

Novel machine learning method allerStat identifies statistically significant allergen-specific patterns in protein sequences

Received for publication, August 9, 2022, and in revised form, April 12, 2023 Published, Papers in Press, April 21, 2023,

<https://doi.org/10.1016/j.jbc.2023.104733>

Kento Goto^{1,‡}, Norimasa Tamehiro^{2,‡}, Takumi Yoshida¹, Hiroyuki Hanada³, Takuto Sakuma¹, Reiko Adachi², Kazunari Kondo^{2,*}, and Ichiro Takeuchi^{3,4,*}

From the ¹Department of Computer Science, Nagoya Institute of Technology, Nagoya, Aichi, Japan; ²Division of Biochemistry, National Institute of Health Sciences, Kawasaki, Kanagawa, Japan; ³Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan; ⁴Graduate School of Engineering, Nagoya University, Furo-cho, Nagoya, Japan

Reviewed by members of the JBC Editorial Board. Edited by Peter Cresswell

Cutting-edge technologies such as genome editing and synthetic biology allow us to produce novel foods and functional proteins. However, their toxicity and allergenicity must be accurately evaluated. It is known that specific amino acid sequences in proteins make some proteins allergic, but many of these sequences remain uncharacterized. In this study, we introduce a data-driven approach and a machine-learning method to find undiscovered *allergen-specific patterns* (ASPs) among amino acid sequences. The proposed method enables an exhaustive search for amino acid subsequences whose frequencies are statistically significantly higher in allergenic proteins. As a proof-of-concept, we created a database containing 21,154 proteins of which the presence or absence of allergic reactions are already known and applied the proposed method to the database. The detected ASPs in this proof-of-concept study were consistent with known biological findings, and the allergenicity prediction performance using the detected ASPs was higher than extant approaches, indicating this method may be useful in evaluating the utility of synthetic foods and proteins.

Food allergies are atopic disorders, which can be classified into immunoglobulin E (IgE)-mediated and non-IgE-mediated disorders. Allergies to cow's milk, egg, wheat, and peanuts, for examples, are IgE-mediated. Food allergies are typically caused by hypersensitivity of the immune system to specific proteins in foods, which brings about various allergic reactions ranging from itchiness, swelling of the tongue, vomiting, diarrhea, hives, trouble breathing, low blood pressure, and systematic anaphylaxis in severe cases (1, 2). Gupta *et al.* (3) showed that children under 18 years in the United States yielded an estimated food allergy prevalence of 8% and that approximately 40% of patients with food allergies have experienced a life-threatening allergic reaction. Food antigen-specific IgE antibodies play a pivotal role in most of immediate hypersensitivity

to food components. The cross-linking of IgE receptors on mast cells and basophils triggers the release of mediators such as histamines and proteases (4). Antigen-specific IgE production from B cells requires for T-cell help, and T-cell-derived cytokine induces the class-switch of B cells into IgE-producing plasma cells (5). The production of such cytokines from helper T cells is regulated by T-cell receptors, which mediate signaling through the affinity to T-cell epitope peptides presented on human leukocyte antigen (HLA) molecules expressed in professional antigen-presenting cells, and thymic microenvironments formed by medullary thymic epithelial cells (mTECs) enable T-cells lineage commitment by presenting self-antigens (6).

Some of the proteins contained in foods are the cause of food allergies. Currently, the allergenicity of proteins has been evaluated by a method based on the homogeneity to known allergenic proteins. In many countries, the method according to the Food and Agriculture Organization–World Health Organization criterion is used, in which proteins having >35% similarity in the 80-aa sliding window of allergen proteins or those identical to six-to-eight contiguous amino acids that are contained in allergen proteins are regarded as allergenic proteins (7). However, novel foods and functional proteins can now be created using genome editing technologies and synthetic biology. The goal of this study is to develop a prediction method that is more reliable even for such novel proteins. Recently, machine learning (ML) has been applied to various research field, and various bioinformatic methods have also been proposed to improve the prediction performance of Food and Agriculture Organization/World Health Organization rules. Most of them are built on whether the protein contains a similar amino acid subsequence to known allergen-specific peptides, which we call allergen-specific patterns (ASPs) (8–16). Unfortunately, these bioinformatic methods are still limited in their ability to identify allergen proteins that does not contain known ASPs; therefore, we developed a ML method to identify unknown ASPs by data-driven approach.

The goal of this study is to identify new ASPs that are responsible for allergenicity from a database of allergen proteins and nonallergen proteins. We propose an ML method

[‡] These authors contributed equally to this work.

* For correspondence: Kazunari Kondo, kondo@nihs.go.jp; Ichiro Takeuchi, ichiro.takeuchi@mae.nagoya-u.ac.jp.

called *allerStat* to efficiently identify statistically reliable ASPs and use them to predict food allergic reactions. This problem is both computationally and statistically challenging. The computational challenge is that, since the number of all possible subsequences in protein amino acid sequences is extremely huge, we need to develop a computational trick that enables an efficient exhaustive search. The statistical challenge is that, when a selection is made from a huge number of candidates, it is difficult to properly evaluate the reliability (*p* values and confidence intervals) of the selection, due to *selection bias* (c.f., multiple-comparison bias). Hence, it is necessary to develop a method to properly mitigate the bias. The main novelty of *allerStat* is that it overcomes these computational and statistical challenges by effectively combining sequence mining (17–24) and multiple testing correction methods (25–37).

As a proof-of-concept (PoC) study, we developed a dataset consisting of 21,154 proteins classified according to whether they have been reported for allergic reactions (2248 allergic and 18,906 nonallergic proteins) by collecting them from multiple databases. We applied *allerStat* to the dataset and successfully identified 5994 statistically significant ASPs after correcting the selection bias at the significance level of $\alpha = 0.05$. We observe that a part of the identified ASPs have high HLA type II binding activity which plays a significant role in atopic diseases and food allergies (38) and are consistent with known IgE epitopes (39). Furthermore, we develop an allergic reaction prediction method based on the identified ASPs and demonstrate that its prediction performance of the method is better than existing methods (40–42). We provide the database for the PoC study in the supplementary materials and the code is provided at <https://github.com/takeuchi-lab/allerStat>.

Results

In this paper, we propose an ML method called *allerStat* for identifying amino acid subsequences highly associated with allergic reactions, which we call ASPs. As a PoC, we constructed a protein dataset that contains both allergic and nonallergic proteins in various biological categories, and we applied *allerStat* to the dataset.

Allergen protein dataset

The input format of the dataset for *allerStat* is illustrated in Figure 1A. Each row of the table represents an individual protein. Each protein consists of three pieces of information: amino acid sequence, biological category, and the presence or absence of allergic reaction (label). We call a protein that causes or does not cause allergic reaction as an *allergenic protein* or a *nonallergenic protein*, respectively. Biological category information is needed to avoid misidentifying category-specific patterns as ASPs. From the original dataset, we extracted the information as illustrated in Figure 1B, in which each biological category in the dataset is classified into three types. The first type contains both allergen and nonallergenic proteins (called *paired category*), the second type only

contains allergenic proteins (called *positive-only category*), and the third type only contains nonallergenic proteins (called *negative-only category*). The distinction of these three types of categories is needed because, e.g., amino acid subsequences frequently observed in a positive-only category cannot be distinguished whether they are specific to the category or to an allergic reaction.

As a PoC, we developed a dataset of 21,154 proteins whose presence or absence of allergic reactions have been already verified (see the [Experimental procedures](#) section). The whole dataset in the form of Figure 1A is given in [Data S1](#), and the biological category information of the dataset in the form of Figure 1B is given in [Data S2](#). The PoC dataset consists of 2248 allergenic proteins and 18,906 nonallergenic proteins. The average, minimum, and maximum lengths of the proteins are 421, 5, and 34,350, respectively. There are 20 paired categories, 204 positive-only categories, and one negative-only category. It is generally difficult to discriminate between allergenic and nonallergenic proteins. Thus, we focused on 20 foods that have been well analyzed for allergens. Allergens and their family proteins were deleted based on reviewed protein information obtained from UniProt, and 17,372 nonallergenic proteins were prepared. Furthermore, the thymic medulla provides a unique microenvironment in which virtually all the self-antigens are presented so that autoreactive T cells are eliminated from the T cell repertoire before emigrating to the periphery, thus establishing central T cell tolerance. mTECs play a pivotal role in self-antigen presentation process by virtue of their promiscuous expression of tissue-restricted proteins. Therefore, such a protein expressed in mTECs basically cannot induce allergic reactions. The published gene and protein expressions were integrated to create 1534 nonallergenic proteins. When this PoC dataset is used for prediction performance evaluations, we need to take care of the fact that many similar proteins are contained in each category. To avoid data leakage issue, we employed, what we call, *leave-category-out cross-validation* (LCO-CV).

Fig. S1 shows the distributions of several physico-chemical features of allergenic and nonallergenic proteins, where, for each protein, these feature values are computed by the average for amino acids in it, with the feature values for each amino acid being presented in [Data S3](#). There is no significant difference in the distributions of any physico-chemical features between them, which suggests that it is impossible to predict allergenicity by simply using those features.

Figure 1C shows examples of amino acid sequences and patterns (subsequences). We call contiguous amino acid subsequences of various lengths *patterns*. The goal of this study is to find the patterns highly associated with allergic reactions as ASPs. The challenges stem from the fact that there exists an extremely large number of candidate patterns to be considered. For example, if we consider patterns up to length 50, because there are 20 amino acids, there are $20 + 20^2 + \dots + 20^{50} = 10^{65}$ possible patterns. In fact, we have to consider only patterns contained in the dataset, but even so, the total number of patterns in the PoC dataset is as large as

A

ID	Amino Acid Sequence	Category	Label
1	LLVAALCALVA...	Oribatida	Allergen
2	EIASQIAQED...	Carnivora	Allergen
3	WQTYVDDHLMC...	Carnivora	Allergen
4	NKLFLVSATLAL...	Bos taurus	Allergen
5	NCLFLVSATLAL...	Bos taurus	Allergen
6	PTSVAVDQGS...	Bos taurus	Allergen
7	LKVAAKCAKVA...	Gallus gallus	Non-Allergen
8	EICSQNAQED...	Oryza sativa	Non-Allergen
9	REKTGAGMQLMS...	Bos taurus	Non-Allergen
⋮	⋮	⋮	⋮

B

ID	Category	Category Type	# of Allergen	# of Non-Allergen	Total
1	Oribatida	Positive-Only	226	0	226
2	Gallus gallus	Paired	18	2262	2280
3	Actinidia deliciosa	Paired	21	11	32
4	Bos taurus	Paired	25	5953	5978
⋮	⋮	⋮	⋮	⋮	⋮

C

An Example: Patterns from Sequence “MKRRELEK”

Length	# of Patterns	Patterns (Subsequences)
1	5	M, K, R, E, L
2	7	MK, KR, RR, RE, EL, LE, EK
3	6	MKR, KRR, RRE, REL, ELE, LEK
4	5	MKRR, KRRE, RREL, RELE, ELEK
5	4	MKRRE, KRREL, RRELE, RELEK
6	3	MKRREL, KRRELE, RRELEK
7	2	MKRRELE, KRRELEK
8	1	MKRRELEK

Figure 1. Illustration of the data format for the proposed method allerStat. A, input data format. Each line indicates a protein, and each protein is represented by an amino-acid sequence, biological category name, and allergenic reaction status. B, data format for biological categories. Three types of categories (*i.e.*, paired category, positive-only category, and negative-only category) are differently treated by the data analysis method. C, examples of patterns. We call a contiguous amino-acid subsequence a *pattern*. The figure indicates that there are 33 patterns within the amino-acid sequence MKRRELEK.

3,783,825,994. Such an extremely large number of candidate patterns causes not only computational but also statistical difficulties. The main contribution in this paper is to develop a method for overcoming these computational and statistical difficulties. The number of patterns contained in the PoC dataset by their lengths is shown in Fig. S2.

Allergen-specific patterns

In this study, we call the patterns that satisfy all the following three conditions as ASPs. The definition of an ASP is that

- (1) the pattern is observed statistically significantly more frequently in allergenic proteins than in nonallergenic proteins,
- (2) the pattern is observed in none of the nonallergenic proteins, and
- (3) the pattern is not specific to particular biological category.

The first condition is needed to guarantee the statistical reliability of the identified ASPs. The second condition means that, if we expect that an ASP is considered as a candidate of a cause of allergic reaction, it should not be observed in any

nonallergenic proteins. The third condition suggests that patterns specific to a particular biological category should not be mistakenly identified as ASPs. In this study, the third condition is verified by checking whether one of the following conditions is met:

- (3a) the pattern is observed in at least one paired category or
- (3b) the pattern is observed in multiple positive-only categories.

Because a paired category contains both allergenic and nonallergenic proteins, a pattern observed more frequently in allergenic proteins can be considered allergen-specific, rather than specific to the biological category (condition 3a). On the other hand, because it is not possible to distinguish whether the pattern contained in a single positive-only category is specific to the biological category or the allergic reaction, we only regard the patterns as ASPs if they are commonly contained in multiple positive-only categories (condition 3b). [Figure 2A](#) shows the definition of ASPs, and [Figure 2B](#) shows examples of ASPs and non-ASPs.

Statistically significant pattern mining method

Because the number of candidate patterns is huge, it is necessary to introduce a computational trick to reduce the computational cost. Efficient algorithms for handling large number of patterns have been studied extensively in the data mining community. In particular, methods for efficiently handling sequential patterns are called sequence mining ([17–24](#)), and our method is built on one of the sequence mining methods called *prefix-span* ([20](#)). The large number of candidate patterns raises not only a computational challenge but also a statistical one. Over the past decade, several methods have been proposed for evaluating the statistical significance of discovered patterns in several data mining tasks ([33, 43–47](#)). In this paper, we evaluate the statistical reliability of ASPs by adapting the techniques developed in these studies to our problem setup.

As mentioned above, because the number of candidate patterns is huge, the computational cost will be extremely large if all the candidate patterns are explicitly handled in the method. In sequence mining, a tree structure of subsequence patterns is constructed, and efficient computation is made possible by pruning the branches of the tree structure, as illustrated in [Figure 3A](#). In the most basic sequence mining task called frequent sequence mining, one can efficiently find patterns whose frequency (called *support*) is greater than a certain threshold. Frequent sequence mining takes advantage of the fact that the frequency of the parent pattern in the tree structure is no less than the frequency of the child patterns. [Figure 3A](#) illustrates the concepts of tree structure and pruning in a frequent sequence mining task.

Statistical hypothesis testing based on contingency tables can be used to statistically evaluate whether certain amino acid subsequence is more frequently observed in allergenic proteins. We employed the Fisher Exact Test (FET) for testing the significance of a contingency table. Because we consider a large number of candidate patterns, we need to consider testing the

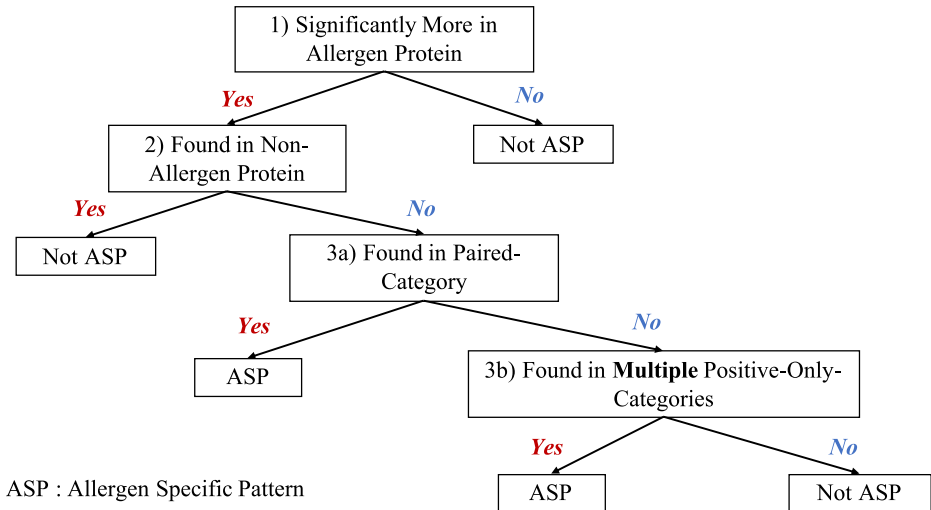
statistical significance of a large number of contingency tables by multiple FETs.

When multiple tests are conducted, a multiple testing correction is needed to adequately control the risk of false positive (FP) findings. In the context of multiple testing, one is often required to control the probability of finding one or more FPs below the significance level, a criterion known as the *family-wise error rate* (FWER). If we employ FWER, we usually specify the desired FWER level α (like $\alpha = 0.05$) and then calculate the significance level per test (called *nominal* significance level) δ . The most basic multiple testing correction method for controlling FWER is called the Bonferroni correction $\delta = \alpha/T$, where T is the number of tests. This can assure that FWER becomes less than α . However, unfortunately, in our problem, T is the number of candidate patterns and is so large that the Bonferroni correction is over-conservative. In fact, because there are 3,783,825,994 candidate patterns in the PoC dataset, only those with the nominal FET p value smaller than $\delta = 0.05/3,783,825,994 < 1.3 \times 10^{-11}$ can be considered statistically significant for $\text{FWER} < \alpha = 0.05$. [Figure 3B](#) illustrates that the statistical reliabilities of the identified ASPs can be quantified by considering multiple testing problems with multiple FETs each of which corresponds to each pattern in the dataset. Various multiple testing correction and other selection bias correction methods have been studied ([25–32, 34–37](#)).

In this study, we introduce a method for multiple testing correction for large number of contingency tables by effectively combining a randomized test called the *Westfall-Young* (WY) method ([48](#)) with a sequence mining method. What we expect for δ is that FWER is less than α when the data is randomly generated. So, in WY method, the labels (*i.e.*, allergen/nonallergen) are randomly shuffled to create a randomized dataset which does not contain any ASPs by construction. Then, we conduct multiple tests for the randomized dataset and calculate the smallest (nominal) FET p value. Because all selected patterns from the shuffled dataset in WY method are interpreted as FP findings, to avoid them, the significance threshold for FWER control must be smaller than all the (nominal) FET p values obtained from the shuffled datasets. By generating multiple (*e.g.*, 10,000) shuffled datasets with different random seeds, we can estimate the distribution of minimal (nominal) FET p values. The adjusted significance level δ is defined to be the lower 100 α % point of the minimal (nominal) FET p value distribution. [Figure 3C](#) schematically illustrates the multiple testing correction for evaluating the statistical reliability of the selected ASPs by WY method.

Unfortunately, because the WY method requires the calculation of (nominal) FET p values for all possible patterns, it cannot be used as it is. Thus, we introduce a trick to avoid computations that does not affect the lower 100 α % of the minimal (nominal) FET p value distribution. Concretely, we compute the lower bound of the (nominal) FET p values of the patterns included in each branch of the tree structure shown in [Figure 3C](#). Then, it is possible to avoid the computation for the patterns which do not affect the lower 100 α % point of the minimum (nominal) FET p value distribution by pruning the

A



B

1. Pattern “KLELS” Not Allergen Specific Pattern

Sequence	Label	Category	Category Type
...TKLELSX.....	Allergen	Poales	Positive-Only
.....ADKLELSP...	Non-Allergen	Bos taurus	Paired

2. Pattern “PSQQ” Allergen Specific Pattern

Sequence	Label	Category	Category Type
...SADPSQXS...	Allergen	Bos taurus	Paired
...MAPSQQEL...	Allergen	Bos taurus	Paired

3. Pattern “RRLE” Allergen Specific Pattern

Sequence	Label	Category	Category Type
...ELTERRRLETX...	Allergen	Poales	Positive-Only
...MLETSRRLEKS...	Allergen	Oribatida	Positive-Only

4. Pattern “MMKLE” Not Allergen Specific Pattern

Sequence	Label	Category	Category Type
...TXMMKLES.....	Allergen	Poales	Positive-Only
.....GABAMMKLEP...	Allergen	Poales	Positive-Only

Figure 2. Definition and examples of allergen-specific patterns. A, decision diagram to define ASPs. B, examples of ASPs and non-ASPs. These examples show the circumstances under which a pattern that satisfies the condition 1 is defined as an ASP. (Example 1) Pattern KLELS is not an ASP because it does not satisfy condition 2. (Example 2) Pattern PSQQ is an ASP because it satisfies conditions 2 and 3a. (Example 3) Pattern RRLE is an ASP because it satisfies conditions 2 and 3b. (Example 4) Pattern MMKLE satisfies condition 2 but is not an ASP because it is found only in a single positive-only category, i.e., satisfies neither condition 3a nor 3b. ASP, allergen-specific pattern.

branches based on the lower bound. The main technical difference of allerStat from existing sequence mining methods is that the tree pruning strategy is designed so that branches only containing patterns not affecting the estimation of the FWER distribution in the WY method are pruned. This computational trick was first introduced in (44), in which the goal was to quantify the statistical significance of itemset mining tasks.

In this study, we adapted the method to our sequence mining task.

Statistical properties of the ASPs in the PoC dataset

The ASPs identified in the PoC dataset are listed in Data S4. In total, 5064 and 5994 statistically significant ASPs were

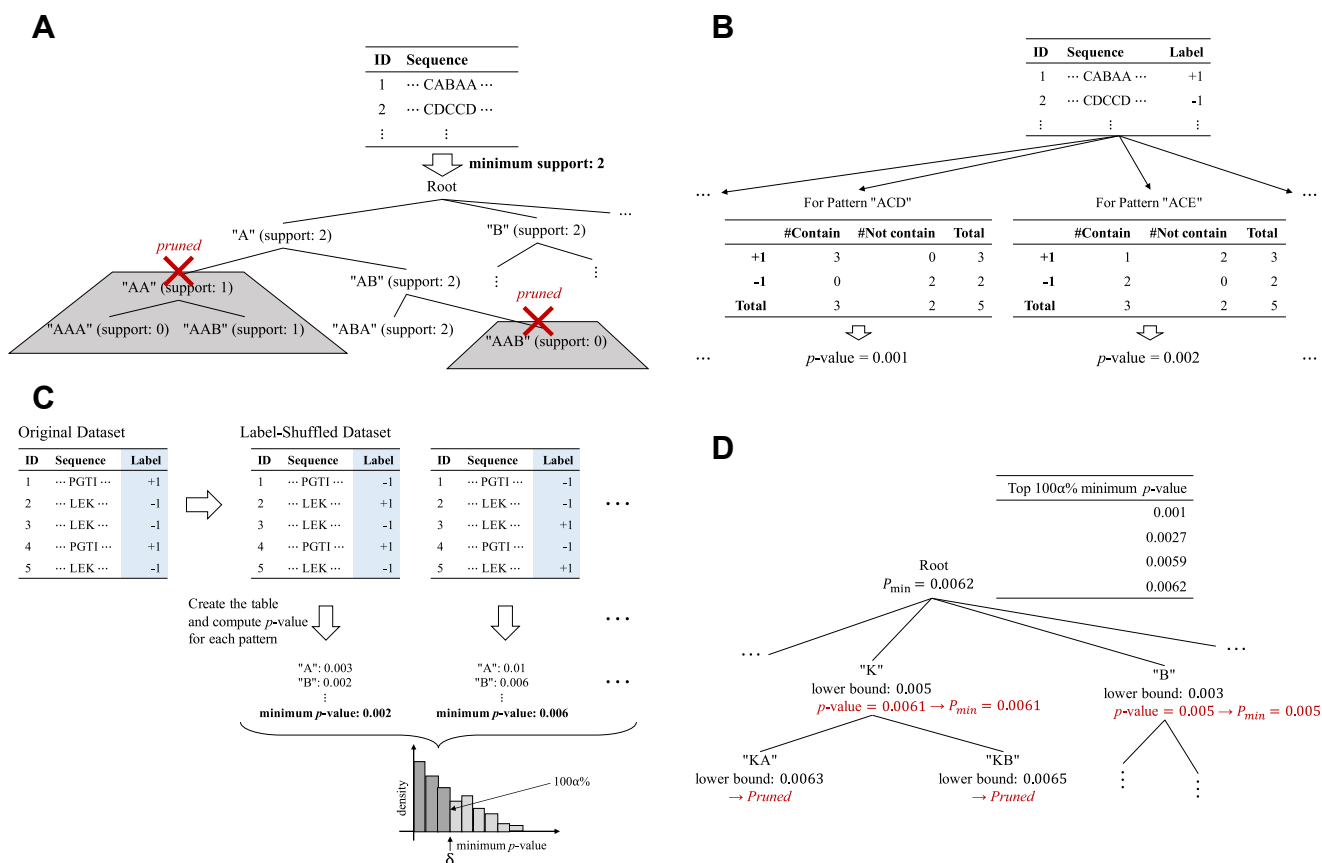


Figure 3. Schematic illustrations of the proposed method. A, illustration of a sequence mining problem and tree pruning. The *support* of a pattern is the number of sequences in the database that contains the pattern. Suppose that we find all patterns with support 2 or more. Then, if we find a pattern whose support is less than 2, the support of any extended pattern must be less than 2, and therefore we can prune (stop extending) the pattern. B, illustration of multiple testing problem with multiple Fisher Exact Test each of which corresponds to each pattern. C, family-wise error rate (FWER) controlled by the Westfall-Young (WY) method. D, illustration of a fast computation of WY method by exploiting the tree structure and its pruning.

found when the statistical significance thresholds were set to be $\alpha = 0.01$ and 0.05 , respectively. Figure 4, A–C show the distributions of the lengths, the adjusted p values, and the supports of the identified ASPs, respectively. Figure 4D shows the distributions of the number of ASPs in each allergenic protein. Figure 4E shows the distribution of how many different biological categories each ASP is included in.

Biological analysis of ASPs

The 5994 ASPs detected by allerStat with $\alpha = 0.05$ were reprocessed to remove overlapped sequences and were concatenated to 1072 patterns (ConcASPs; Data S5) to make it easier for further analysis (see the Experimental procedures section and Text S1). Among the ConcASPs, 687 patterns had a length of 15 amino acids or more. Therefore, we next examined HLA type II binding activities. A number of recent studies suggest that HLA plays a significant role in atopic diseases and food allergies (49–51). Protein antigens in foods are internalized by antigen-presenting cells such as dendritic cells and B cells. These antigens are processed into short peptides (12–25 aa long) that can bind to HLA-II and presented to T cells. The binding to HLA-II molecules, such as HLA-DR, is considered as the first step to antigen-specific IgE

antibody production. The association of the ConcASP and typical 15 HLA-DR alleles such as HLA-DRB1 was investigated using NetMHCIIpan 4.0 (38). We used the default setting of % Rank of Strong (2%) or Weak (10%), which means, according to NetMHCIIpan's instructions, "The percentile rank for a peptide is generated by comparing its score against the scores of 200,000 random natural peptides of the same length of the query peptide. For example, if a peptide is assigned a rank of 1%, it means that its predicted affinity is among the top 1% scores for the specified molecule." A total of 82% of the 687 ConcASPs with ≥ 15 aa long carried a HLA-DRB1-binding activity we examined (Fig. 5A). Of the 687 ConcASPs, 40.8% for HLA-DRB1*13.01 to 59.1% for HLA-DRB1*04.01 showed HLA-DRB1-binding activity for each major allele. Consensus motifs of core sequences to each HLA-DRB1 allele are shown in Figure 5B as Web logo. The full result of the existences of binding for the 687 ConcASPs and 15 HLA-DR alleles is presented in Data S6.

Furthermore, we examined the relationship between identified ASPs and known IgE epitopes. The web server program called allergen database for food safety provides analytical tools for searching the similarity with these validated B-cell epitopes for which site binds to the IgE from patients with food allergy. We performed a homology search with online

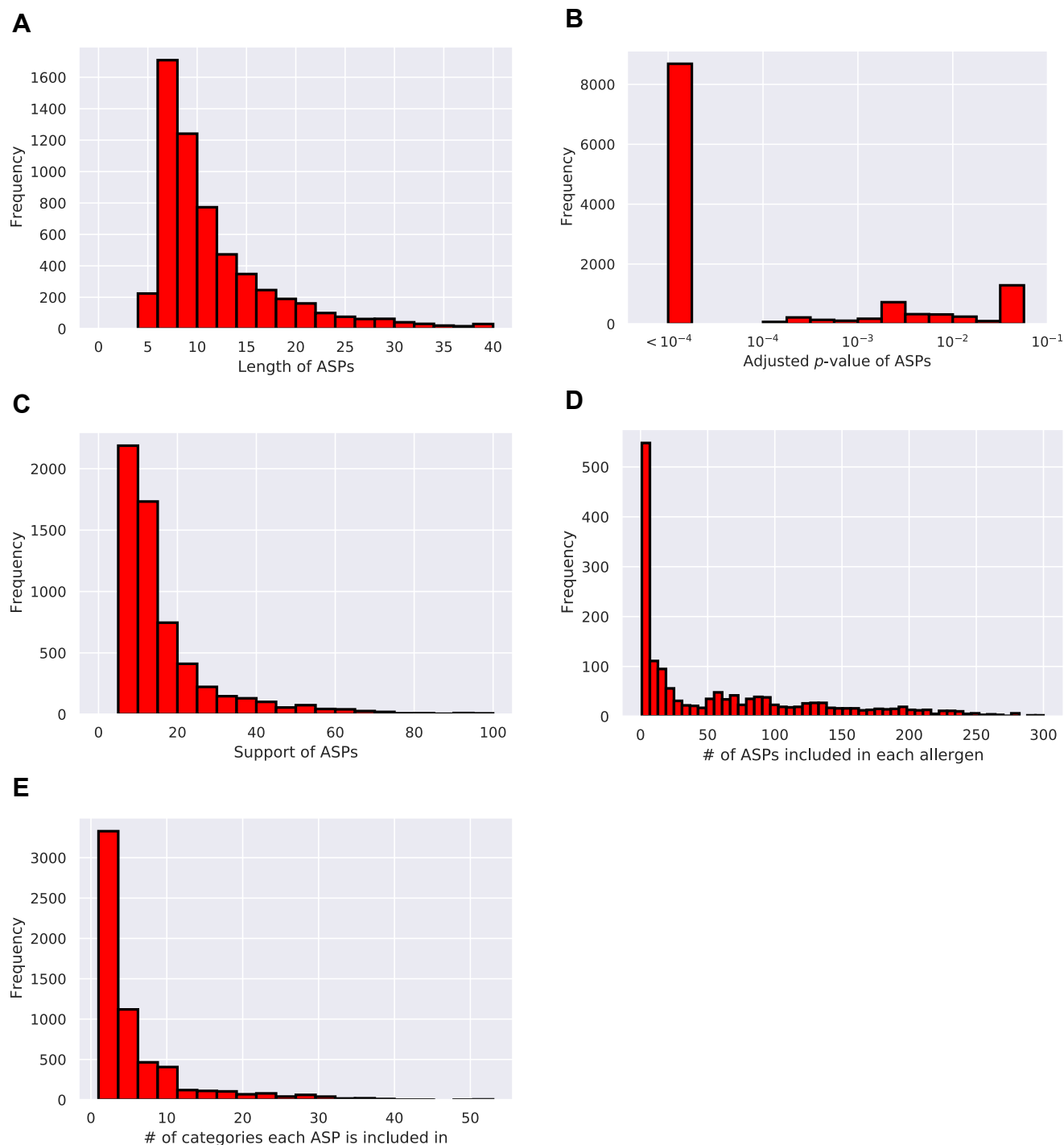


Figure 4. Statistical properties of allergen-specific patterns in the proof-of-concept dataset. A–C, distributions of the lengths, the adjusted p -values, and the supports of the identified ASPs, respectively. D, distributions of the number of ASPs in each allergenic protein. E, distribution of how many different biological categories each ASP is included in. ASP, allergen-specific pattern.

alignment search program called Protein Blast (BlastP). As a result, 24.3% of ConcASP were highly homologous to the sequences of B-cell epitopes (Data S7). For instance, pattern 537 had a high degree of sequence homology with seven epitopes identified from tree nut and bean allergens, suggesting an allergen cross reactivity (Fig. 5C). Remarkably, patterns 80, 497, and 539 shared homologies with epitope sequences identified in two completely different species: bee and peanut.

Some more specific examples are presented in Fig. S3. Additionally, 225 out of 1072 ConcASP (21.0%) were perfectly matched with the known food allergen epitopes.

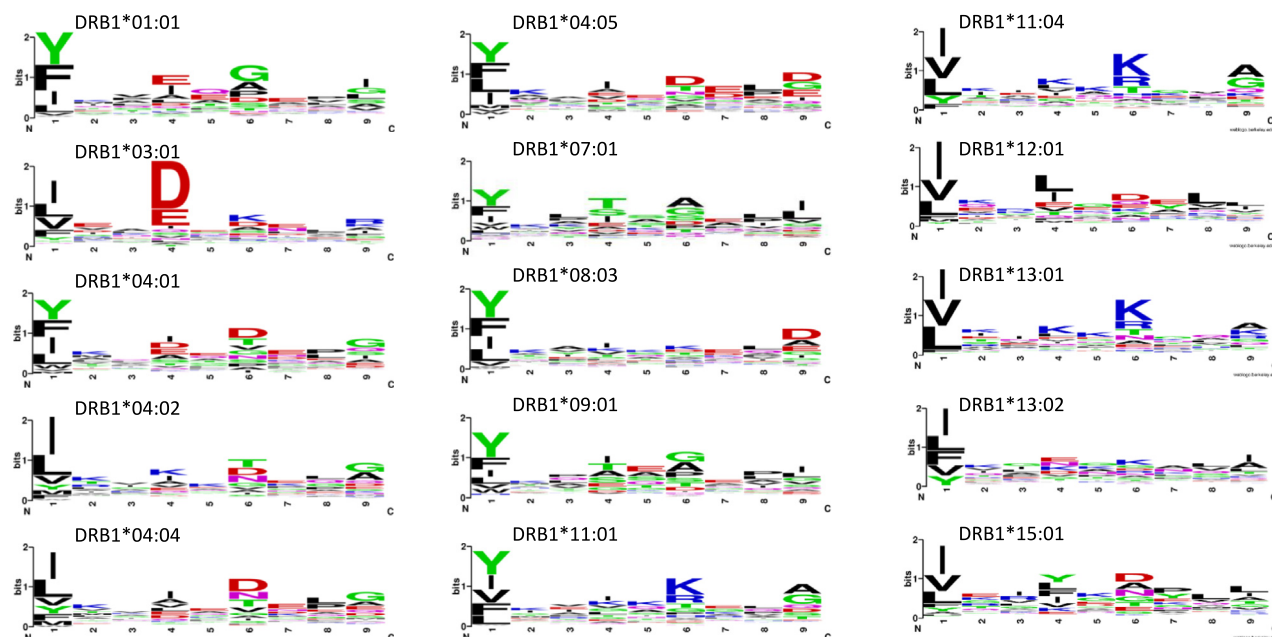
Predicting allergic reaction using ASPs

We considered the problem of predicting allergic reactions of unknown proteins using the ASPs identified by allerStat.

A

	DRB1_0101	DRB1_0301	DRB1_0401	DRB1_0402	DRB1_0404	DRB1_0405	DRB1_0701	DRB1_0901
Positive motifs (Ratio)	49.20%	41.90%	59.10%	55.60%	54.60%	54.30%	49.50%	53.40%
Positive motifs (Number)	338	288	406	382	375	373	340	367
	DRB1_0803	DRB1_1101	DRB1_1104	DRB1_1201	DRB1_1301	DRB1_1302	DRB1_1501	
Positive motifs (Ratio)	50.10%	51.50%	44.40%	44.70%	40.80%	45.10%	46.00%	
Positive motifs (Number)	344	354	305	307	280	310	316	

B



C

Pattern 537	1	---	QDRHQKIRHFR	EGDIIAFPAGV
Car i	1	---	QDRHQKIRHFR	---
Jug r 4	1	REFQ	QDRHQKIRHFR	---
Cor a 9	1	QRSE	QDRHQKIRHFR	---
Ana o 2	1	---	QDRHQKIRHFR	---
Pattern 537	1	QDRHQKIRHFR	EGDIIAFPAGV	---
Jug r 4	1	---	GDIIAFPAGVAHWSY	---
Cor a 9	1	---	GDIIAFPAGVAHWSY	---
Gly m	1	---	GDIIAFPAGVAHWSY	---
Pattern 80	1	---	SESGPFNLRSRNP	PIYSNKGKFFETPEK
Pattern 497	1	---	SESGPFNLRSRNP	PIYSNKGKFFETPEK
Pattern 539	1	---	SESGPFNLRSRNP	PIYSNKGKFFETPEK
Api c 1	1	---	SESGPFNLRSRNP	PIYSNKGKFFETPEK
Ara h 1	1	---	SESGPFNLRSRNP	PIYSNKGKFFETPEK
Pattern 80	40	ALMLPHFNSKAMVIV	---	---
Pattern 497	40	ALMLPHFNSKAMVIV	---	---
Pattern 539	40	ALMLPHFNSKAMVIV	---	---
Api c 1	40	ALMLPHFNSKAMVIV	---	---
Ara h 1	40	ALMLPHFNSKAMVIV	---	---

Figure 5. Biological characteristics of the sequence patterns specific to allergenic proteins. Protein sequences that have amino acids with a length of 15 or more were examined to predict human leukocyte antigen (HLA)–DRB1 binding activities. A, percentage of HLA-binding motifs contained in the extracted sequence patterns. B, consensus motifs of core sequences to each HLA-DRB1 allele are shown as sequence logos. C, ConcASP patterns contain validated B-cell epitope sequence. BLAST search was performed to find similarity with B-cell epitope. ASP, allergen-specific pattern.

Allergic reaction prediction from amino acid sequences is needed for safety assessments of biotechnology-based synthetic foods such as genome editing.

Cross-validation (CV) is often employed in the evaluation of predictive modeling. CV evaluates predictive performance by

removing some instances from the dataset, training an ML model with the remaining instances, and applying the trained model to the removed instances for performance evaluation. It is important to note that CV is based on the assumption that each instance is independently identically distributed (i.i.d.).

Because the proteins used in this study are strongly associated within the same biological category, we used what we call *LCO-CV*. In *LCO-CV*, all proteins in each category were removed, an ML model is trained with the proteins in the remaining categories, and the prediction performance is

investigated by applying the trained model to the proteins in the removed category. We considered *LCO-CV* for each paired category. [Figure 6A](#) is a schematic illustration of *LCO-CV*.

To predict the presence or absence of an allergic reaction, it is desirable to use not only the ASPs but also the patterns that

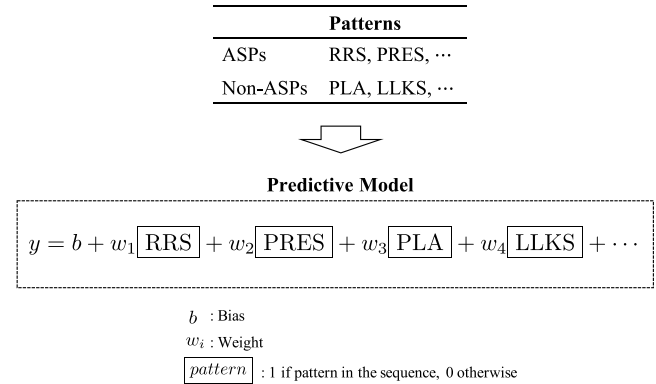
A

Gallus gallus Out				
	Gallus gallus	Actinidia deliciosa	Bos taurus	...
# of Allergen	18	21	25	...
# of Non-Allergen	2262	11	5953	...
Train or Test	Test	Train	Train	...

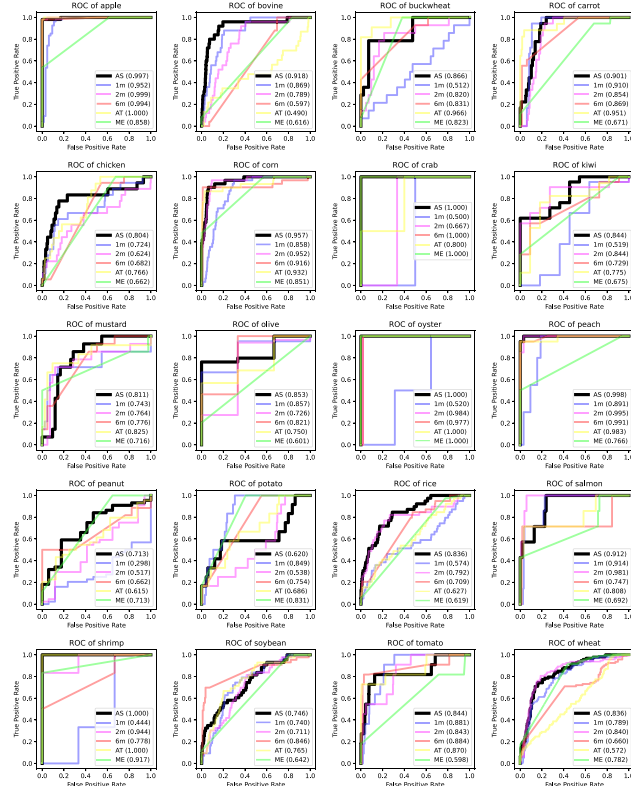
Actinidia deliciosa Out				
	Gallus gallus	Actinidia deliciosa	Bos taurus	...
# of Allergen	18	21	25	...
# of Non-Allergen	2262	11	5953	...
Train or Test	Train	Test	Train	...

Bos taurus Out				
	Gallus gallus	Actinidia deliciosa	Bos taurus	...
# of Allergen	18	21	25	...
# of Non-Allergen	2262	11	5953	...
Train or Test	Train	Train	Test	...

B



C



D

Performances by category & method, in AUC						
Criterion	Category	allerStat	1mer	2mer	6mer	Allertop
AUC	Apple	0.9969	0.9524	0.9991	0.9943	0.9997
	Bovine	0.9182	0.8689	0.7894	0.5966	0.4898
	Buckwheat	0.8656	0.5119	0.8197	0.8308	0.9659
	Carrot	0.9014	0.9099	0.8543	0.8690	0.9515
	Chicken	0.8037	0.7241	0.6242	0.6817	0.7662
	Corn	0.9569	0.8583	0.9522	0.9165	0.9321
	Crab	1.0000	0.5000	0.6667	1.0000	0.8000
	Kiwi	0.8442	0.5195	0.8442	0.7294	0.7754
	Mustard	0.8112	0.7432	0.7636	0.7755	0.8254
	Olive	0.8532	0.8571	0.7262	0.8214	0.7500
	Oyster	1.0000	0.5195	0.9844	0.9766	1.0000
	Peach	0.9984	0.8906	0.9953	0.9906	0.9828
	Peanut	0.7126	0.2981	0.5174	0.6618	0.6151
	Potato	0.6201	0.8492	0.5376	0.7540	0.6859
	Rice	0.8362	0.5744	0.7915	0.7086	0.6265
	Salmon	0.9124	0.9143	0.9806	0.7475	0.8079
	Shrimp	1.0000	0.4444	0.9444	0.7778	1.0000
	Soybean	0.7463	0.7395	0.7107	0.8461	0.7648
	Tomato	0.8439	0.8807	0.8432	0.8842	0.8697
	Wheat	0.8364	0.7889	0.8401	0.6604	0.5723
	Average	0.8729	0.7173	0.8092	0.8111	0.8090

Performances by method, in various criteria						
Criterion	Category	allerStat	1mer	2mer	6mer	Allertop
AUC	Average	0.8729	0.7173	0.8092	0.8111	0.8090
AUC-10%	Average	0.7697	0.5839	0.6859	0.7404	0.7561
F1	Average	0.5174	0.0066	0.4051	0.4939	0.4577
MCC	Average	0.4495	0.0007	0.3160	0.4388	0.4249

Figure 6. Illustrations and results of prediction analysis. *A*, schematic illustration of leave-category-out cross validation (*LCO-CV*). *B*, schematic illustration of prediction analysis by sparse support vector machine (SVM). *C*, receiver operating characteristic (ROC) curves for 20 paired categories in *LCO-CV*. Names of feature extraction methods are abbreviated as follows: AS (AllerStat; proposed method), 1m (1-mer), 2m (2-mer), 6m (6-mer), AT (Allertop), and ME (MEME). Parenthesized numbers are AUC scores. *D*, prediction performances in four criteria for 20 paired categories in *LCO-CV*. The full table (performances by each category for all criteria) is given in [Data S8](#). AUC, area-under-curve.

AllerStat: Significant allergen-specific sequences

are more frequently observed in nonallergens. Therefore, we also considered non-ASPs which were similarly identified as ASPs. Non-ASPs differ from ASPs, in that non-ASPs should be significantly more frequently appeared in nonallergenic proteins but may also be appeared in a small number of allergenic proteins. Therefore, without the condition corresponding to the second condition 2) in the definition of ASP, a non-ASP is defined as

- (1') the pattern is observed statistically significantly more frequently in nonallergenic proteins than in allergenic proteins and
- (3') the pattern is not specific to particular biological category.

We note that, from the reasons above, we employ non-ASPs only for allergenicity prediction tasks.

When predicting an allergic reaction of a new protein, it is desirable to interpret why the protein is or is not determined to have allergic reaction. Some ML models, such as deep neural networks, are called black-box models in that are too complex to interpret how the predictions are made. Because the interpretation of the prediction process is important in the current problem, we adopted a linear prediction model in which a feature is defined by the existence of each ASP or non-ASP in the amino acid sequence. In the training set, each row corresponds to a protein and each column corresponds to each ASP or non-ASP, where the ASPs and non-ASPs in the training set are patterns that were identified without using the proteins in the removed category by LCO-CV. [Figure 6B](#) illustrates the concept of a predictive model based on the ASPs and the non-ASPs selected by support vector machine (SVM). See the Methods section for the detail of prediction model development by SVM.

[Figure 6, C and D](#) show the results of the predictive analysis based on the LCO-CV for the PoC dataset. To evaluate the effectiveness of the ASPs identified by allerStat, we compared the prediction performances with the methods proposed in previous studies. There are several previous studies on ML-based allergen prediction from amino acid sequences ([10–16, 40–42, 52, 53](#)). An ML-based allergen prediction model consists of two steps: Step 1) feature extraction from amino acid sequences and Step 2) model training based on the extracted features. Because the goal of the current study is to evaluate the effectiveness of the ASPs identified by allerStat, we only compared the feature extraction method at Step 1 with those of previous studies and used the same ML model (sparse SVM) at Step 2.

We compared the ASPs obtained by allerStat with other sequence-derived features used in existing allergenicity prediction studies: *Allerdictor* ([41](#)), *AlgPred2* ([53](#)), *Allertop* ([42](#)), and *MEME* ([40](#)). Note that, for fair comparison, we only used information that can be obtained from amino acid sequences in the PoC dataset. The sequence-derived features in *Allerdictor* and *AlgPred2* are contiguous sequence of amino acid with fixed length k , which is often called k -mer. The main drawback of k -mers is that, different from allerStat, it could only consider patterns of fixed length. We investigated the

performances of k -mer with $k = 1, 2, 6$, because *Allerdictor* uses 6-mer features, while *AlgPred2* uses either 1-mer features or 2-mer features. *MEME* is a method for extracting frequent amino acid subsequences based on a probabilistic model called the *hidden Markov model*. The main drawback of the feature extraction method by *MEME* is that it is an unsupervised method and therefore cannot extract patterns observed at different frequencies in allergenic and nonallergenic proteins. *Allertop* differs from *allerStat*, *Allerdictor*, and *MEME* in that the features extracted by *Allertop* are constructed from the physico-chemical features of the entire amino acid sequence. The main drawback of the feature extraction method by *Allertop* is that it is difficult to interpret which parts of the amino acid subsequence are associated with the allergic reaction. Note that, unlike *allerStat*, the statistical reliabilities of the patterns extracted in these three previous methods cannot be properly evaluated.

[Figure 6C](#) shows the receiver operating characteristic (ROC) curves of the prediction results based on LCO-CV. [Figure 6D](#) shows the area-under-curve (AUC) scores of the ROC for various prediction models. Also, we evaluated three other criteria to evaluate binary classification performances: AUC-10%, F1 score, and Matthews Correlation Coefficient (MCC) for the average of 20 categories. Details of these criteria are presented in Methods section, and the full table of the results is presented in [Data S8](#). When interpreting the results, it is important to note that the sample size of each category varies greatly, and the numbers of allergenic and nonallergenic proteins are highly unbalanced. We employed these four criteria because they are commonly used for evaluating prediction performances for such unbalanced data. It can be observed that the prediction performances of *allerStat* are consistently better than the three existing methods. The information on the prediction model trained with the entire PoC dataset is provided in [Data S9](#).

Discussions

The problem of finding ASPs based on a database of allergenic/nonallergenic proteins can be regarded as a typical feature selection problem for a binary classification problem. However, it is important to note that the proteins in the database are not i.i.d., which is a common assumption taken in conventional classification problem. In particular, proteins belonging to the same category (apple, bovine, etc.) have amino acid sequences that are specific to that category, and if the search is not conducted carefully, category-specific amino acid sequences will be identified instead of allergen-specific ones. To address this problem, we divided the proteins in the database into categories and used different criteria for ASPs depending on whether each category was paired, positive-only, or negative-only. Furthermore, in the allergen prediction task, care must be taken not to make predictions based on category-specific amino acid sequences by introducing LCO-CVs in which each category is excluded from the training data. In the protein allergen study, it is important to carefully take into account for the “non-i.i.d.ness” of proteins.

Our results show that 82% of the sequence patterns specific to allergenic proteins obtained by allerStat carried the binding activity to any of the 15 major HLA-DRB1 alleles. Almost 50% of the patterns had the HLA-DRB1-binding affinity for either of the 15 HLA-DRB1 alleles. This indicates that allerStat can extract biologically significant patterns. Sequence logo analysis indicates that the core sequences had characteristic motifs at the position of 1, 4, 6, and 9 in 9 core sequences, especially DRB1*01.01, *04.01, *04.02, *07.01, and *11.01 (Fig. 5B). This suggests that our extracted patterns include many core sequences important for major HLA-DRB1 binding. We did not evaluate HLA-binding activities of the ConcASPs having a length less than 15 aas. The remaining 18% of the patterns that do not have any HLA-DRB1-binding activity need to be investigated to clarify their importance, because they may include biologically significant sequences. On the other hand, nonallergenic patterns have sequences with three or six amino acid long, which do not have any consensus motif.

Among the identified ASPs, 1083 allergen-specific IgE epitopes on 247 food allergens have been experimentally elucidated. Among the obtained 1072 ConcASPs from allerStat, 225 of which (21%) included any of the entire length of 125 epitope sequences. On the other hand, only 19 epitope sequences were found in 17,372 nonallergen sequence (0.1%). Although three epitopes including Pen c 3 epitope (ISSK and YGVA) and Fag e 1 epitope (QQPGQ) had overlapping in between, all 19 epitopes length is in three to seven amino acids considering to have low specificity. Moreover, 24.3% of ConcASP showed high homology with known food allergen epitopes. Interestingly, there are many patterns homologous with epitopes among different species (Fig. 5C). These results indicated that allerStat could properly extract B-cell epitope sequences characteristic to the allergen from our dataset, even though epitope information was not held in the training data. The remaining 75% patterns are thought to contain T-cell epitopes and derived unexpected sequences, such as those related to immunological tolerance and antibody production, in addition to unknown T- and B-cell epitopes.

Experimental procedures

Allergen and nonallergen data collection

Allergen data were retrieved from full-length amino acid sequence list in the COMprehensive Protein Allergen Resource (COMPARE) 2020 data (<https://comparedatabase.org/>). The COMPARE database, a collaborative effort of the HESI Protein Allergenicity Technical Committee, is a curated database comprising allergen-sequence-associated peer-reviewed publications. Nonallergen data were collected by reviewing protein sequence expressed in 20 foods including Wheat (*Triticum aestivum*), Bovine (*Bos taurus*), Chicken (*Gallus gallus*), Soybean (*Glycine max*), Crab (*Scylla serrata*), Shrimp (*Penaeus monodon*), Peanut (*Arachis hypogaea*), Buckwheat (*Fagopyrum esculentum*), Salmon (*Salmo salar*), Kiwi (*Actinidia deliciosa*), Mustard (*Sinapis alba*), Olive (*Olea europaea*), Carrot (*Daucus carota*), Apple (*Malus domestica*),

Tomato (*Solanum lycopersicum*), Peach (*Prunus persica*), Potato (*Solanum tuberosum*), Corn (*Zea mays*), Rice (*Oryza sativa*), and Oyster (*Crassostrea gigas*) obtained from UniProt (<https://www.uniprot.org/>). It is known that there are many food-allergic patients for these selected foods, and allergens in them that commonly cause an allergic reaction have been well analyzed historically. Therefore, not only major allergen but also minor allergens have been shown in previous studies such as component analysis. It is indicated that the presence of unknown allergens is considered low as compared with foods that few food allergy patients are reported. Since the prediction performance of ML strongly rely on the data quality, known allergens in these foods were listed by adding the data from Allergen Database for Food Safety to the COMPARE database. Furthermore, to eliminate the possibility of cross-reactivity induced by proteins belonging to the same family for allergens in each of the foods, nonallergen dataset was finally created by excluding allergen and its similar protein from the total protein. mTEC dataset were generated by integrating gene and protein expression profiles (54, 55). Data reduction of gene expression in SGLT1+ mature mTECs (GEO# GSE49625) were selected with the cutoff level of $t = 5$ (adjp ≈ 0.01).

HLA-II-binding activity of motifs specific to allergenic proteins

Among 1072 concatenated ASPs, 687 of which had a length of 15 aa or more. HLA-DRB1-binding activity was predicted by NetMHCIIpan server ver.4 (<https://services.healthtech.dtu.dk/service.php?NetMHCIIpan-4.0>), which requires amino acid sequences with 15 aa or longer. The HLA-DRB1-binding activity of the allergen-specific sequences was examined at 15 aa windows from the N terminus to the C terminus. ASPs having strong and weak binding activities were analyzed using sequence logo to clarify specific motifs to each HLA-DRB1 allele using Web Logo (<https://weblogo.berkeley.edu/logo.cgi>).

Similarity between the identified ASPs and known IgE-epitopes

Sequence similarity was searched by running BlastP with following parameters: E value = 2000 and cutoff of 10^{-4} . The sequences of exactly matched targets were extracted from BlastP results (E value: 20,000). As a preprocessing for this purpose, we concatenated ASPs in the following criteria. For two sequences x and y , let $x \diamond y$ indicates the concatenated sequence. We define *overlapped concatenation* between two sequences x and y as $x' \diamond w \diamond y'$ if there exists $x', y', w \in S$ (S : set of all sequences) such that $x = x' \diamond w$ and $y = w \diamond y'$. Note that the overlapped concatenation is just the ordinary concatenation $x \diamond y$ when we take w as an empty sequence. The overlapped concatenation of three or more sequences is defined as the repetition of an overlapped concatenation for two sequences. Then, we define a sequence x as a *concatenated ASP* (ConcASP) if (i) there exists an allergen sequence y in the dataset such that x is included in y as a subsequence (denoted by $x \subseteq y$) and (ii) x is represented as an overlapped concatenation of any number of ASPs. Finally, after retrieving all

ConcASPs, we extracted only the maximal ones, that is, we excluded those that are included in other concatenated ASPs. In order to find completely matching IgE epitopes to concatenated ASPs, we simply examined the inclusion, that is, checked whether $x \sqsubseteq y$ for each x in IgE epitope sequences and each y in ConcASPs. Note that, since the database of IgE epitopes we employed includes both sequence and non-sequence (e.g., conformational) epitopes, we extracted only sequence epitopes of total 1104. Among the 1072 concatenated ASPs, 225 completely matched at least one epitope sequence. We present the algorithm for overlapped concatenation as Algorithm 1 in [Text S1](#).

Fisher Exact Test

Let D , D^+ , and D^- represent the sets of all proteins, allergenic proteins, and nonallergenic proteins, respectively. Furthermore, the sizes of D , D^+ , and D^- are respectively written as n , n^+ , and n^- . To quantify whether there is a difference in the frequencies of patterns between allergenic and nonallergenic proteins, we consider the following contingency table:

	Containing pattern q	Not containing pattern q	Sum
Allergen	$D^+[q]$	$n_+ - D^+[q]$	n_+
Nonallergen	$D^-[q]$	$n_- - D^-[q]$	n_-
Sum	$D[q]$	$n - D[q]$	n

where $D[q]$, $D^+[q]$, and $D^-[q]$ denote the number of proteins in D , D^+ , and D^- which contains pattern q . To test the statistical significance, FET is used. The p value of FET is computed by considering all possible realizations of the contingency table under the condition that n^+ , n^- , and $D[q]$ are fixed. Specifically, let

$$g(n_+, n_-, D[q], a) := \frac{\binom{n_+}{a} \binom{n_-}{D[q]-a}}{\binom{n}{D[q]}}$$

be the probability of finding a pattern q in a of n_+ allergenic proteins conditional on n^+ , n^- , and $D[q]$. The two-tailed FET p value is then calculated as

$$p(D, q) = 2 \sum_{a \in \mathcal{A}(D, q)} g(n_+, n_-, D[q], a), \quad (1)$$

where,

$$\mathcal{A}(D, q) := \left\{ a \mid \max\{0, D[q] - n_-\} \leq a \leq \min\{n_+, D[q]\}, \right. \\ \left. g(n_+, n_-, D[q], a) \leq g(n_+, n_-, D[q], D^+[q]) \right\}.$$

Multiple testing correction by WY method

When we conduct only one hypothesis test, the type-I error, that is, the probability of FP finding, can be controlled at a significance level $\alpha \in (0, 1)$ (typically $\alpha = 0.05$) by rejecting the

null hypothesis when the p value is less than α . However, when we conduct many hypothesis tests simultaneously, we often need to control the FWER: the probability of having one or more FP findings. To control the FWER at the significance level of α , we need to reject the null hypothesis more conservatively, that is, a null hypothesis is rejected when the p value is less than the adjusted significance level $\delta \in (0, 1)$, where δ is usually much smaller than α . Methods for finding the adjusted significance level δ such that the FWER can be controlled to α is discussed in the context of multiple comparisons. The most commonly used multiple comparison method is Bonferroni correction in which the adjusted significance level δ is set to be $\alpha/\{\text{# of hypotheses}\}$, where $\{\text{# of hypotheses}\}$ is the number of hypotheses to be tested simultaneously. Bonferroni correction is known to be overly conservative, especially when the number of hypotheses is large. In allerStat, we employ a random permutation test called the WY method (48) as a multiple test correction method. In the WY method, multiple randomized datasets are constructed by permuting the labels (allergenic or nonallergenic in our problem setup) at random. For each randomized dataset, the minimum p value is computed. Then, the adjusted significance level δ is set to be the $[(\alpha + 1)/M]^{\text{th}}$ smallest value among M smallest p values, where M is the number of randomized datasets (typically $M = 10,000$). [Figure 3C](#) illustrates the notion of WY method.

Sequence mining and efficient computation of the WY method

Unfortunately, since allerStat needs to handle an extremely large number of patterns, it is impossible to calculate the minimum p value for each of the M randomized datasets for the WY method. In allerStat, we consider a tree structure among patterns as shown in [Figure 3](#) for efficiently computing the minimum p value. Specifically, we consider a node corresponding to a pattern q and the set of its descendant sequence patterns $\mathcal{S}(q)$. Then, our basic idea is to compute a lower bound of the FET p values of the descendant sequence patterns in the form of

$$p_{\text{low}}(D, q) \leq p(D, q') \text{ for any } q' \in \mathcal{S}(q). \quad (2)$$

If we can find such a lower bound, we can reduce the computational cost because we do not need to calculate the p values of all descendant nodes when we encounter a node whose lower bound $p_{\text{low}}(D, q)$ is greater than the current minimum p value. Here, we employed a lower bound that satisfies the property in (2), written as

$$p_{\text{low}}(D, q) =$$

$$\min \left\{ g(n_+, n_-, \min\{D[q], n_+\}, \min\{D[q], n_+\}), \right. \\ \left. g(n_+, n_-, \min\{D[q], n_-\}, 0) \right\}. \quad (3)$$

The lower bound in (3) was proposed in FastWY method in (43, 44), in which the authors studied itemset mining tasks to

find statistically significant combinations of multiple genetic factors. In allerStat, we adapted the technique in the FastWY method for sequence mining setting. Therefore, the proof that the bound (3) satisfies the property in (2) can be simply done as described in (43, 44). We present the algorithm for AllerStat as Algorithm 2 in Text S2.

Prediction model for allergic reaction using ASPs and non-ASPs

In order to predict allergic reactions by proteins using ASPs and non-ASPs, we employed the SVM with features being defined by ASPs and non-ASPs. We follow the formulation of SVM employed by SVC class with linear kernel in *scikit-learn* (<https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>), a python implementation that we used. For an overview of SVM, see, for example, Chapter 7 of (56).

SVM assumes that we are given n instances, where each instance consists of d input variables $\mathbf{x}_i \in \mathbb{R}^d$ and binary output variable $y_i \in \{-1, 1\}$ ($1 \leq i \leq n$). Here, $\mathbb{R}$ denotes the set of all real numbers, and $\mathbb{R}^d$ denotes the set of all d -dimensional real vectors. Then, given a test instance whose d -dimensional input variables are $\mathbf{x}'$ while its output variable is unknown, suppose that its prediction result is given by the linear model

$$f_{\mathbf{w}^*, b^*}(\mathbf{x}') := \mathbf{w}^{*\top} \mathbf{x}' + b^*, \quad (4)$$

where $\mathbf{w}^* \in \mathbb{R}^d$ and $b^* \in \mathbb{R}$ are to be learned by SVM. If $f_{\mathbf{w}^*, b^*}(\mathbf{x}')$ is larger than (resp. smaller than) zero, the prediction function conjectures that the output variable for $\mathbf{x}'$ is 1 (resp. -1). From the training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ defined above, SVM learns optimal $\mathbf{w}^*$ and b^* by

$$(\mathbf{w}^*, b^*) := \operatorname{argmin}_{\mathbf{w}, b} \frac{1}{2} \mathbf{w}^\top \mathbf{w} + C \sum_{i=1}^n \max\{0, y_i f_{\mathbf{w}, b}(\mathbf{x}_i)\}$$

where C is a hyperparameter to control overfitting or underfitting. For C , in this computational experiment, we chose the best one in the specified prediction performance (AUC, AUC-10%, F1 score, or MCC; explained later) by LFO-CV, except that we used common C for AUC and AUC-10% where C is chosen by AUC. We chose C from $\log_{10} C \in \{-27/9, -23/9, -19/9, \dots, 9/9\}$ (10 cases), however, we extended to smaller or larger C as long as the smallest or the largest C is the best. Note that, since this LFO-CV is conducted for the training set composed by an LFO-CV, this LFO-CV for the selection of C is conducted for a remained set of biological categories.

Then, we show how to apply ASPs by allerStat and protein sequences to the formulation of SVM above. Let $\mathcal{Q}_j \in \mathcal{S}$ ($1 \leq j \leq d$) be an ASP or a non-ASP, $z_i \in \mathcal{S}$ ($1 \leq i \leq n$) be training sequences, and $y_i \in \{-1, 1\}$ ($1 \leq i \leq n$) be their allergenicity (1 if allergenic or -1 otherwise). Then we define $\mathbf{x}_i \in \mathbb{R}^d$ as the indicator variables of including patterns in z_i , that is, $x_{ij} = 1$ if

$\mathcal{Q}_j \subseteq z_i$ or $x_{ij} = 0$ otherwise. As a result, we can employ SVM to learn an allergenicity prediction function according to ASPs and non-ASPs. Here, we can see that the j^{th} element in $\mathbf{w}^*$ is the importance of the corresponding pattern: if it is a large positive number, the pattern is expected to contribute the allergenicity. Note that the training sequences are defined by LFO-CV stated in the Results section.

Then, we show how the existing methods are compared to allerStat stated in the Results section. For k -mers and MEME, we employed the same process as that for allerStat, except for the retrievals of patterns. For 6-mers (following Allerdicator), we used length-six patterns whose supports (number of proteins that includes the pattern) are 15 or more, rather than ASPs and non-ASPs. The threshold of the support 15 is chosen so that the total number of patterns is similar to that of ASPs and non-ASPs. This is almost the same for 1-mers and 2-mers, however, we did not apply the threshold of the support, and we specified x_{ij} as the “ratio of j (k -mer) among all k -mers in z_i ” (i.e., $0 \leq x_{ij} \leq 1$) instead of specifying $x_{ij} \in \{0, 1\}$, following AlgPred2. For MEME, since it is an unsupervised method, we ran MEME by providing only allergenic proteins to retrieve ASPs, then by providing only nonallergenic proteins to retrieve non-ASPs. For Allertop, since it provides a fixed-length vector from a protein sequence, we just used it for the SVM training and the prediction.

Finally, we show the prediction performance criteria used in the experiment. First, for a set of test proteins $\{(\mathbf{x}'_i, y'_i)\}_{i=1}^m$ (converted from sequence to d -dimensional vector as above) and a prediction function (we use (4) in this experiment), we calculate true positive (TP), FP, true negative (TN), false negative (FN), true positive rate (TPR), and false positive rate (FPR) as follows:

$$TP(\theta) := \#\{i | 1 \leq i \leq m, f_{\mathbf{w}^*, b^*}(\mathbf{x}'_i) \geq \theta \wedge y'_i = +1\},$$

$$FP(\theta) := \#\{i | 1 \leq i \leq m, f_{\mathbf{w}^*, b^*}(\mathbf{x}'_i) \geq \theta \wedge y'_i = -1\},$$

$$TN(\theta) := \#\{i | 1 \leq i \leq m, f_{\mathbf{w}^*, b^*}(\mathbf{x}'_i) < \theta \wedge y'_i = -1\},$$

$$FN(\theta) := \#\{i | 1 \leq i \leq m, f_{\mathbf{w}^*, b^*}(\mathbf{x}'_i) < \theta \wedge y'_i = +1\},$$

$$TPR(\theta) := TP(\theta) / [TP(\theta) + FN(\theta)],$$

$$FPR(\theta) := FP(\theta) / [FP(\theta) + TN(\theta)].$$

Then, the ROC curve is defined as the curve in a two-dimensional space $\{(FPR(\theta), TPR(\theta))\}_{\theta=-\infty}^{+\infty}$, and its AUC score is defined by the integral of the curve over $[0, 1]$ (note

AllerStat: Significant allergen-specific sequences

that FPR must be between 0 and 1). AUC-10% is defined by the integral of the curve over $[0,0.1]$ multiplied by 10: it observes the AUC for low FPR, and the multiplication of 10 is employed to ensure the AUC-10% is 1 for completely correct classification. The F1 score and MCC are defined as follows:

$$F_1 := \frac{2 \cdot TP(0)}{2 \cdot TP(0) + FP(0) + FN(0)},$$

$$MCC := \frac{TP(0)TN(0) - FP(0)FN(0)}{\sqrt{TPFP_0 \cdot TPFN_0 \cdot TNFP_0 \cdot TNFN_0}},$$

Where $TPFP_0 := TP(0) + FP(0)$, $TPFN_0 := TP(0) + FN(0)$, $TNFP_0 := TN(0) + FP(0)$, and $TNFN_0 := TN(0) + FN(0)$. As shown above, the F1 score and MCC assumes that the threshold θ is fixed as zero. So, in order to improve them, the prediction function f must be learned to classify proteins correctly at $\theta = 0$. On the other hand, in order to improve AUC and AUC-10%, the prediction function f must be learned to classify proteins correctly for changing θ .

Data availability

All the data are provided as Supplementary Data. The software is available at <https://github.com/takeuchi-lab/allerStat>.

Supporting information—This article contains supporting information (44).

Author contributions—N. T., R. A., K. K., and I. T. conceptualization; K. G., T. Y., H. H., T. S., and I. T. methodology; N. T., R. A., and K. K. resources; K. G., N. T., T. Y., H. H., and K. K. visualization; K. G., T. Y., H. H., T. S., and I. T., software; K. G., N. T., H. H., K. K., and I. T. writing-original draft.

Funding and additional information—This work was supported in part by the following funds:

- The Ministry of Health, Labor and Welfare grant H30-Syokuhin-Ippan-002 (N. T., K. K., I. T.),
- The Ministry of Health, Labor and Welfare grant 21KA1002 (K. K.),
- MEXT KAKENHI grant 20H00601 (I. T.),
- JST Moonshot R&D grant JPMJMS2033-05 (I. T.),
- NEDO grants JPNP18002, JPNP20006 (I. T.),
- JST CREST JPMJCR21D3, JPMJCR22N2 (I. T.), and
- RIKEN Center for Advanced Intelligence Project (I. T., H. H.).

Conflicts of interest—The authors declare that they have no conflicts of interest with the contents of this article.

Abbreviations—The abbreviations used are: ASP, allergen-specific patterns; AUC, area-under-curve; BlastP, Protein Blast; CV, Cross-validation; FET, Fisher Exact Test; FP, false positive; FWER, family-wise error rate; HLA, human leukocyte antigen; IgE,

immunoglobulin E; i.i.d., independently identically distributed; LCO-CV, leave-category-out cross-validation; MCC, Matthews Correlation Coefficient; ML, machine learning; mTEC, medullary thymic epithelial cell; PoC, proof-of-concept; ROC, receiver operating characteristic; SVM, support vector machine; WY, Westfall-Young.

References

1. Barry Kay, A. (2001) Allergy and allergic diseases. *New Engl. J. Med.* **344**, 30–37
2. Kagan, R. S. (2003) Food allergy: an overview. *Environ. Health Perspect.* **111**, 223–225
3. Gupta, R. S., Springston, E. E., Warrier, M. R., Smith, B., Kumar, R., Pongracic, J., *et al.* (2011) The prevalence, severity, and distribution of childhood food allergy in the United States. *Pediatrics* **128**, e9–e17
4. Menon, A. K., Holowka, D., Webb, W. W., and Baird, B. (1986) Cross-linking of receptor-bound ige to aggregates larger than dimers leads to rapid immobilization. *J. Cel. Biol.* **102**, 541–550
5. Mongini, P. K., Paul, W. E., and Metcalf, E. S. (1982) T cell regulation of immunoglobulin class expression in the antibody response to trinitrophenyl-ficoll. evidence for t cell enhancement of the immunoglobulin class switch. *J. Exp. Med.* **155**, 884–902
6. Klein, L., Robey, E. A., and Hsieh, C.-S. (2019) Central cd4⁺ t cell tolerance: deletion *versus* regulatory t cell differentiation. *Nat. Rev. Immunol.* **19**, 7–18
7. FAO/WHO (2001) Report of a joint fao/who expert consultation on allergenicity of foods derived from biotechnology. *Eval. Allergen. Genet. Modified Foods*, 12–13
8. Ladics, G. S., Cressman, R. F., Herouet-Guicheney, C., Herman, R. A., Privalle, L., Song, P., *et al.* (2011) Bioinformatics and the allergy assessment of agricultural biotechnology products: industry practices and recommendations. *Regul. Toxicol. Pharmacol.* **60**, 46–53
9. Nicola, J. S., Ghantous, H. N., Ladics, G. S., House, R. V., Gendel, S. M., and Hastings, K. L. (2013) Workshop proceedings: challenges and opportunities in evaluating protein allergenicity across biotechnology industries. *Int. J. Toxicol.* **32**, 4–10
10. Stadler, M. B., and Stadler, B. M. (2003) Allergenicity prediction by protein sequence. *FASEB J.* **17**, 1141–1143
11. Riaz, T., Ley Hor, H., Krishnan, A., Tang, F., and Li, K.-B. (2005) Weballergen: a web server for predicting allergenic proteins. *Bioinformatics* **21**, 2570–2571
12. Li, K.-B., Issac, P., and Krishnan, A. (2004) Predicting allergenic proteins using wavelet transform. *Bioinformatics* **20**, 2572–2578
13. Martinez Barrio, A., Soeria-Atmadja, D., Anders, N., Gustafsson, M. G., Hammerling, U., and Bongcam-Rudloff, E. (2007) Evaller: a web server for *in silico* assessment of potential protein allergenicity. *Nucl. Acids Res.* **35**, W694–W700
14. Soeria-Atmadja, D., Lundell, T., Gustafsson, M. G., and Hammerling, U. (2006) Computational detection of allergenic proteins attains a new level of accuracy with *in silico* variable-length peptide extraction and machine learning. *Nucl. Acids Res.* **34**, 3779–3793
15. Zhang, L., Huang, Y., Zou, Z., He, Y., Chen, X., and Tao, A. (2012) Sortaller: predicting allergens using substantially optimized algorithm on allergen family featured peptides. *Bioinformatics* **28**, 2178–2179
16. Saha, S., and Raghava, G. P. S. (2006) Allpred: prediction of allergenic proteins and mapping of ige epitopes. *Nucl. Acids Res.* **34**, W202–W209
17. Fournier-Viger, P., Gomariz, A., Campos, M., and Thomas, R. (2014) Fast vertical mining of sequential patterns using co-occurrence information. In *Pacific-asia Conference on Knowledge Discovery and Data Mining*, Springer Cham, Cham, Switzerland: 40–52
18. Yang, Z., and Kitsuregawa, M. (2005). In *Lapin-spam: an improved algorithm for mining sequential pattern*, IEEE, Los Alamitos, CA, 1222–1222
19. Ayres, J., Flannick, J., Gehrke, J., and Yiu, T. (2002) Sequential pattern mining using a bitmap representation. In *Proceedings of the Eighth ACM*

- SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY: 429–435
20. Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., *et al.* (2004) Mining sequential patterns by pattern-growth: the prefixspan approach. *IEEE Trans. knowledge Data Eng.* **16**, 1424–1440
21. Zaki, M. J. (2001) Spade: an efficient algorithm for mining frequent sequences. *Machine Learn.* **42**, 31–60
22. Fu, T. C. (2011) A review on time series data mining. *Eng. Appl. Artif. Intelligence* **24**, 164–181
23. Srikant, R., and Agrawal, R. (1996) Mining sequential patterns: generalizations and performance improvements. In International conference on extending database technology, Springer, Berlin, Germany: 1–17
24. Fournier-Viger, P., Lin, J. C. W., Uday Kiran, R., Koh, Y. S., and Thomas, R. (2017) A survey of sequential pattern mining. *Data Sci. Pattern Recognition* **1**, 54–77
25. Benjamini, Y., and Yekutieli, D. (2005) False discovery rate–adjusted multiple confidence intervals for selected parameters. *J. Am. Stat. Assoc.* **100**, 71–81
26. Leeb, H., and Pötscher, B. M. (2005) Model selection and inference: facts and fiction. *Econometric Theor.*, 21–59
27. Leeb, H., Pötscher, B. M., *et al.* (2006) Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Stat.* **34**, 2554–2591
28. Benjamini, Y., Heller, R., and Yekutieli, D. (2009) Selective inference in complex research. *Philosophical Trans. R. Soc. A: Math. Phys. Eng. Sci.* **367**, 4255–4271
29. Pötscher, B. M., Schneider, U., *et al.* (2010) Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electron. J. Stat.* **4**, 334–360
30. Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., *et al.* (2013) Valid post-selection inference. *Ann. Stat.* **41**, 802–837
31. Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014) A significance test for the lasso. *Ann. Stat.* **42**, 413
32. [preprint] Taylor, J., Lockhart, R., Tibshirani, R. J., and Tibshirani, R. (2014) Post-selection adaptive inference for least angle regression and the lasso. *arXiv* **354**. <https://doi.org/10.48550/arXiv.1401.3889>
33. Suzumura, S., Nakagawa, K., Umez, Y., Tsuda, K., and Takeuchi, I. (2017) Selective inference for sparse high-order interaction models. In , **70**. *Proceedings of the 34th International Conference on Machine Learning*, Journal of Machine Learning Research: 3338–3347
34. Tanizaki, K., Hashimoto, N., Yu, I., Hontani, H., and Takeuchi, I. (2020). In *Computing valid p-values for image segmentation by selective inference*, IEEE, Los Alamitos, CA: 9553–9562
35. Le Duy, V. N., Toda, H., Sugiyama, R., and Takeuchi, I. (2020) Computing valid p-value for optimal changepoint by selective inference using dynamic programming. In *Advances in Neural Information Processing Systems* **33**. Curran Associates, Inc., Red Hook, NY: 11356–11367
36. [preprint] Le Duy, V. N., Iwazaki, S., and Takeuchi, I. (2020) Quantifying statistical significance of neural network representation-driven hypotheses by selective inference. *arXiv*. <https://doi.org/10.48550/arXiv.2010.01823>
37. Sugiyama, K., Le Duy, V. N., and Takeuchi, I. (2021) More powerful and general selective inference for stepwise feature selection using homotopy method. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research: 9891–9901
38. Reynisson, B., Barra, C., Kaabinejad, S., Hildebrand, W. H., Peters, B., and Nielsen, M. (2020) Improved prediction of mhc ii antigen presentation through integration and motif deconvolution of mass spectrometry mhc eluted ligand data. *J. Proteome Res.* **19**, 2304–2315
39. Matsuo, H., Yokooji, T., and Taogoshi, T. (2015) Common food allergens and their ige-binding epitopes. *Allergol. Int.* **64**, 332–343
40. Bailey, T. L., and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymer. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, American Association for Artificial Intelligence, Palo Alto, California: 28–36
41. Dang, H. X., and Lawrence, C. B. (2014) Allerdicator: fast allergen prediction using text classification techniques. *Bioinformatics* **30**, 1120–1128
42. Dimitrov, I., Bangov, I., Flower, D. R., and Doytchinova, I. (2014) Allertop v.2 — a server for in silico prediction of allergens. *J. Mol. Model.* **20**, 1–6
43. Terada, A., Kim, H., and Sese, J. (2015) High-speed westfall-young permutation procedure for genome-wide association studies. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, Association for Computing Machinery, New York, NY: 17–26
44. Terada, A., Tsuda, K., and Sese, J. (2013) Fast westfall-young permutation procedure for combinatorial regulation discovery. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, IEEE, Los Alamitos, CA: 153–158
45. Terada, A., Okada-Hatakeyama, M., Tsuda, K., and Sese, J. (2013) Statistical significance of combinatorial regulations. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12996–13001
46. Pellegrina, L., and Vandin, F. (2018) Efficient mining of the most significant patterns with permutation testing. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery, New York, NY: 2070–2079
47. Llinares-López, F., Sugiyama, M., Papaxanthos, L., and Borgwardt, K. (2015) Fast and memory-efficient significant pattern mining via permutation testing. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY: 725–734
48. Westfall, P. H., and Stanley Young, S. (1993) *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Vol. 279). John Wiley & Sons, New York, NY
49. Pino-Yanes, M., Corrales, A., Acosta-Herrera, M., Pérez-Rodríguez, E., Cumpido, J., Campo, P., *et al.* (2014) Hla-drb1*15:01 allele protects from asthma susceptibility. *J. Allerg. Clin. Immunol.* **134**, 1201–1203
50. Mishra, M. N., Dudeja, P., and Gupta, R. K. (2014) Association of hla-class ii and ige serum levels in pediatric asthma. *Iranian J. Immunol.* **11**, 21–28
51. Dimitrov, I., and Doytchinova, I. (2016) Associations between milk and egg allergens and the hla-drb1/dq polymorphism: a bioinformatics approach. *Int. Arch. Allergy Immunol.* **169**, 33–39
52. Dimitrov, I., Naneva, L., Doytchinova, I., and Bangov, I. (2013) AllergenFP: Allergenicity prediction by descriptor fingerprints. *Bioinformatics* **30**, 846–851
53. Sharma, N., Patiyal, S., Dhali, A., Pande, A., Arora, C., and Raghava, G. P. S. (2020) AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Brief. Bioinform.* **22**, bbaa294
54. Pinto, S., Michel, C., Schmidt-Glenewinkel, H., Harder, N., Rohr, K., Wild, S., *et al.* (2013) Overlapping gene coexpression patterns in human medullary thymic epithelial cells generate self-antigen diversity. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E3497–E3505
55. Skogberg, G., Lundberg, V., Berglund, M., Gudmundsdottir, J., Telemeo, E., Lindgren, S., *et al.* (2015) Human thymic epithelial primary cells produce exosomes carrying tissue-restricted antigens. *Immunol. Cell Biol.* **93**, 727–734
56. Christopher Bishop, M. (2006) *Pattern Recognition and Machine Learning*. Springer-Verlag New York, New York, NY