53:947–958

19. Vita R, Mahajan S, Overton JA, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2019; 47:D339–D343

20. Boeckmann B, Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003; 31:365–370

21. Crooks GE, Hon G, Chandonia J-M, et al. WebLogo: a sequence logo generator. Genome Res. 2004; 14:1188–1190

22. Pande A, Patiyal S, Lathwal A, et al. Computing wide range of protein/peptide features from their sequence and structure. bioRxiv 2019; 599126

23. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 2011; 12:2825–2830

24. Patiyal S, Agrawal P, Kumar V, et al. NAGbinder: An approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence. Protein Sci. 2020; 29:201–210

25. 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

26. Patiyal S, Dhall A, Raghava GPS. DBpred: A deep learning method for the prediction of DNA interacting residues in protein sequences. bioRxiv 2021; 2021.08.05.455224

27. Roy T, Sharma K, Dhall A, et al. In-silico method for predicting infectious strains of Influenza A virus from its genome and protein sequences. bioRxiv 2022; 2022.03.20.485066

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; 22:936–945

29. Dhall A, Patiyal S, Sharma N, et al. Computer-aided prediction of inhibitors against STAT3 for managing COVID-19 associated cytokine storm. Comput. Biol. Med. 2021; 137:104780

30. 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:

31. Lazar-Molnar E, Snyder M. The Role of Human Leukocyte Antigen in Celiac Disease Diagnostics. Clin. Lab. Med. 2018; 38:655–668

32. do Vale RR, Conci N da S, Santana AP, et al. Celiac Crisis: an unusual presentation of gluten-sensitive enteropathy. Autops. case reports 2018; 8:e2018027

33. Lauret E, Rodrigo L. Celiac disease and autoimmune-associated conditions. Biomed Res. Int. 2013; 2013:127589

34. Troncone R, Jabri B. Coeliac disease and gluten sensitivity. J. Intern. Med. 2011; 269:582–590

35. Sollid LM, Qiao S-W, Anderson RP, et al. Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules. Immunogenetics 2012; 64:455–460

36. Rai S, Kaur A, Chopra CS. Gluten-Free Products for Celiac Susceptible People. Front. Nutr. 2018; 5:116

37. Kumar J, Kumar M, Pandey R, et al. Physiopathology and Management of Gluten-Induced Celiac Disease. J. Food Sci. 2017; 82:270–277

38. Alves TO. Determination of Gluten Peptides Associated with Celiac Disease by Mass Spectrometry. 2017; Ch. 4