NAIST-IS-MT1651001

修士論文

配列類似度に基づいたアレルゲンネットワーク分析

青山泰枝

2018 年 3 月 15 日

奈良先端科学技術大学院大学
情報科学研究科

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士(工学) 授与の要件として提出した修士論文である。

青山泰枝

審査委員：

金谷 重彦教授	(主指導教員)
笠原 正治教授	(副指導教員)
MD.ALTAF-UL-AMIN准教授	(副指導教員)
小野 直亮准教授	(副指導教員)
黄 銘助教	(副指導教員)

配列類似度に基づいたアレルゲンネットワーク分析*

青山泰枝

内容梗概

アレルギーは先進国に多くの患者を持つ疾患である．アレルゲンの分類はアレルギーの起こる仕組みの解明に役立ち，アレルゲンの予測は新規の食品の開発に役立つ．近年の研究でアレルゲンは，進化的に近いタンパク質のファミリーに分類されるが，このファミリーの分類はアレルギーの交差反応の予測には不十分である．アレルゲンの予測のための機械学習は，手法の複雑化に伴って，時間的コストが高いものが多くなりつつある．これらの問題解決のため，本研究では局所的な配列類似度に基づくアレルゲンネットワークの構築を提案する．このネットワークにDPClusアルゴリズムを適用することによって，類似性の高い配列のみからなるクラスターを得ることができる．これによりタンパク質のファミリーをより細かく分類し，交差反応性をより起こしやすいグループを作成できる．また，各クラスターから代表配列を抽出したモデルは，少ない配列で，既存のアレルゲンの特徴を表現でき，機械学習の時間コストを下げるのに貢献できる．また，本研究によって得られたクラスターの分析により，植物のアレルゲンからなるクラスターに動物のアレルゲンが入ることがほとんどなく，これらのアレルゲンは構造的に異なることがわかった．各クラスターは共通のタンパク質の機能を持つこともわかった．

キーワード

アレルゲン，バイオインフォマティクス，DPClus algorithm，ネットワーク解析

*奈良先端科学技術大学院大学 情報科学研究科 修士論文,

NAIST-IS-MT1651001, 2018 年3 月15 日.

Multifaceted analysis of sequence similarity-based network of allergens*

Yasue Aoyama

Abstract

Abstract

Allergy is one of the most common diseases in developed countries. Classification of allergen is helpful for understanding the immune system. Prediction of allergen is useful for production of food. Allergens are classified into different groups based on their evolutionally related groups. However, the conventional classification approach is insufficient for allergen cross-reactivity prediction. The prediction of allergen by machine learning method becomes time consuming as the prediction method becomes complex. In order to solve these problems, we construct allergen networks depending on local sequence similarities. The allergens are classified into small clusters so that members in the same cluster are densely connected using DPClus algorithm. This research shows that the problems about allergen prediction may be improved by (1) developing more detailed classification method, (2) constructing small allergen model set that will represent allergens property more precisely. Our clustering results show that most of clusters contain either plant proteins or animal proteins solely. Each cluster can be characterized by one protein function/domain. Based on these insights, we proposed a new allergen prediction model from the clusters.

Keywords:

Allergen, Bioinformatics, DPClus algorithm, Network analysis

---

*Master's Thesis, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1651001, March 15, 2018.

Index

Figure index

Chapter 1 Introduction

1.1 Explanation of allergy

Allergy is one of the most common diseases in developed countries [1], which is caused by inappropriate immune responses. Allergenicity is the potential of any material to cause sensitization and allergic reaction and is frequency associate with the Immunoglobulin E (IgE) antibody [2]. Over the past decade, basic research has intensified in an attempt to better understand the molecular-biological mechanisms leading to inflammation reactions [3-4]. However, the reason why allergens can cause T-cell and B-cell responses remains largely unanswered [5]. Figure 1 shows schematic illustration of the mechanism of allergic diseases.

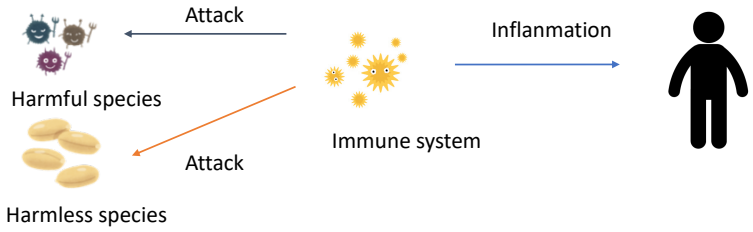

Figure 1 Schematic illustration of the mechanism of allergic diseases

The immune system is a host defense system comprising many biological structures and processes within an organism that protects against disease. If a host is infected by hazardous species (i.e. virus, bacteria, parasite), inflammation is triggered. The immune system detects the infection of hazardous species by detecting non-self proteins. Immune tolerance, or immunological tolerance, or immunotolerance, is a state of unresponsiveness of the immune system to substances or tissue that have the capacity to elicit an immune response in given organism. Proteins included in foods are non-self proteins. But inflammation is not triggered by food proteins because immune tolerance systems is working proper. Immune tolerance systems are not working proper in people with a predisposition to allergies.

Metal allergy and drug allergy are also major allergies. Metals itself does

not cause allergic reaction. But metal ions bind to human proteins and change the structure of binding proteins. A drug molecule is a low molecular weight compound. A drug molecule itself is not able to cause allergic reaction, but it binds to a human protein and change the structure of binding protein. These structurally changed human proteins can be allergens. Therefore, allergic reactions are triggered by a variety of substances such as proteins, metals and low-molecular compounds. But allergic reaction is triggered by foreign proteins or structurally changed proteins. In these reasons, we are paying attention to allergen protein sequences. Immune tolerance is important for allergenicity. But the number of proteins known to be allergen is very few. Swiss-prot [6], a major protein database contains 556,388 proteins, but the numbers of allergens included in AllergenOnline database are 2034 [7]. Analyzing and classification of allergens can help to prove allergy mechanism.

Allergens include proteins in food, air, metals, and so on. Avoiding allergic materials is very important for the people with a predisposition to allergies. Therefore, assessment of the potential allergenicity of proteins is important for food, and drug production.

1.2 Immune system

Before explain about allergy, we will explain about immune system, which is important to understand the mechanism of allergy. Immune system is the major defense system, which will recognize foreign proteins as hazardous substances and then try to expel or neutralize them before they can do harm. IgE are one kind of antibodies to detect pollen and parasite proteins. Among numerous types of IgEs. there is a type of IgE binding to specific proteins with common protein surface structure. Although the IgE is the rarest type of immunoglobulins among the major Ig, it has a high blood concentration in Allergy patients.

When hazardous species enter to a human body, Immune systems recognize and attack hazardous species in the ways shown below (Figure 2). (1) Macrophages, a kind of phagocytes that patrol the body, catch hazardous species. (2) A fragment of protein of hazardous species is send to a helper T-cell, which then transform to th1 cell or th2 cell in response to the specific type of the hazardous specie. (3-1) Th1 cell activate cell-mediated immunity. As a result, killer T-cells attack the cells which are polluted by hazardous species. Once cell-mediated immunity is activated, Th1 cells remain in the body and prepare for second or later hazardous species invasion. (3-2) Th2 cell activate humoral immunity. As a result, B-cells make IgE antibodies which attack proteins of hazardous species. Once humoral immunity is activated, Th2 cells remain in the body and prepare for second or later hazardous species invasion. Remained Th1 or Th2 cells can attack its hazardous species skipping (1) ~ (3) steps. In these ways, we are able to combat hazardous species very fast and effectively.

As mentioned in 1.1, immune tolerance system works to avoid inflammation to harmless proteins, but this system does not work well in people with a predisposition to allergies. As a result, inflammatory response is caused when harmless proteins enter to the patient's body. In addition, allergy patient's Th2 cells are hyperactive.

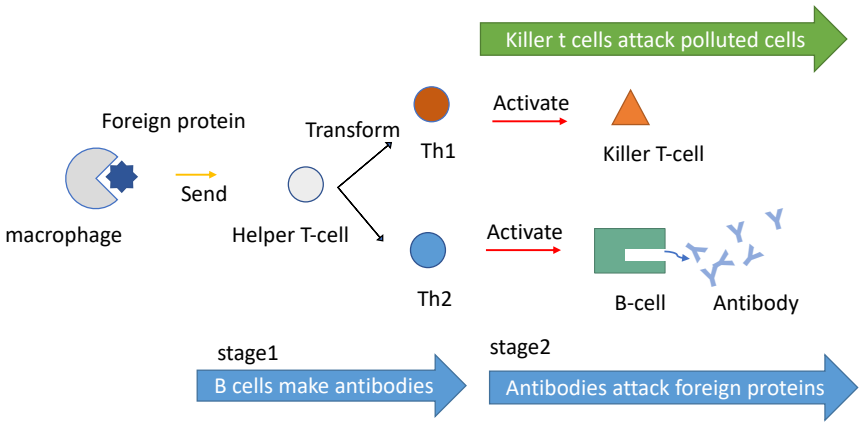

Figure 2 Schematic illustration of immune system

1.3 Defining allergens

Allergens represent a small fraction of the proteins that humans are frequently exposed to. Proteins are large biomolecules, or macromolecules, consisting of one or more long chains of amino acid residues (Figure 3). The reason why allergic proteins can cause allergic reaction remains largely unanswered. The number of allergens is very small compared to total number of proteins. Not all proteins in allergic material are allergens. We assume that we can find new things by analyzing allergens.

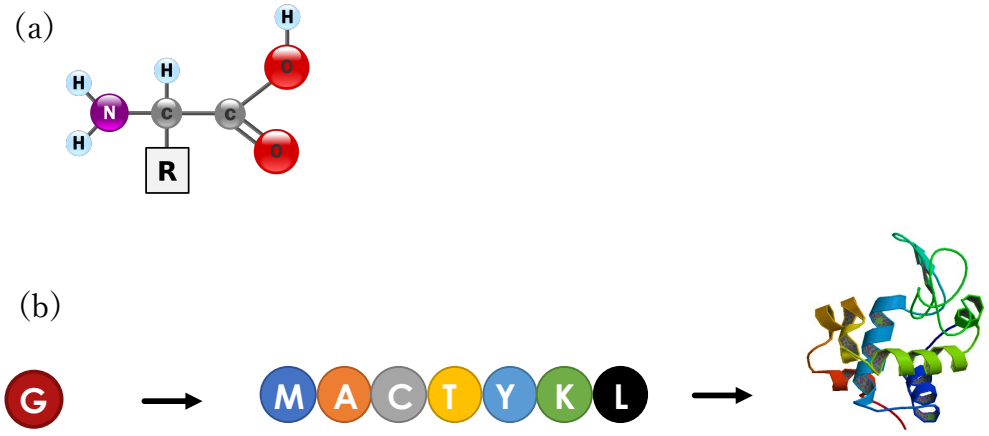

Figure 3 The structure of amino acid and protein

(a) **Molecular structure of an amino acid:** R denotes a side chain that characterizes each amino acid. There are 20 standard amino acids. (b) A protein is a long chain of amino acids. An amino acid is represented as an alphabet. A sequence of amino acid in a protein is represented by an arrangement of alphabets. A protein structure depends on its amino acid sequence. The 3D structure is important for protein function. But information about protein structure is included in amino acid sequence information.

1.4 Allergen nomenclature

The official nomenclature of allergenic proteins is based on the Linnaean binominal nomenclature identifying genus and species of all organisms and was first published in 1986 [8] and revised in 1994 [9-13]. The allergen nomenclature is maintained by the IUIS Allergen Nomenclature Sub-Committee under the auspices of the World Health Organization (WHO) and the International Union of Immunological Societies (IUIS). The committee maintains the database of approved allergen names (www.allergen.org). Allergen names are composed of an abbreviation of the scientific name of its source (genus: 3–4 letters; species: 1–2 letters) and an Arabic numeral, for example Der p 1 for the first allergen to be described from the house dust mite *Der matophagoides pteronyssinus*.

1.5 Allergen domain

Protein sequences are the fundamental determinants of biological structure and function. One of the most important information related to protein function is the protein domain. A protein domain is a spatially localized unit of the protein structure, and there is often a specific function associated with it. Most of proteins have one characteristic domain, while some proteins have more than one domains. Protein family is a group of evolutionally related proteins. Each members of a family have high sequence similarity and highly similar function. It has been reported that most allergens belong to a limited number of protein families [14]. And the protein sequence and its domain have something to do with allergic reaction.

1.6 Allergen prediction

In clinical view, analyzing and classification of allergen is important because it is help of knowing the pattern of allergy. Prediction of potential allergen is important for food and drug production. A variety of machine learning methods have been developed to predict potential allergens. The United Nations Food and Agriculture Organization (FAO) and the World Health Organization (WHO) have developed guidelines for evaluating the potential allergenicity of novel food proteins [15-16]. According to these guidelines, a protein is a potential allergen if it has either an identity of 6 to 8 consecutive amino acids or greater than 35 % overall sequence similarity over a window of 80 amino acids when compared with known allergens [17]. This method has high sensitivity, but it has very low specificity [18].

Transformation of protein sequence to numeric vector is frequently used in allergen prediction. Support vector machine using amino acid, dipeptide and tripeptide composition was developed to predict allergen [19,20]. Amino acid chemical property was used for prediction [21].

1.7 Purpose

In some resent researches, allergens have been classified into evolutionally related protein groups (protein families) [14]. However, not all members in an allergen family are cross-reactive, which means that the classification is not enough to predict cross-reactivity.

In most machine learning methods, amino acid sequences of allergen proteins are characterized by a vector of numeric features of the same length, in order to avoid the difficulty to compare the sequence of allergen proteins with various length of amino acid polymers. Most of such transforming processes are computationally quite time consuming. The number of allergens is increasing because technologies to specify allergens are developing. Allergen prediction become more difficult as the number of allergens increases.

Our purpose is to provide clues for these problems. In this research, we constructed a network of allergens depending on local similarity and classified allergens into clusters using DPClus algorithm. DPClus algorithm is a graph-clustering algorithm that can be used to extract densely connected vertexes as a cluster. In this method, members of a protein family are able to be separated.

In addition, we developed an allergen sequence set by choosing one representative allergen from each cluster. The representative sequence sets are helpful to mitigate the time consuming problem in allergen prediction.

To prove that our model is useful for allergen prediction, we predicted potential allergenicity of query proteins. The calculation process was displayed in Figure 4.

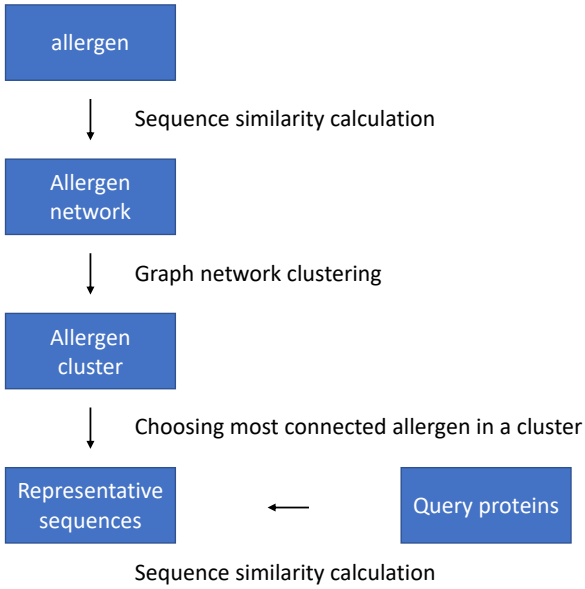

Figure 4 Flow chart of calculations in this thesis

The calculation process was displayed in the flow chart.

1.8 Introduction of BLAST

BLAST (Basic Local Alignment Search Tool) is an algorithm that calculates similarity between two sequences [22]. Local similarity within two sequences is calculated. A query protein sequence is compared with proteins in a database (Figure 5). As shown in Figure 6, two sequences are aligned to maximize alignment score. The Expect value (*E*-value) is an index that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. The lower the E-value, or the closer it is to zero, the more "significant" the match is. Alignment score is a parameter that describes similarity between two sequences. This requires scoring matrix (Figure 7), or a table of values that describes the probability of a biologically meaningful amino-acid or nucleotide residue-pair occurring in an alignment. Scores for each position are obtained frequencies of substitutions in blocks of local alignments of protein sequences.

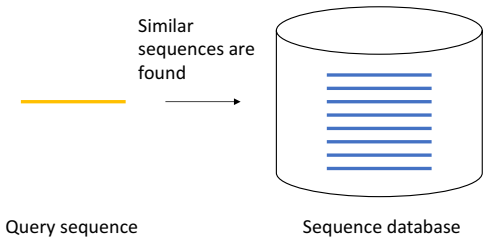

Figure 5 Explanation of BLAST algorithm

A query sequence is compared with all sequences in a database. The output includes the sequence names which have high similarity with a query sequence, alignment scores.

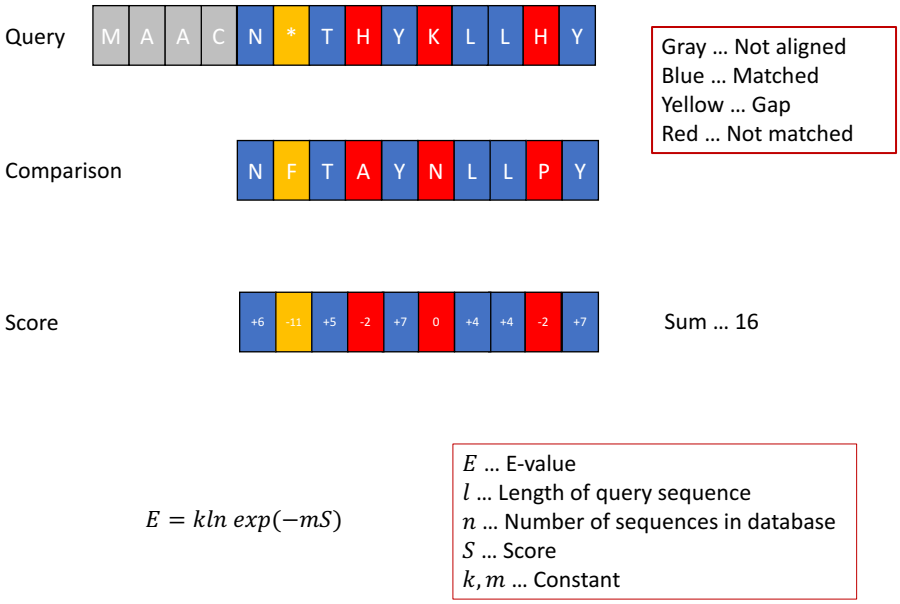

Figure 6 Explanation of alignment score

The alignment score denotes the sum of aligned amino acid substitution scores. If a pair of amino acid are chemically similar, substitution score is high. If that's not the case, matching score is low. The substitution score matrix is shown in Figure 7.

E-value is an error rate of an alignment result. Small E-value is credible. *E*-value is calculated from the length of query sequence, the number of sequence in database and the alignment score.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 7 BLOSUM 62 Substitution matrix

Alignment score is calculated by this matrix.

1.9 Introduction of DPClus

DPClus algorithm is a graph-clustering algorithm that can be used to extract densely connected vertexes as a cluster [23, 24]. This algorithm can be applied to an undirected simple graph $G = (N, E)$ that consists of a finite set of vertexes N and a finite set of edges E. Two important parameters are used in this algorithm (i.e., density d and cluster property cp). Density dk of any cluster k is the ratio of the number of edges present in the cluster (E) to the maximum possible number of edges in the cluster (E_{max}). The cluster property of a vertex n with respect to cluster k is represented as

$$cp_{nk} = \frac{E_{nk}}{d_k \times N_k}$$

where N_k is the number of vertexes in k and E_{nk} is the total number of edges between n and each vertex of k. Simple example of cp calculation is shown in Figure 8.

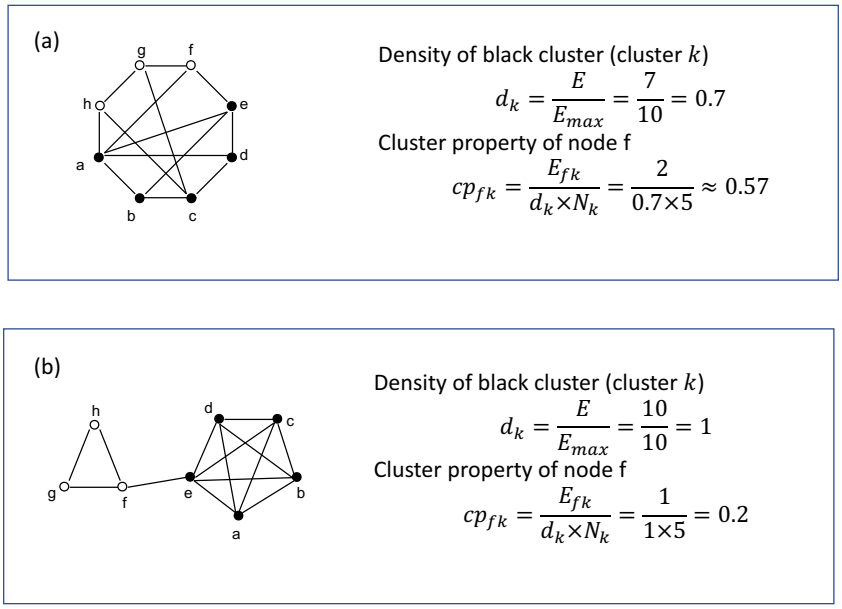

Figure 8 Examples of cluster property calculation

The number of cluster in a graph is decided by density and cluster property (*cp*). Density of graph (a) and graph (b) are both 0.8. However, graph (a) seems to consist of one cluster, and graph (b) seems to consist of two clusters.

A cluster to be belonged is decided by *cp* (cluster property). When *cp* of vertex f with respect to black cluster (cluster k) is higher than a threshold, vertex f belongs to cluster k. When *cp* threshold is 0.5, vertex f of graph (a) belongs to cluster k, but vertex f of graph (b) does not belongs to cluster k. As a result, graph (b) consists of two clusters.

1.10 Cross-reactivity of allergen

It is known that IgE recognize similar protein having similar structure as the antigen. This phenomenon is called cross-reactivity(Figure 9). Allergy cross-reactivity is difficult to find because allergen cross-reactivity works usually, but not always. The cross-reactivity of IgE is important for some reasons. Firstly, from clinical viewpoint, it is important to know the patterns of cross-reactivity. Secondly, prediction of cross-reactivity is important for novel food and drug production. Homology in allergen sequences results in similar 3D structure and potentially, in cross-reactivity. Homologous allergens from phylogenetically related grasses tend to be cross-reactive [25]. IgE antibodies to allergens from peanuts are often cross-reactive with homologous proteins in soybean and other legumes.

However, cross-reactivity between much more distantly related organism have also been known. Some examples of them are ragweed/banana [25], birch/apple [25], profilin [25], mugwort/celergy [25], mite/snail/shrimp [25]. They are distantly related organism, but their allergens are similar. Calculating similarities between allergens are important to indentify the unkown cross-reactivity.

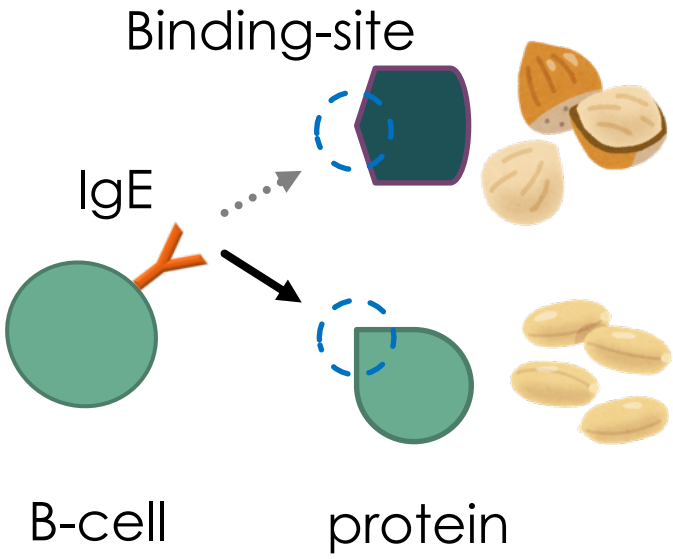

Figure 9 Systematic illustration of cross-reactivity

Antibodies bind to specific allergens and structural similar protein with the antibodies. This is called cross-reactivity. For example, IgE in the figure is specified to peanut allergens (below). The IgE can bind to similar proteins with peanut allergens(above).

Chapter 2 Materials and Methods

2.1 Data preparation

The allergens were obtained from AllergenOnline [7] database (http://www.allergenonline.org). The network consists of 2034 allergen proteins. AllergenOnine data includes information about allergen species. The data example of AllergenOnline dataset is shown in Figure 10. The dataset for allergen prediction is obtained from AllerTOP [21] (http://www.ddg-pharmfac.net/AllerTOP). It has 2432 allergens and 2432 non-allergens. The non-allergens were collected from widely used food species: *Solanum lycopersicum* (tomato), *Capsicum annuum* (pepper), *Solanum tuberosum* (potato), *Triticum aestivum* (bread wheat) and *Oryza sativa* (Asian rice) and *Oryza glaberrima* (African rice) using searches in Swiss-Prot for proteins annotated with the label "evidence for the existence of protein—evidence at protein level". Proteins containing the keyword "allergen" in their description were excluded. Additionally, a set of non-allergens was collected from Swiss-Prot to include proteins from Homo sapiens, again labeled with "evidence for the existence of protein—evidence at protein level". The proteins with keywords "allergen" in their description, as well as proteins with unidentified amino acids in their sequences, were excluded [21]. There are some allergen sequences included in both of the datasets. Overlapping between two datasets is shown in Figure 11.

Species	Common	IUIS.Allergen	Type	Group.	Length	Accession
Acarus siro	Mite	Aca s 13	Aero Mite	Acarus Aca s 13	131	ABL09307.1

Figure 10 Data example of AllergenOnline data

Data Example of AllergenOnline data. Species means scientific name of allergen source. Common means common name of allergen source. IUIS Allergen means allergen name decided by IUIS. Type is decided depending on allergen source and the way of ingestion. Groups means similar allergen group determined by AllergenOnline database. Length means the length of protein, the number of amino acids. Accesssion means ID in NCBI database.

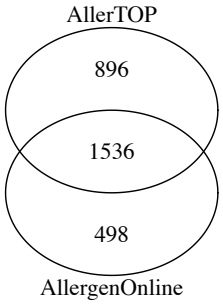

Figure 11 Overlapping between two data

The overlapping between two datasets are shown in the venn diagram.

2.1 Protein domain data collection

Information about allergen family was collected from NCBI's (National Center for Biotechnology Information) Conserved Domain Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) [26].

2.3 Network construction

A vertex in the allergen network is a protein. When a pair of allergen proteins have high sequence similarity, two vertexes are connected by an edge (Figure 12). Sequence similarity is calculated by BLASTP. Global alignment of AllergenOnline [25] allergen sequences was done. Allergen pairs which has E-value lower than 10^{-40} were selected. The numbers of allergen pairs which has high sequence similarity are 71227.

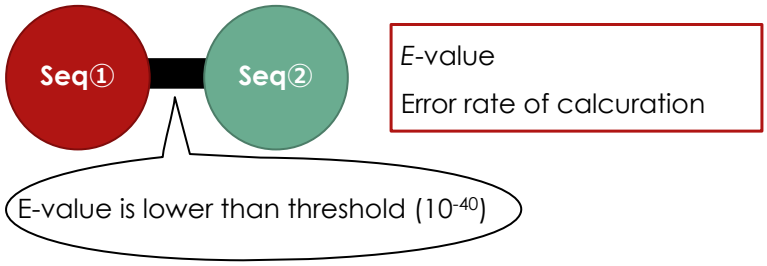

Figure 12 Illustration of network construction method

A vertex represents an allergen protein. Two vertexes are connected by an edge when the E-value of two sequences is lower than threshold. The allergen network is constructed by repeating this step.

2.4 Clustering

In the sequence similarity-based network, some sequences which has relatively low similarities might be connected, like Figure 13. To avoid the case, allergen sequences are classified into the clusters by DPClus. Allergens are classified into 209 allergen clusters. Cluster size is ranged from 2 to 158.

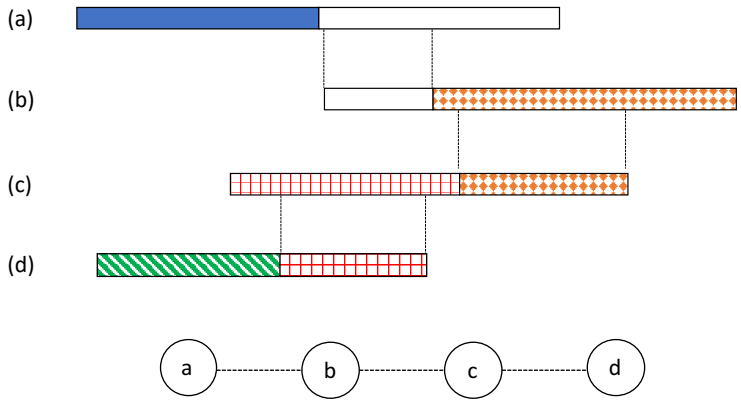

Figure 13 A problem of similarity-based network

(a)~(d) represent protein sequences. If two sequences have similar parts, two sequences are connected by break line. (a) and (d) is not similar each other. But (a) and (d) are connected in a network.

2.5 Allergen prediction

Most of allergens are attributed to clusters. One cluster consists of allergens which has high sequence similarities between each other. A protein which has sequence similarity with any known allergen is a potential allergen. We then proposed that by choosing one representative allergen from the each cluster, we could develop a smaller allergen model of known allergens.

2.6 Choosing representative sequences

We selected two types of representative sequences (Figure 14), that is, (1) one allergen from each cluster which has the most connections with other allergens in the cluster, and (2) isolated allergens in the network. For prediction purpose, we utilized 3 allergen networks using 3 different *E*-value thresholds (10^{-40}, 10^{-20} and 10^{-5}).

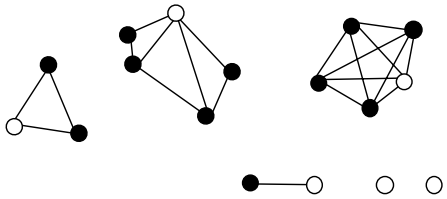

Figure 14 How to decide representative allergen

White vertexes are representative allergens. One allergen from each cluster which has most connection between allergens in the cluster and isolated allergen in the network are selected.

2.7 Prediction

In our method, a query protein is a potential allergen when its alignment *E*-value is lower than threshold value (10^{-1},10^{-5},10^{-10},10^{-20} and 10^{-40}) compared with representative allergens by BLASTP. The performance was measured as sensitivity and specificity. Sensitivity measures the ability to pick out true allergens. Specificity measures the ability to pick out non-allergens. Sensitivity and specificity is represented in the equation.

$$sensitivity = \frac{TP}{TP + FN}, \qquad specificity = \frac{TN}{TN + FP}$$

TP is a true positive, a true allergen with an *E*-value lower than the prediction threshold. TN is a true negative, a non-allergen with an *E*-value higher than the prediction threshold. FP is a false positive, a true-allergen with an *E*-value higher than the prediction threshold. FN is a false negative, a non-allergen with an *E*-value lower than the prediction threshold.

Chapter 3 Result

3.1 Network property

The overview of network is shown in Figure 15~17. A vertex in the allergen network is a protein. If a pair of allergen proteins have high sequence similarity, two vertexes are connected by an edge. The number of edges is 71227. We used 2034 allergen sequence to construct a network, but the total number of vertexes in the network is 1790. vertexes are separated into small groups. These groups construct almost perfect graph.

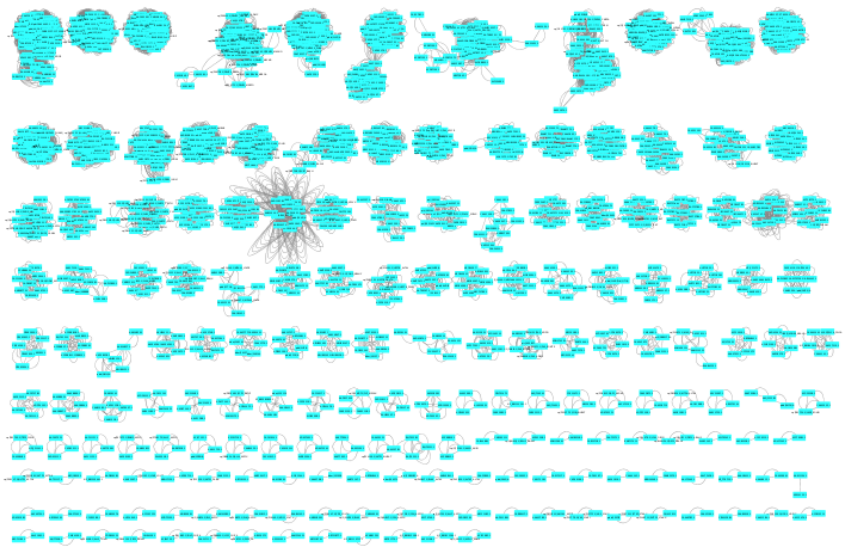

Figure 15 Sequence similarity-based network of allergen (1)

The E-value threshold to make network is 10^{-40}.

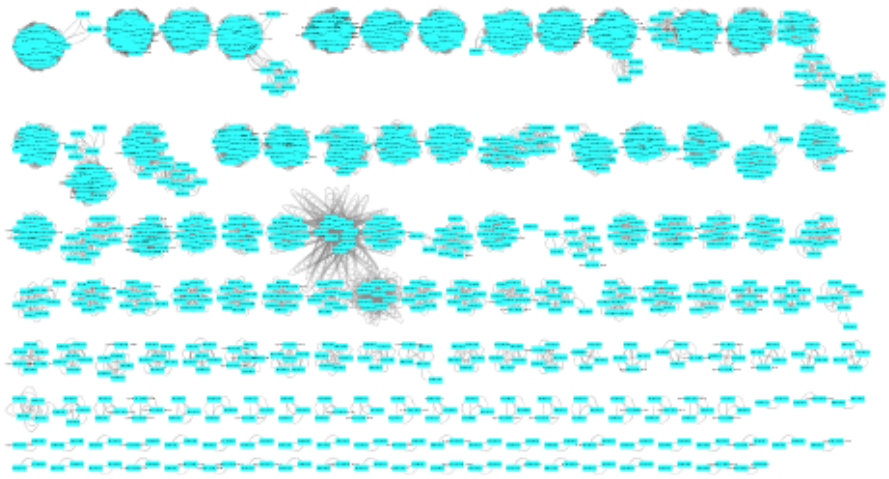

Figure 16 Sequence similarity-based network of allergen (2)

The E-value threshold to make network is 10^{-20}.

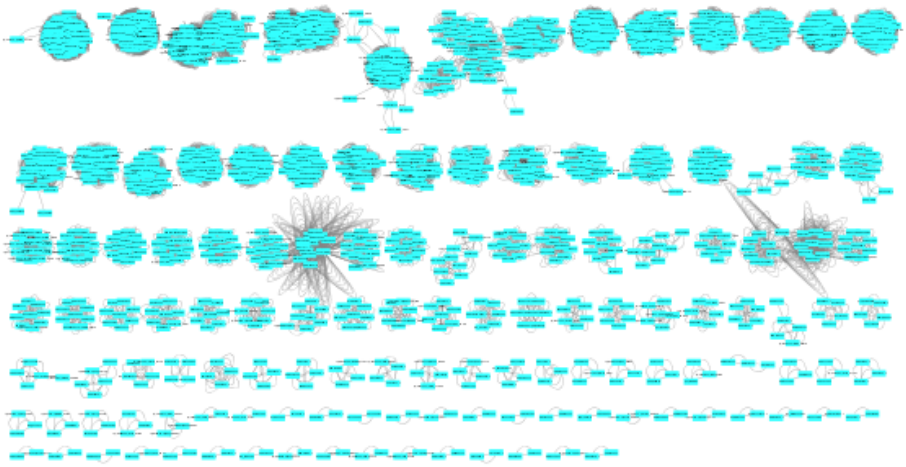

Figure 17 Sequence similarity-based network of allergen (3)

The E-value threshold to make network is 10^{-5}.

3.2 Clusters

We classified 1790 allergens into 209 clusters by DPClus algorithm. The size of clusters is ranged from 2 to 158. Most clusters are not connected with any other cluster, but some clusters are connected by one or two allergens. The members of a cluster are highly connected. It suggests that we could attribute allergens into groups of which each member has similar sequence. This may suggest that each member of a cluster have similar function and these functions may be related to allergenicity.

3.3 Representative sequences

For prediction purpose, we constructed 3 allergen networks using 3 different *E*-value thresholds (10^{-40}, 10^{-20} and 10^{-5}). The number of representative sequences corresponding to 3 networks are 495, 364 and 299. As *E*-value becomes lower, the numbers of edges decrease and the numbers of cluster increases.

3.4 Prediction

Representative sequences were compared with query sequences using BLASTP. The prediction sensitivity and specificity reached around 90%. The detailed result is shown in Figure 18.

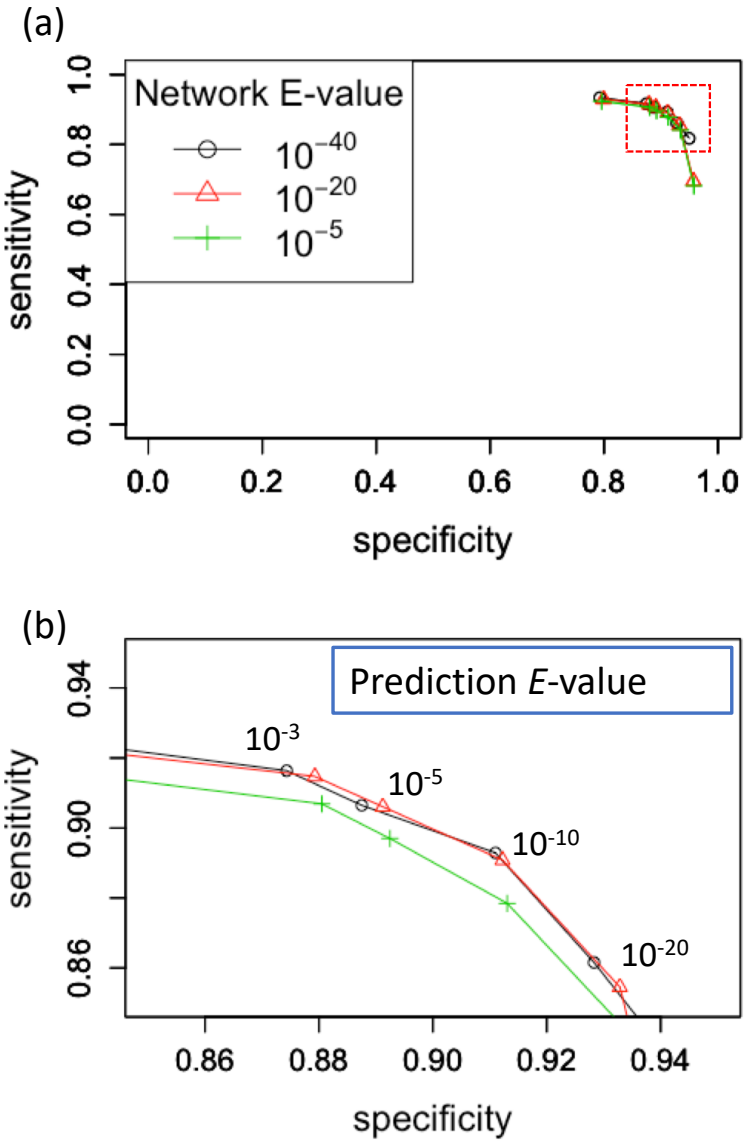

Figure 18 Result of allergen prediction

Each line shows results corresponding to an *E*-value. *E*-value change does not affect result so much. The best *E*-value for query prediction is $10^{-3} \sim 10^{-5}$.

Chapter 4 Discussion

4.1 Species Types of allergens

AllergenOnline dataset includes information about allergen source and allergen type. Allergen types are a classification of allergens according to allergen source and the way by which allergen transmits. They are 14 kinds of allergen types, "Aero Plant", "Food Plant", "Gliadin", "Contact", "Aero Animal", "Food Animal", "Aero Insect", "Worm (parasite)", "Aero Mite", "Aero Fungi", "Bacteria airway", "Bacteria skin", "Venom or Salivary", "Unassigned" [25]. "Contact" includes allergens of latex, fungus and yeast.

We inspect types distribution of allergens in each cluster. Examples of clusters are shown in Figure 19. Plant allergens are included in clusters that consist of plant allergens only. Species included in each cluster is shown in Figure 20. Only a few clusters include both plant and non-plant allergens. For example, cluster 58 includes plant, mite and fungi allergens, though the majority of cluster 58 is cyclophilin proteins. This kind of protein widely exists in many species including human. This result suggest that plant allergens and non-plant allergens are structurally different. But non-plant allergens are not completely different from each other.

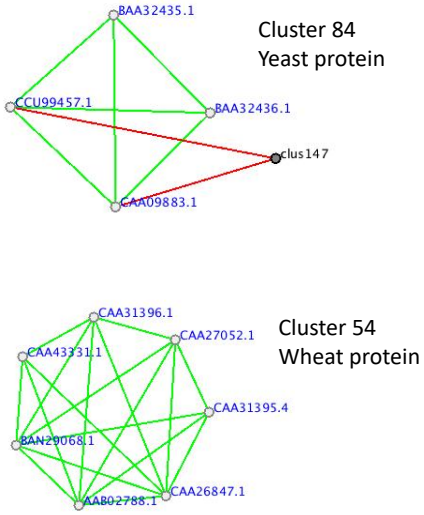

Figure 19 Samples of allergen clusters

Densely connected clusters are obtained from DPClus algorithm. Each cluster consists of allergens of similar species. For example, Cluster 84 is consisting of yeast allergens. Cluster 54 is consisting of wheat allergens.

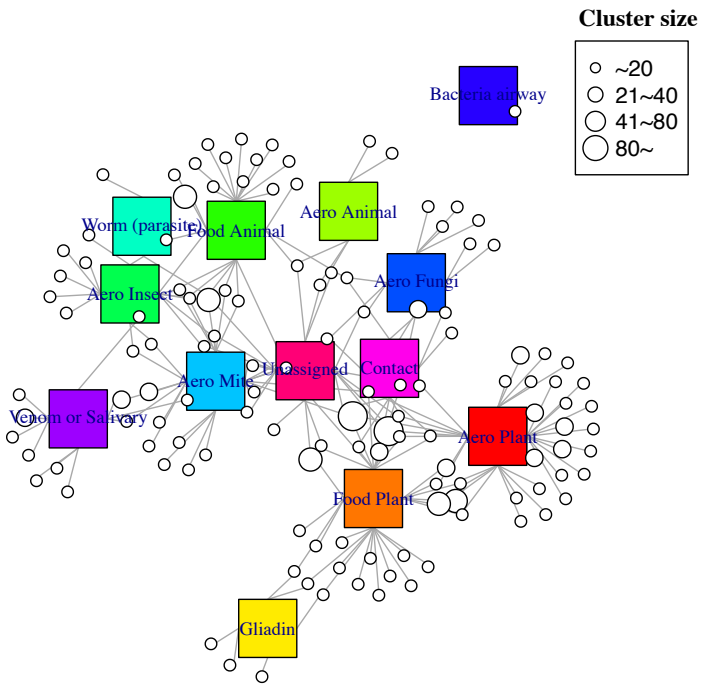

Figure 20 Allergen type distribution in cluster

Circles represent clusters, and rectangles represent allergen type. Cluster vertex and type vertex is connected when the cluster have at least one allergen of the type. “Gliadin”, “Food Plant” and “Aero Plant” are plant allergen types. “Bacteria Airway”, ”Aero Animal“, ”Food Animal”, “Worm”, ”Aero Insect”, “Aero Mite “ and “Venom or Salivary” are animal allergen types. “Contact” includes animal and plant allergens. One cluster consists only animal or plant allergens.

4.2 Allergen domain in a cluster

Protein domain information about protein is collected by NCBI’s Conserved Domain Database. Protein domain is related to protein function. If more than 90% of a cluster member has same domain, that domain represents the cluster. The relationship between cluster and domains are shown in Figure 21.

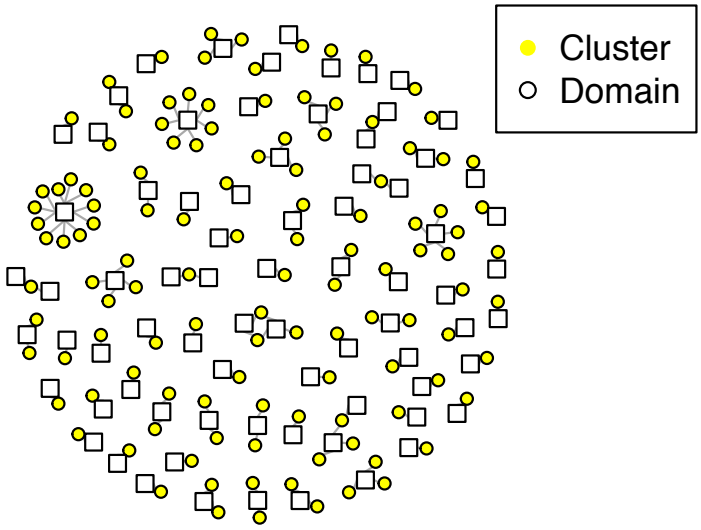

Figure 21 Representative domains in clusters

Circles represent clusters, and rectangles represent protein domains. Cluster vertex and domain vertex is connected when the cluster is representated by the domain (90% members of one domain has one domain). Most of clusters are represented by one cluster.

4.3 Cross-reactivity

IgE binds to some allergens. These allergens are characterized by similar local structures. We constructed sequence similarity-based network of allergen. After that, graph clustering has been done. We extracted 209 clusters. Each members of a cluster have high sequence similarity each other because *cp* is higher than threshold. We hypothesize that each members of a cluster are potentially cross-reactive. To explain that, we searched cross-reactivity of the clusters and some papers support the theory. *Penaeus aztecus* (Brown shrimp) is known to cross-reactive with mite, cockroach and crab. Pen a 1, which is an allergen of shrimp is known to be cross-reactive to Der p 10 (mite allergen), Per a 7 (cockroach allergen), Cha f 1 (crab allergen) [27]. These allergens are belonging to cluster 3. Many food plant allergens are known to cross-reactive with pollen allergens. Bet v 1, which is birch pollen allergen, belongs to cluster 1. Cluster 1 includes many

pollen allergens and food plant allergens. Mal d 1 (apple allergen), Gly m 4 (soybean allergen) and Pul c 1 (peach allergen) are known to be cross-reactive with Bet v 1 and belong to cluster 1[28]. Api g 1 (celery allergen) is known to be cross-reactive with Bet v 1[28]. Api g 1 belongs to cluster 22. But cluster 1 is connected to cluster 22 by some vertexes. As a result, it is said that allergens in same cluster or allergens in connected clusters are potentially cross-reactive.
Our clustering method can classify cross-reactive group into different clusters(Figure 22).

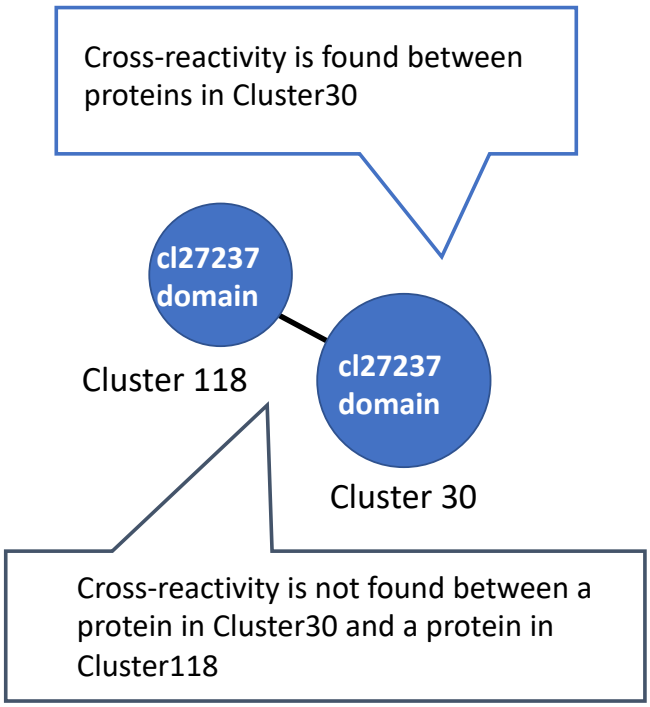

Figure 22 Protein domain and cross-reactivity

Our clustering method classified members of cross-reactive allergen group into a same cluster. Cluster 30 and cluster 118 are represented by one same domain, cl27237. Both cluster are connected by some proteins. Cross-reactivities within cluster 30 or cluster 118 are reported. But cross-reactivity between cluster30 and cluster 118 is not reported.

4.4 Prediction result analysis

The prediction sensitivity and specificity reached around 90%, that implies features of allergens are limited. For prediction purpose, we evaluated many *E*-values. The change of *E*-value to make network does not affect prediction result strongly. That means the number of representative sequences is enough. But *E*-value to predict affect prediction result very much. That means how similar query protein is with representative allergens is very important.

Conclusion

Similarity between allergens is very important in predicting allergens and predicting allergen cross-reactivity. We have developed similarity-based network of allergens. We developed small model of allergen which is useful for allergen prediction and cross-reactivity prediction. Comparing preceding methods of allergen models, which have already showed high predictivity, our model can show the utility of local sequence similarity in prediction. Many researchers continue to develop new allergen methods because allergen recognition mechanism is unanswered. The number of allergens is increasing because technologies to specify allergens are developing. In many machine learning method, allergen sequences should be transformed into same length numeric vectors because allergen proteins have different amino acid sequence length. Most of transforming process is time consuming. As allergen sequences increasing, the time of these process will increase. Our clustering method provided small model of allergens that represent allergen structural features. Allergens belonging to same protein family con divided by our clustering method. We developed allergen sequence model consists of 299~495 allergens from 2037 sequences. That help to solve time consuming problem in allergen prediction. Our small model will provide useful clues to understand those problems such a problem.

Acknowledgement

私の研究を熱心にご指導してくださった金谷重彦教授、ゼミナールの場において様々なアドバイスをいただいた親身になって研究の相談に乗っていただいた Md. Altaf-Ul-Amin 准教授，ゼミナールなどの場において鋭い指摘やアドバイスをいただいた小野直亮准教授，研究に関して様々なアドバイスをいただいた佐藤哲大客員准教授，黄銘助教授，研究室生活を支えてもらった金谷研究室の学生へ心から感謝の気持ちと御礼を申し上げたく、謝辞にかえさせていただきます。

Reference

[1]. Pawankar, R., Canonica, G. W., Holgate, S. T., & Lockey, R. F. (2011). World Allergy Organization (WAO) white book on allergy. *Wisconsin: World Allergy Organisation*.

[2]. Verhoeckx, K. C., Vissers, Y. M., Baumert, J. L., Faludi, R., Feys, M., Flanagan, S., ... & Wichers, H. (2015). Food processing and allergenicity. *Food and Chemical Toxicology*, *80*, 223-240.

[3]. Helm, R. M., & Burks, A. W. (2000). Mechanisms of food allergy. *Current Opinion in Immunology*, *12*(6), 647-653.

[4]. Kristiansen, M., Dhami, S., Netuveli, G., Halken, S., Muraro, A., Roberts, G., ... & Ansotegui, I. J. (2017). Allergen immunotherapy for the prevention of allergy: A systematic review and meta‐analysis. *Pediatric Allergy and Immunology*, *28*(1), 18-29.

[5]. Hayes, M., Rougé, P., Barre, A., Herouet-Guicheney, C., & Roggen, E. L. (2015). In silico tools for exploring potential human allergy to proteins. *Drug Discovery Today: Disease Models*, *17*, 3-11.

[6]. Pundir, S., Martin, M. J., & O'Donovan, C. (2017). Uniprot protein knowledgebase. *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*, 41-55.

[7]. Goodman, Richard E., et al. "AllergenOnline: A peer‐reviewed, curated allergen database to assess novel food proteins for potential cross‐reactivity." Molecular nutrition & food research 60.5 (2016): 1183-1198.

[8]. Marsh, D. G., Goodfriend, L., Lewenstein, H., & Platts-Mills, T. A. (1987). Allergen nomenclature. *Journal of allergy and clinical immunology*, *80*(5), 639-645.

[9]. King TP, Hoffman D, Lowenstein H, Marsh DG, Platts-Mills TA, Thomas W. Allergen nomenclature. Int Arch Allergy Immunol 1994;105:224–233.

[10]. King TP, Hoffman D, Lowenstein H, Marsh DG, Platts-Mills TA, Thomas W. Allergen nomenclature. Bull World Health Organ 1994;72:797–806.

[11]. King TP, Hoffman D, Lowenstein H, Marsh DG, Platts-Mills TA, Thomas W.

Allergen nomenclature. Clin Exp Allergy 1995;25: 27–37.

[12]. King TP, Hoffman D, Lowenstein H, Marsh DG, Platts-Mills TA, Thomas W. Allergen nomenclature. Allergy 1995;50:765–774.

[13]. King TP, Hoffman D, Lowenstein H, Marsh DG, Platts-Mills TA, Thomas W. Allergen Nomenclature. J Allergy Clin Immunol 1995;96:5–14.

[14]. Radauer, C., Bublin, M., Wagner, S., Mari, A., & Breiteneder, H. (2008). Allergens are distributed into few protein families and possess a restricted number of biochemical functions. *Journal of Allergy and Clinical Immunology*, *121*(4), 847-852.

[15]. FAO/WHO Agriculture and Consumer Protection (2001) Evaluation of allergenicity of genetically modified foods. Report of a Joint FAO/ WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology, Rome

[16]. FAO/WHO Codex Alimentarius Commission (2003) Codex principles and guidelines on foods derived from biotechnology. Joint FAO/ WHO Food Standards Programme, Rome Stadler MB, Stadler BM (2003) FASEB J 17:1141−1143

[17]. Zorzet A, Gustafsson M, Hammerling U (2002) In Silico Biol 2:525−534

[18]. Stadler, M. B., & Stadler, B. M. (2003). Allergenicity prediction by protein sequence. The FASEB Journal, 17(9), 1141-1143.

[19]. umar, K. K., & Shelokar, P. S. (2008). An SVM method using evolutionary information for the identification of allergenic proteins. *Bioinformation*, *2*(6), 253.

[20]. Muh, H. C., Tong, J. C., & Tammi, M. T. (2009). AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins. *PloS one*, *4*(6), e5861.

[21]. Dimitrov, Ivan, et al. "AllerTOP v. 2—a server for in silico prediction of allergens." Journal of molecular modeling 20.6 (2014): 2278.

[22]. Altschul, Stephen F., et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids.

[23]. Altschul, Stephen F., et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids.

[24]. Altaf-Ul-Amin, Md, et al. "DPClus: a density-periphery based graph clustering software mainly focused on detection of protein complexes in interaction networks." *Journal of Computer Aided Chemistry* 7 (2006): 150-156.

[25]. Aalberse, R. C., Akkerdaas, J., & Van Ree, R. (2001). Cross‐reactivity of IgE antibodies to allergens. *Allergy*, *56*(6), 478-490.

[26]. Marchler-Bauer, Aron, et al. "CDD/SPARCLE: functional classification of proteins via subfamily domain architectures." *Nucleic acids research* 45.D1 (2016): D200-D203.'

[27]. Ayuso, Rosalia, et al. "Molecular basis of arthropod cross-reactivity: IgE-binding cross-reactive epitopes of shrimp, house dust mite and cockroach tropomyosins." *International archives of allergy and immunology* 129.1 (2002): 38-48.

[28]. Popescu, Florin-Dan. "Cross-reactivity between aeroallergens and food allergens." *World journal of methodology* 5.2 (2015): 31.