Regular Paper

Epitope Prediction of Antigen Protein Using Attention-based LSTM Network

Toshiaki Noumi[1,a)] Seiichi Inoue[1] Haruka Fujita[1] Kugatsu Sadamitsu[1]
Makoto Sakaguchi[2] Akiko Tenma[2] Hironori Nakagami[3]

Received: July 29, 2020, Accepted: January 12, 2021

Abstract: B-cells inducing antigen-specific immune responses in vivo produce large amounts of antigen-specific antibodies by recognizing the subregions (epitope regions) of antigen proteins. These antibodies can inhibit the functioning of antigen proteins. Predicting epitope regions is beneficial for the design and development of vaccines aimed to induce antigen-specific antibody production. However, prediction accuracy requires improvement. The conventional epitope region prediction methods have focused only on the target sequence in the amino acid sequences of an entire antigen protein and have not thoroughly considered its sequence and features as a whole. In the present paper, we propose a deep learning method based on long short-term memory with an attention mechanism to consider the characteristics of a whole antigen protein in addition to the target sequence. The proposed method achieves better accuracy compared with the conventional method in the experimental prediction of epitope regions using the data from the immune epitope database.

Keywords: B-cell epitope prediction, protein, amino acid sequence, epitope, LSTM, attention

1. Introduction

B-cells, which induce antigen-specific immune responses in vivo, produce large amounts of antigen-specific antibodies by recognizing the subregions (epitope regions) of antigen proteins. Antibodies can inhibit the functioning of an antigen protein by binding it to its epitope region [1]. A substance that mimics the structure and function of an epitope can be considered as a "vaccine" to an organism that is aimed to induce specific antibodies in vivo. Various studies focused on epitopes have been conducted to design safe and effective vaccines. Conducting the three-dimensional structural analysis of antibody-antigen complexes by X-ray [2] or nuclear magnetic resonance (NMR) spectroscopy [3] is deemed as the most reliable way to identify epitopes recognized by B-cells. However, this procedure is expensive in terms of time, cost, and labor. Therefore, to address this problem, a computer-based epitope prediction has been introduced. Recently, various linear B-cell epitope prediction methods have been proposed [4], [5], [6], [7], [15]. Although their performance has been improved, the employed features are limited to those associated with the target amino acid sequence, and therefore, the representation capability of such models is insufficient.

In the present paper, we propose a method for linear B-cell epitope prediction using an attention-based long short-term memory network (LSTM) [16] (a deep learning approach) to incorporate not only the target amino acid sequence but also the long-range features of a whole protein.

In addition, the attention mechanism [17] is realized to enable automatically estimating the points to be emphasized while predicting each amino acid in and out of candidate epitopes. Furthermore, to address the problem of data sparseness, we extend the method to enable the simultaneous consideration of the physical and chemical features of an entire antigen protein in a deep learning network.

Our empirical results on the Immune Epitope Database (IEDB) [18], [19] indicate that the proposed method achieves better prediction accuracy than the existing method BepiPred2.0 [4]. Some of the datasets used in the present study are available to the public [*1].

2. Task Definition

2.1 Task Definition

In the present study, the prediction of linear B-cell epitopes was based on the long-chain amino acid sequences that constituted antigenic proteins. The approximate length range of epitope regions registered in IEDB is from 5 to 20 amino acids [18], [19], [20]. Although the proposed method could be applied regardless of the length of an epitope region, in the problem setting of this research, we limited the length of a candidate epitope peptide (corresponding to the short amino acid sequences) to 8-14 amino acids. This is because the number of data for peptides with less than 7 amino acids is too small, and the number of data for peptides with 15 amino acids is too large to compare fairly. We addressed the problem of classifying peptides of 8–14 amino acids into two categories: with antibody inducing activity (posi-

[1] Future Corporation, Shinagawa, Tokyo 141–0032, Japan
[2] FunPep Co., Ltd., Shibuya, Tokyo 151–0051, Japan
[3] Department of Health Development, Osaka University Graduate School of Medicine, Suita, Osaka 565–0871, Japan
[a)] t.nomi.pb@future.co.jp

[*1] https://www.kaggle.com/futurecorporation/epitope-prediction.

© 2021 Information Processing Society of Japan

Table 1 Statistics of the used dataset.

length	Number of peptides	Number of proteins	Ratio of positives
8	2,271	135	0.197
9	242	128	0.675
10	3,276	180	0.227
11	346	125	0.204
12	758	172	0.492
13	245	186	0.736
14	337	127	0.458
total	7,475	646	0.293

Table 2 One of the datasets after homology deletion.

length	Number of peptides	train peptides	test peptides
8	1,967	1,860	107
9	161	144	17
10	3,121	2,981	140
11	152	135	17
12	643	537	106
13	207	182	25
14	244	205	39
total	6,495	6,044	451

tive) and without such activity (negative).

2.2 Employed Dataset

Information on whether or not an amino acid peptide exhibited antibody-inducing activity (marked by an activity label) could be obtained from IEDB [18], [19], [21], which was used in many previous studies. Accordingly, this information was used as the label data. We also obtained the epitope candidate is part of a protein (called "peptide" in this paper) and its activity label data from the B-cell epitope data provided in IEDB. The presented antibody proteins were restricted object type linear peptides with no defects in parent protein and also restricted to IgG that constituted the most recorded type in IEDB. For convenience, we excluded records representing different quantitative measures of antibody activity for the same peptide from experiments. The epitope data obtained from IEDB corresponded to the five types of activity: "Positive-High," "Positive-Intermediate," "Positive-Low," "Positive," and "Negative." However, due to the limited number of data elements marked with the "Positive-High," "Positive-Intermediate," and "Positive-Low" labels, we equally considered these labels as "Positive", thereby attributing the task to a binary estimation. In **Table 1**, we represented "the number of peptides," "the number of proteins," and "the ratio of positives" in the considered data including each length. Notably, the number of significant digits was three. As shown in column "the ratio of positives" in Table 1, we noted that there was a difference between the positive ratio values corresponding to various lengths.

Concerning this population, we extracted the dataset corresponding to 10 different random states with the ratio of the learning data to the evaluation data at approximately 10:1. The data were split without duplication across all sets. In each dataset, there were no duplications of proteins across train and test data. Furthermore, we excluded the peptides with high homology from each dataset, using CD-HIT [23]. CD-HIT is the tool, which determines degree of homology between two sequences. In this study, we calculated degree of homology for all combinations of peptide sequences in the train and test dataset by CD-HIT. Similar to the previous study [24], we set 40% as a maximum degree of homology and used the data as no homology sequences data.

One of the 10 data sets created as described above is shown in **Table 2**.

3. Related Work and Issues

The early computer-based linear B-cell epitope prediction methods focused solely on the physicochemical properties of the amino acids constituting protein [22]. In contrast to these manually derived predictions based on specific indices, machine learning methods incorporating the information from the amino acid sequences themselves achieved high performance [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. Various methods based on machine learning algorithms were proposed, including those using random Forest, BepiPred-2.0 [4], support vector machine (SVM) method, LBtope [5], SVMTriP [6], COBEpro [7], BCpred [8], AAPpred [9],Bayesb [10], LEPS [11], BEST [12], BEORACLE [13], SVM and AdaBoost-Random forest method, LBEEP [14], Recurrent Neural Network (RNN) [25] and ABCpred [15]. Although the performance of these methods could not be unconditionally compared due to different datasets used in the experiments, BepiPred-2.0 achieved the best results [4] in that there is a significant difference in AUC between BepiPred-2.0 and others. In several previous studies, short amino acid sequences (1-3 amino acids), before and after the target peptide, were added as features. However, the long-range features of antigen proteins were not incorporated.

To process the long-distance information outside of peptides in antigenic proteins, several approaches considering amino acid sequences as series data were proposed. For example, ABCpred based on RNN corresponding to a type of deep learning did not include the amino acid sequence information outside an epitope. One of the related problems lied in inability to handle long-distance information appropriately, as it caused the problem of vanishing gradient during the learning process of RNN. In the present study, we consider the target peptide and the amino acid sequences before and after the target peptide as the new target peptide and address the problem of vanishing gradient by applying LSTM [16].

In addition, one of the essential problems associated with machine learning is that it is difficult to derive accurate predictions when the amount of training data is limited. In the previous study [5], physicochemical properties, such as type, composition, hydrophobicity, polarity, and the stability of amino acids in peptides, were employed as features in SVM. The proposed method further solves this problem by extending the model to include in a neural network the physical and chemical features of an entire antigen protein in addition to the amino acid sequence information.

4. Proposed Method

4.1 Overview of Proposed Method

While predicting epitopes, the features and amino acid sequences within an epitope may not provide sufficient information for prediction. In the present study, we develop a method based on LSTM with an attention model in which the amino acid sequences inside and outside the epitope are considered as series

© 2021 Information Processing Society of Japan

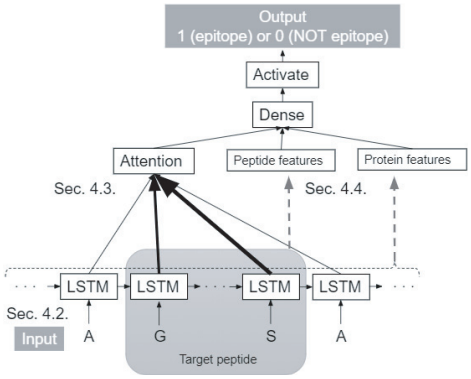


Fig. 1 Network structure of the proposed model. "Sec." number in the figure indicates the corresponding section that describes each part.

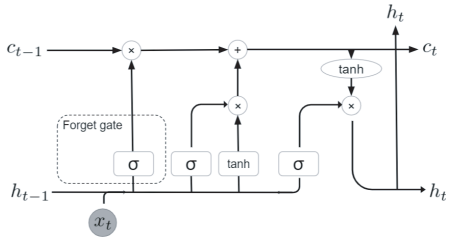


Fig. 2 Structure of an LSTM block at a certain series point t, where x_t is the input at point t corresponding to the tth amino acid in this case; c_t is the memory cell that preserves the long-range information; h_t is the embedded vector of the hidden state; σ is sigmoid function; *tanh* is the hyperbolic tangent function; $\times$ is the Hadamard product of the matrix, and + is the simple addition.

data to handle the long-range information on proteins.

Figure 1 represents the overall scheme of the proposed method. In the following sections, we introduce LSTM and attentional mechanisms, and then describe the proposed method in detail.

4.2 LSTM

RNN [25] is a deep learning method that can be used to handle series data. However, it is not applicable to long-range series information due to the problem of vanishing gradient, meaning that the gradient becomes small as a result of backpropagation during training.

LSTM [16] is an RNN model that can handle the long-range sequence information using memory cells and gates. **Figure 2** represents a schematic diagram of blocks (rectangles described as "LSTM" in Fig. 1) at the point t, where t denotes a time series point in the focus of attention in LSTM. Concerning memory cells (denoted as c_t in Fig. 2), we can observe that they pass only through simple addition and the Hadamard product by following a backpropagation flow (the flow to c_{t-1} in Fig. 2). The backpropagation of simple addition is set to 1 on partial differentiation so that the gradient does not change, while the partial differentiation of the Hadamard product depends only on the output of a forgetting gate (the "Forget gate" in Fig. 2). That is, the gradient of an element which is considered to be forgotten by the gate becomes small, while being maintained and transmitted in the past direction otherwise. Therefore, a memory cell can propagate the information that should be stored for a long time without losing it. In the present study, bi-directional LSTM [26] that is defined as a combination of the forward and backward LSTM models is implemented to combine the information obtained from the series corresponding to opposite directions.

4.3 Attention Model

Although LSTM has the ability to retain the long-range information in contrast to RNN, the detailed information in an input series tends to be lost, as it cannot be represented as a compressed vector (h_T), where T denotes the last point in the series. Therefore, we implement an attention mechanism that can refer directly to the information of an input series. Using such an attention mechanism, it is possible to memorize the vector output from an LSTM block at each series point and then multiply them by weights to obtain a context vector defines what element in the context is to be focus on (the "Attention" part in Fig. 1). LSTM with an attention mechanism has achieved remarkable results in various application fields, including natural language processing [17]. In the present study, we introduce an attention mechanism to consider which parts of proteins should be emphasized in the model.

4.4 Utilization of the Physical and Chemical Features of Amino Acid Sequences

In general, the problem of machine learning methods is that they cannot be used to learn sufficiently in the cases when the amount of training data is limited. In the present study, we incorporate the physical and chemical features within peptides and whole antigenic proteins to enable robust predictions even when the amount of data is limited. The physical and chemical features of the peptides considered in this study are the following: β-turn [27], relative surface accessibility [28], antigenicity [29], and hydrophilicity [30]. They are obtained using an epitope prediction application programming interface (API) provided by IEDB [19]. The total antigen protein features are considered as they are expected to affect the ease of binding between proteins and the epitopes. The following four features: isoelectric point, aromaticity, hydrophobicity, and stability, are obtained using the Biopython library [31]. A total of eight of the physical and chemical features are integrated into the network of the proposed method, as described the next section.

4.5 Epitope Prediction of Antigen Protein Using Attention-based LSTM Network

LSTM described in Section 4.2 is employed to preserve the long series information, and the attention mechanism introduced in Section 4.3 enables the prediction mechanism of notable locations. Then, incorporating the physical and chemical features, as suggested in section 4.4 allows enabling robust estimation even when the amount of training data is limited. In this section, we demonstrate how these three features can be combined in a single model to perform epitope predictions that are the subject of the present research.

In Fig. 1, the input sequence information describes each amino acid (A, G,..., S) inside and outside a peptide, and a LSTM block at each sequence point receives the amino acid information. In LSTM, the amino acid information at the point in a series and

© 2021 Information Processing Society of Japan

that before and after this point are combined into an embedded vector (h_t). This allows capturing the long-range amino acid sequence information.

Then, we apply the attention mechanism to estimate which amino acids are particularly noteworthy for the purposes of epitope estimation. For example, as the information corresponding to an inside peptide is considered to be more important than that associated with an outside one, it is possible to assign a higher weight to an inside peptide to propagate the information.

The final embedding vector obtained using LSTM with the attention mechanism and the physical and chemical features of the amino acid sequences, described in Section 4.4 (denoted as "peptide/protein features" in Fig. 1) are combined with the dense layer, and finally the presence of an epitope is judged using the sigmoid function (the "Activate" layer). In the model learning phase, the model parameters at each learning step are employed to perform the above estimation, and then the model backpropagates the loss to epitope determination and updates the model parameters in each layer.

5. Evaluation Experiments

5.1 Experimental Conditions

To evaluate the applicability of the proposed methods, we compared with the baseline approach on the same dataset. In this experiment, we used BepiPred-2.0 [4] as the baseline model. **Table 3** represents the comparison of the features considered in BepiPred-2.0 and in the proposed method. Here, in this experiment, we adopted 16 amino acid residues as the window size that scored the highest in our preliminary experiments.

Each model finally outputted a value in the range between 0.0 and 1.0 that indicates epitopicity. When this value was greater than 0.5, it is regarded as epitope, and when it was less than 0.5, it was considered as non-epitope. BepiPred-2.0 calculated predictions on the basis of amino acids, so that we used the average of the prediction scores for each amino acid as the predicted score of a target peptide. The trained model for BepiPred-2.0 was published as API [19], and we employed it to obtain results for evaluation, as described in Section 2.2.

We analyzed the performance concerning each antigen in terms of binary accuracy, similarly as Ref. [15]. To compare the performance of two models, the paired t-test was applied to their accuracy estimates on individual datasets. A confidence interval of 95% was considered to identify a significant difference between two compared models. Then, to focus on the prediction of positive labeling, we utilized the three indicators of positive labeling as described below:

- Sensitivity (Sens.) = True Positive / (True Positive + False Negative)
- Positive Prediction Value (PPV) = True Positive / (True Positive + False Positive)
- F1 value = 2 × PPV × Sensitivity / (PPV + Sensitivity)

Additionally, we evaluated the area under the curve (AUC) to address the problem of bias in datasets. The chance rate in the present study was set as shown in the "Positive ratio" in Table 1 if all predictions were 1. Otherwise, it was set to 1− "positive ratio" if all predictions were 0. Notably, the chance rate was also high due to the bias in labels in the data.

Table 4 Accuracy and p-value of BepiPred-2.0 and the proposed methods. The highest scores among the compared methods in "Accuracy" columns and the value in "p-value" columns where there is a 5% significant difference are shown in bold. Macro average of accuracy for each method is shown in the first line as "Total."

Length	Method	Accuracy	p-value
Total	BepiPred-2.0	0.489	—
	proposed	**0.695**	
	w/o Attention	0.673	
8	BepiPred-2.0	0.526	0.204
	proposed	0.649	
	w/o Attention	**0.774**	
9	BepiPred-2.0	0.462	**0.002**
	proposed	**0.688**	
	w/o Attention	0.638	
10	BepiPred-2.0	0.550	**0.005**
	proposed	**0.673**	
	w/o Attention	0.657	
11	BepiPred-2.0	0.503	**0.002**
	proposed	**0.677**	
	w/o Attention	0.644	
12	BepiPred-2.0	0.395	**0.001**
	proposed	**0.790**	
	w/o Attention	0.582	
13	BepiPred-2.0	0.488	**0.000**
	proposed	**0.783**	
	w/o Attention	0.708	
14	BepiPred-2.0	0.495	**0.005**
	proposed	**0.725**	
	w/o Attention	0.705	

Table 5 Sensitivity, positive prediction value, F1 value, and Area under the curve of the BepiPred-2.0 and proposed methods. The highest scores among the compared methods are shown in bold. Macro average of the score in each method is shown in the first line as "Total."

Length	Method	Sens.	PPV	F1	AUC
Total	BepiPred-2.0	0.384	**0.768**	0.479	0.569
	proposed	**0.775**	0.725	**0.705**	0.706
	w/o Attention	0.644	0.752	0.656	**0.713**
8	BepiPred-2.0	0.530	**0.736**	0.546	0.576
	proposed	**0.858**	0.547	0.610	**0.822**
	w/o Attention	0.733	0.640	**0.644**	0.801
9	BepiPred-2.0	0.337	**0.898**	0.479	0.595
	proposed	**0.819**	0.788	**0.790**	0.626
	w/o Attention	0.697	0.839	0.732	**0.709**
10	BepiPred-2.0	0.285	0.449	0.311	0.499
	proposed	**0.393**	0.578	**0.400**	**0.792**
	w/o Attention	0.292	**0.580**	0.350	0.745
11	BepiPred-2.0	0.409	**0.980**	0.574	**0.694**
	proposed	**0.763**	0.838	**0.780**	0.628
	w/o Attention	0.682	0.855	0.744	0.693
12	BepiPred-2.0	0.339	0.689	0.409	0.470
	proposed	**0.698**	**0.790**	**0.676**	**0.739**
	w/o Attention	0.557	0.736	0.581	0.619
13	BepiPred-2.0	0.406	**0.900**	0.553	0.605
	proposed	**0.951**	0.805	**0.871**	0.661
	w/o Attention	0.752	0.859	0.777	**0.716**
14	BepiPred-2.0	0.381	0.726	0.484	0.545
	proposed	**0.940**	0.730	**0.806**	0.677
	w/o Attention	0.796	**0.756**	0.764	**0.712**

5.2 Experimental Results and Discussions

The experimental results corresponding to the BepiPred-2.0 and the proposed methods are provided in **Tables 4** and **5**. As a reference, the results of the proposed method without the attention model (in each table, "w/o Attention") were also included in a bottom line for each length.

We observed that the proposed method achieved the highest ac-

© 2021 Information Processing Society of Japan

Table 3 Comparison of the features used in BepiPred-2.0 and in the proposed method.

	BepiPred-2.0	Proposed
Embedding	Not used	Each amino acid in the target peptide and 16 amino acid before and after the peptide are embedded into a 20-dimensional vector. By concatenating the decoded vector (256-dimension) via LSTM and the attention decoder and peptide and protein features (8-dimension) as below, the output is obtained in the form of 264-dimensional embeddings.
Features about peptide	Molecular weight in amino acids, Hydrophobicity, Polarity	β-turn, Relative surface accessibility, Antigenicity, Hydrophilicity
Features about protein	Relative surface accessibility, Secondary structure	Isoelectric point, Aromaticity, Hydrophobicity, Stability

Table 6 Sensitivity, positive prediction value, F1 value, and Area under the curve of the proposed methods. The highest scores among the compared methods are shown in bold. Macro average of the score in each method is shown in the first line as "Total."

Length	Method	Sens.	PPV	F1	AUC
Total	proposed	0.775	0.725	0.705	0.706
	w/o 8 features	0.665	0.615	0.600	0.686
	w/o 4 protein features	0.714	0.667	0.644	0.695
	w/o 4 peptide features	0.757	0.734	0.704	0.704

curacy concerning all lengths compared to BepiPred-2.0, as represented in Table 4.

These results indicated that the proposed method significantly outperformed BepiPred-2.0 except for length 8, when p-values below 0.05 were considered statistically significant. Moreover, as shown in Table 5, the proposed method outperformed BepiPred-2.0 in terms of sensitivities, F1 scores, and AUC for almost all lengths. Although the scores at particular lengths were below BepiPred-2.0 in terms of PPV, the AUC scores of the proposed model were higher than those of BepiPred-2.0 in almost all lengths. Therefore, we concluded that there was room for improvement through adjusting the threshold.

Next, we compared the results of the proposed method with and without the attention model. Although only a small difference in PPV was observed between them, the sensitivity was less than that of the proposed method, suggesting that the importance of the attention model could play an important role in epitope prediction. These issues are planned to be addressed in the future research.

5.3 Contribution of the Features

Table 6 shows the results of the experiments without four peptide features, four protein features, and all of them in order to examine the contribution of the features. It confirms that both the peptide and protein features are effective. In addition, the protein features were more effective than the peptide features, because the peptide features are derived from the peptide sequence, while the protein features are completely different information.

5.4 Example of Detection by the Proposed Method

In this section, we discuss a prediction case for the proposed method. First, we considered the predicted results of BepiPred-2.0 and LSTM, including the peptides with no label data existing, as represented in **Fig. 3**, respectively. As an example of a target protein, a 10-length peptide in protein P02662 was analyzed. The horizontal axis in these figures represents the starting point of a peptide in a protein, and the vertical axis denotes the predicted value of the target peptide. Here red, blue and gray maps indicate the positive data, negative data, and no data for a label respectively. For example, a plotting point at the point, where the horizontal axis is 1 denotes the point where the amino acids from the 1st one to 10th are regarded as a peptide representing an estimation value, and it does not have the label data in the dataset.

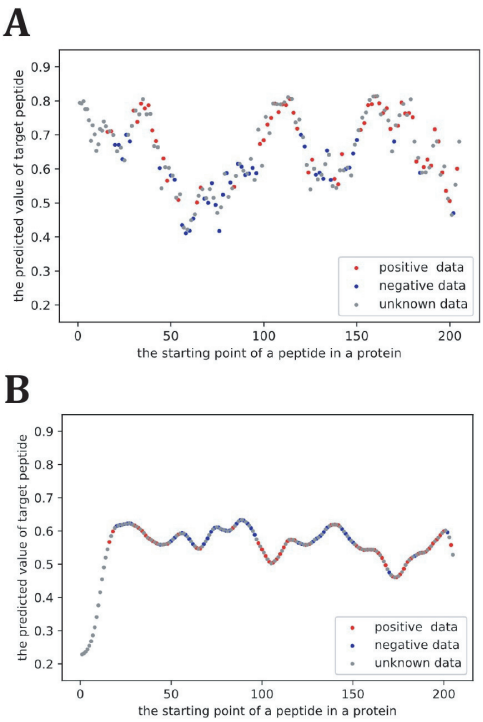


Fig. 3 A. Plot of the predicted value of protein P02662 by the proposed method (panel **A**) and BepiPred-2.0 (panel **B**).

The results of BepiPred-2.0, as represented in Fig. 3 demonstrated that the positive red and negative blue examples tended to be placed in opposite positions for corresponding to the movements of the peaks and valleys of the predictions on the vertical axis. However, concerning the proposed method, as shown in Fig. 3, it could be seen that the tendency of peaks and valleys, and the tendency of the red/blue color classification were relatively similar. In addition, if the threshold is determined appropriately, rather than being set identical (0.5) for all proteins, it is expected that the accuracy may be improved further. **Table 7** represents the peptides and their correct and estimated values concerning the three protein examples: P02662, P62314, and P22796. Here, P02662 is the same target as Fig. 3. As indicated in Table 7, the proposed method was able to recognize a peptide as non-epitope, unlike BepiPred-2.0. Concerning P62314, the answer labels were

© 2021 Information Processing Society of Japan

Table 7 Example of epitope prediction: "Start" denotes the start position of a peptide. "Ans." is an answer label indicating that the target peptide is an epitope. Each value in the "Proposed" and "BepiPred-2.0" columns represents an estimated value, and estimated value is marked in bold if it matched to an answer label.

Protein	Start	Peptide	Ans.	Proposed	BepiPred
P02662	56	SKDIGSESTE	0	**0.435**	0.594
P02662	58	DIGSESTEDQ	0	**0.411**	0.588
P02662	74	QMEAESISSS	0	**0.495**	0.610
P02662	76	EAESISSSEE	0	**0.417**	0.611
P02662	168	FYQLDAYPSG	1	**0.718**	0.497
P02662	170	QLDAYPSGAW	0	0.680	**0.475**
P02662	172	DAYPSGAWYY	1	**0.727**	0.463
P02662	174	YPSGAWYYVP	1	**0.795**	0.461
P02662	176	SGAWYYVPLG	1	**0.763**	0.469
P02662	178	AWYYVPLGTQ	1	**0.764**	0.487
P62314	70	LPDSLPLD	0	**0.428**	**0.497**
P62314	71	PDSLPLDT	0	**0.441**	0.502
P62314	72	DSLPLDTL	0	**0.419**	0.505
P62314	73	SLPLDTLL	0	**0.398**	0.504
P62314	74	LPLDTLLV	0	**0.376**	0.500
P62314	77	DTLLVDVE	0	**0.381**	**0.488**
P62314	78	TLLVDVEP	0	**0.380**	**0.488**
P62314	80	LVDVEPKV	0	**0.379**	0.504
P62314	81	VDVEPKVK	0	**0.392**	0.516
P62314	91	KREAVAGR	0	**0.399**	0.663
P22796	34	IVNTLNGFYRSL	1	**0.523**	0.425
P22796	37	TLNGFYRSLNIL	0	**0.496**	**0.450**
P22796	46	NILISLTDLEIW	1	**0.573**	0.418
P22796	49	ISLTDLEIWSNQ	1	**0.669**	0.438
P22796	55	EIWSNQDLINVQ	0	0.627	0.542
P22796	58	SNQDLINVQSAA	0	**0.398**	0.558
P22796	79	WRERVLLNRISH	0	**0.419**	0.583
P22796	82	RVLLNRISHDNA	0	**0.410**	0.565
P22796	94	QLLTAIDLADNT	1	**0.645**	0.453
P22796	157	CSASFCIMPPSI	1	**0.701**	0.454

often negative; however BepiPred-2.0 incorrectly gave a high estimated score, while the proposed method outputted the negative one. However, concerning the peptides QLDAYPSGAW in P02662 and EIWSNQDLINVQ in P22796, the proposed method also provided negative incorrectly. It is necessary to investigate and mitigate the causes of such cases of incorrect estimation in the future.

6. Conclusion

In the present paper, we proposed a new model for B-cell epitope prediction considering the following features:

(1) Not only peptides but also proteins were processed as the amino acid sequence data and were modeled using an LSTM method based on attention model.

(2) The combination of the physical and chemical features of peptides and whole proteins enabled robust predictions even on a limited amount of learning data.

We demonstrated that the proposed model achieved superior performance compared with the existing method (BepiPred-2.0). This proposed method can also be applied to the prediction of interactions between the partial and whole protein sequences.

The issues to be addressed in the future research include the need to confirm the effectiveness of the proposed method through conducting experiments based on applying it to antibody proteins other than IgG, as well as through considering the case when the physical and chemical features of an amino acid sequence are unified with the comparative method. In the experiments, we excluded peptides which were either too short or too long from the experiment for stable training. Dealing with the prediction of peptide with chain lengths other than 8–14 will be considered as an issue to focus on in future research.

Furthermore, we will focus on modeling the protein three-dimensional structures used for the prediction of nonlinear epitopes [34] that are deemed to be a predicting source of information. In addition, we would like to further improve the prediction accuracy of the proposed method by utilizing the protein three-dimensional structures. Some of the datasets used in this paper are now available on https://www.kaggle.com/futurecorporation/epitope-prediction.

References

[1] Van Regenmortel, M.H.: The concept and operational definition of protein epitopes, *Philosophical Trans. Royal Society of London, Series B, Biological Sciences*, Vol.323, pp.451–466 (1989).

[2] Rux, J.J. and Burnett, R.M.: Type-Specific Epitope Locations Revealed by X-Ray Crystallographic Study of Adenovirus Type 5 Hexon, *Molecular Therapy*, Vol.1, pp.18–30 (2000).

[3] Mayer, M. and Meyer, B.: Group Epitope Mapping by Saturation Transfer Difference NMR To Identify Segments of a Ligand in Direct Contact with a Protein Receptor, *Journal of the American Chemical Society*, Vol.123, pp.6108–6117 (2001).

[4] Jespersen, M.C., Peters, B., Nielsen, M. and Marcatili, P.: BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes, *Nucleic Acids Research*, Vol.45, pp.24–29 (2017).

[5] Singh, H., Ansari, H.R. and Raghava, G.P.S.: Improved Method for Linear B-Cell Epitope Prediction Using Antigen's Primary Sequence, *PLoS ONE*, Vol.8, e62216 (2013).

[6] Yao, B., Zhang, L., Liang, S. and Zhang, C.: SVMTriP: A method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity, *PLoS ONE*, Vol.7, e45152 (2012).

[7] Sweredoski, M.J. and Baldi, P.: COBEpro: A novel system for predicting continuous B-cell epitopes Protein Engineering, *Design and Selection*, Vol.22, pp.113–120 (2008).

[8] El-Manzalawy, Y., Dobbs, D. and Honavar, V.: Predicting linear B-cell epitopes using string kernels, *Journal of Molecular Recognition*, Vol.21, pp.243–255 (2008).

[9] Davydov, I. and Tonevitskiĭ, A.G.: Linear B-cell epitope prediction, *Molekuliarnaia Biologiia*, Vol.43, pp.166–174 (2009).

[10] Wee, L.J., Simarmata, D., Kam, Y.W., Ng, L.F. and Tong, J.C.: SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction, *BMC Genomics*, Vol.11, S21 (2010).

[11] Wang, H.W., Lin, Y.C., Pai, T.W. and Chang, H.T.: Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification, *Journal of Biomedicine and Biotechnology*, Vol.2011 (2011).

[12] Gao, J., Faraggi, E., Zhou, Y., Ruan, J. and Kurgan, L.: BEST: Improved prediction of B-cell epitopes from antigen sequences, *PLoS One*, Vol.7, e40104 (2012).

[13] Wang, Y., Wu, W., Negre, N.N., et al.: Determinants of antigenicity and specificity in immune response for protein sequences, *BMC Bioinformatics*, Vol.12, (2011).

[14] Saravanan, V. and Gautham, N.: Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor, *OMICS: A Journal of Integrative Biology*, Vol.19, pp.648–658 (2015).

[15] Saha, S. and Raghava, G.P.S.: Prediction of continuous B-cell epitopes in an antigen protein using recurrent neural network, Proteins: Structure, *Function and Genetics*, Vol.65, pp.40–48 (2006).

[16] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, Vol.9, pp.1735–1780 (1997).

[17] Bahdanau, D., Cho, K. and Bengio, Y.: Neural machine translation by jointly learning to align and translate, arXiv 1409.0473 (2014).

[18] available from ⟨http://www.iedb.org⟩.

[19] Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A. and Peters, B.: The Immune Epitope Database (IEDB): 2018 update, *Nucleic acids research*, Vol.47, pp.D.339–343 (2019).

[20] available from ⟨http://www.thinkpeptides.com/bcell.html⟩.

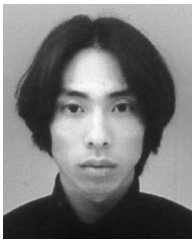
[21] available from ⟨http://www.iedb.org/database_export_v3.php⟩.

[22] Sanchez-Trincado, J. L., Gomez-Perosanz, M. and Reche, P.A.: Fundamentals and Methods for T- and B-Cell Epitope Prediction, *Journal*

© 2021 Information Processing Society of Japan

of Immunology Research, Vol.2017, (2017).
[23] Li, W. and Godzik, A.: Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, Vol.22, pp.1658–1659 (2006).
[24] Xu, Z., Li, S., Rozewicki, J., Yamashita, K., Teraguchi, S., Inoue, T., Shinnakasu, R., Leach, S., Kurosaki, T. and Standley, D.M.: Functional clustering of B cell receptors using sequence and structural features, *Molecular Systems Design & Engineering*, pp.769–778 (2019).
[25] Rumelhart, D.E., Hinton, G.E. and Williams, R.J.: Learning representations by back-propagating errors, *Nature*, Vol.323, pp.533–536 (1986).
[26] Graves, A. and Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Networks*, Vol.18, pp.602–610 (2005).
[27] Chou, P.Y. and Fasman, G.D.: Prediction of the Secondary Structure of Proteins From Their Amino Acid Sequence, *Advances in Enzymology and Related Areas of Molecular Biology*, Vol.47, pp.45–148 (1978).
[28] Emini, E.A., Hughes, J.V, Perlow, D.S. and Boger, J.: Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide, *Journal of Virology*, Vol.55, pp.836–839 (1985).
[29] Kolaskar, A.S. and Tongaonkar, P.C.: A semi-empirical method for prediction of antigenic determinants on protein antigens, *FEBS Letters*, Vol.276, pp.172–174 (1990).
[30] Parker, J.M.R., Guo, D. and Hodges, R.S.: New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites, *Biochemistry*, Vol.25, pp.5425–5432 (1986).
[31] available from ⟨https://biopython.org/⟩.
[32] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree, *Proc. Advances in Neural Information Processing Systems*, pp.3149–3157 (2017).
[33] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed representations of words and phrases and their compositionality, *Proc. Advances in Neural Information Processing Systems*, Vol.26, pp.3111–3119 (2013).
[34] Sun, J., Wu, D., Xu, T., Wang, X., Xu, X., Tao, L., Li, Y.X. and Cao, Z.W.: SEPPA: A computational server for spatial epitope prediction of protein antigens, *Nucleic Acids Research*, Vol.37, pp.W.612–616 (2009).

Toshiaki Noumi received his M.S. degree from Tokyo University in 2020, and is currently a researcher at Future Corporation. His current research interests is bioinformatics

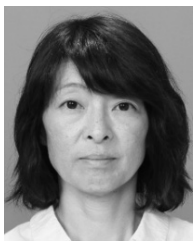


Seiichi Inoue is student at Soka University, and is currently a part-time researcher at Future Corporation. His current research interest is Natural Language Processing.

Haruka Fujita is an IT consultant at Future Corporation.

Kugatsu Sadamitsu received his Ph.D. from University of Tsukuba in 2009, and is currently a chief AI engineer at Future Corporation. His current research interests is natural language processing and bioinformatics.

Makoto Sakaguchi is currently a director in division of Translational Research at Funpep Co., Ltd. His current research interest is research and development toward peptide vaccine.



Akiko Tenma is currently a director in division of Discovery Research at Funpep Co., Ltd. Her current research interest is research and development of peptide vaccine.

Hironori Nakagami received his M.D. from Nara Medical University in 1994, and Ph.D. from Osaka University in 2003. His current research interest is translational research toward clinical application.

© 2021 Information Processing Society of Japan