

Chapter 12

Bioinformatic Classifiers for Allergen Sequence Discrimination

Yuyi Huang and Ailin Tao

Abstract With the rapid development of biotechnology and sequencing techniques, the volume of various protein and sequence data has grown exponentially. However, compared to the rapid breakthrough in large-scale sequencing, management and mining of these data has made little progress. Proteins are a collection of macromolecules that play essential roles in organisms. Prediction of the functional classification of a protein is a major task of bioinformatics. Recently, with the rapid developments in bioinformatics, computational classification of proteins has revolutionized protein research by guiding experimental design.

Keywords Allergen · Bioinformatics · Classifiers · Protein · Prediction

12.1 Introduction

12.1.1 *The Primary Structure of Proteins*

Proteins are a collection of macromolecules that play essential roles in organisms. Predicting the functional classification of a protein is a major task of bioinformatics since the primary structure of a protein determines its characteristics

Y. Huang · A. Tao (✉)

Guangdong Provincial Key Laboratory of Allergy and Clinical Immunology,
The State Key Clinical Specialty in Allergy, The State Key Laboratory of Respiratory Disease,
The Second Affiliated Hospital of Guangzhou Medical University,
250# Changgang Road East, Guangzhou 510260, Guangdong Province,
People's Republic of China
e-mail: taoailin@gzhmu.edu.cn

Y. Huang
e-mail: chvipdata@126.com

and function (Wu et al. 2013). Protein structure is dependent on its amino acid sequence and it is widely recognized that this sequence can be used to predict the function of a protein (Popov et al. 2014).

All proteins are made of a set of 20 amino acids, each of which is assigned a unique letter, and can be represented as $AA = [A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, \text{ and } Y]$. Different amino acid alignments make different proteins. When two amino acids link through esterification between the carboxyl group of the first amino acid and the amino group of the next, the connection is called a peptide bond. The linear structure formed by linked amino acids is called a peptide, and when a peptide becomes longer and consists of several amino acids, it is called polypeptide. One end of a polypeptide has a free amino group and is called the amino terminus and the other end has a free carboxyl group and is called the carboxyl terminus. In a sense, proteins are basically a strand or several strands of polypeptides. For example, ferredoxin and myoglobin are composed of a single polypeptide chain, insulin is composed of two polypeptide chains that are connected by disulfide bonds, and hemoglobin is composed of four polypeptides that are connected non-covalently. An amino acid sequence is translated from messenger RNAs, which were transcribed from DNA, and thus ultimately, it is the DNA sequence that determines the characteristics and function of proteins.

12.2 Sequence Data Mining and Classification

With the rapid development of biotechnology and sequencing techniques, the volume of various protein and sequence data has expanded greatly but management and mining of these data has made little progress. Data itself is not knowledge. Some point mutations or even large-scale deletions may exert little effect on the biological function of a certain protein. These amino acid fragments are called noise sequences. Other amino acids or sequences are largely conserved, however, once replaced or deleted, severe consequences may occur that lead to partial or complete loss of protein function and these fragments are called characteristic sequences (Zhang et al. 2012). This difference is the basis of sequence data mining and classification.

The traditional way to understand the function of a protein is through experimentation that can be expensive and time consuming; therefore, it is not possible to test every protein experimentally. Meanwhile, various sequence databases have been filled by an ocean of sequence data waiting to be mined for useful information. This data could have great clinical significance in understanding the potential correlation between the sequence and function of proteins as well as predict protein function from its sequence through machine learning and data mining. Nobel Prize winner W. Gilbert had pointed out that traditional biology seeks answers by experiments, while emerging approaches would be based on full knowledge of the genome electronically stored in databases. The next generation of biology research should be initiated by theories, with the scientist starting from theoretical

assumptions and then testing the assumptions by experimentation (Whelan et al. 2013). In recent years, with the rapid development in bioinformatics, computational classification of proteins has revolutionized protein research by guiding experimental design (Yu et al. 2015; Jones et al. 2014; Nevado and Perez-Enciso 2015; Srinivasan et al. 2013).

12.3 Prediction of Allergen Classification

12.3.1 Introduction

Sequence data is a discrete set of ordered data containing genetic information as an ordered series of a certain length. In fact, conservation and divergence of the sequence data contain very valuable information. They reflect the adaptive changes of the genetic information during the long period of evolution, thus representing the “memory” of a species. If we can quantitatively describe such characteristic “memory,” we would be able to predict the unknown; therefore, analysis of the signature of the sequence data becomes the key to predicting sequence classification. Prediction of allergen protein classification is basically performed in two ways, the feature extraction-based method and the similarity-based method. The feature extraction-based method extracts the characteristic parameters of the amino acid sequence based on conservation and divergence data. The amino acid sequence of a protein can be abstractly described by the characteristic parameters, which can then be used for machine learning and classification. The similarity-based method collects a sufficiently large database of allergen sequences, which is a prerequisite for the sequence similarity model to work. The sequences with high similarities can be retrieved and, through certain approaches, correlations among the sequences can be identified and thereby an unknown function of a sequence can be predicted.

12.3.2 Allergen Classification

Allergen classification can be fulfilled by several methods as follows.

12.3.3 Similarity-Based Method

The criteria proposed by FAO/WHO (FAO/WHO 2001) to distinguish allergens and nonallergens are based on similarity. A protein is identified as an allergen if (i) it contains a 6-aa sequence of an already known allergen, or (ii) it shows at least 35 % identity with an already known allergen in an 80-aa segment. Such definition

requires a pre-established database for sequence alignment. That is, a database of known allergens would be required before similarity testing using BLAST (Altschul et al. 1997), FASTA (Pearson 1988), or other alignment tools. The criteria above belong to the classical model of allergen classification, which in practice may be adjusted.

Construction of a similarity model includes the following three steps: (i) Establish an allergen sequence database, excluding repeated and redundant data to ensure accuracy and applicability; (ii) Submit the target sequence to the database and search for similar sequences using BLAST or FASTA; and (iii) Filter the results according to existing or customized standards for classification.

Accuracy of the similarity model depends on the accuracy of the allergen sequence database, thus, only if the sequences in the database are confirmed to be allergens, proteins having similarities with the database sequences would be allergens. So it is of utmost importance to establish an accurate sequence database, examples of which are described below.

Allermatch

Allermatch (Fiers et al. 2004) is an online allergen prediction tool using FAO/WHO criteria. The prediction steps are as follows: (1) Amino acid sequences of known allergens are obtained in FASTA format; (2) The submitted sequence is read into overlapping 80-aa slices, also in FASTA format; (3) Each of the slices is aligned with sequences of known allergens in the database; and (4) Using the FAO/WHO standard, determine whether the target sequence is a potential allergen or not. As mentioned above, the accuracy of the similarity model depends on accuracy of the allergen sequence database that is used. In this regard, the Allermatch database has collected 730 known allergens from WHO-IUIS or SwissProt.

WebAllergen

WebAllergen (Riaz et al. 2005) is another online allergen prediction service. It provides a database of 664 known allergens. Users can submit their sequences to align with the sequences in the database or other allergen sequence databases of their own. The server converts submitted allergens into allergen motifs, which are aligned with the target sequence to be classified.

12.3.4 Feature Extraction-Based Method

Prediction of allergen classification using the feature extraction-based method consists of the following steps: (1) Construct a sequence database of allergens and non-allergens using data that is accurate, objective, and typical; (2) Extract characteristic parameters according to conservation and divergence of the sequence among related protein family so that the amino acid sequence of the protein can be classified by characteristic parameters; (3) Design a classifier by matching learning; and (4) Predict the classification of an unknown protein using the classifier. Due to the diversity and complexity of the allergen sequences in sequence length

and order, conservation and divergence of each site of the amino acid sequences differs, making it difficult to find a characteristic parameter representative of known allergens and to construct a proper classifier. However, many advances have been made in this field as described below:

EVALLER

EVALLER (Martinez Barrio et al. 2007) is an interactive web program based on the filtered length-adjusted allergen peptides (DFLAP) algorithm. DFLAP is characterized by two features: (1) highly specific allergen peptides can be selected by comparison to nonallergens and (2) an allergen discrimination model is established using the support vector machine (SVM) model. This algorithm shows very high sensitivity and specificity in identification of potential allergens. The EVALLER web server accepts FASTA protein sequences as input for allergenicity assessment and the results are presented as both text and graphics so that the user can conveniently evaluate accuracy of the results.

AllerHunter

AllerHunter (Muh et al. 2009) is a web-based program for the assessment of potential allergenicity and cross-reactivity. This program integrates FAO/WHO and the SVM learning model and has been tested using 1356 known allergens and nonallergens. This system works by, first, evaluating the potential allergenicity and cross-reactivity of a protein sequence using the SVM learning model, with a likelihood score as output, and then, the output is further evaluated by FAO/WHO criteria, resulting in a comprehensive assessment of the allergenicity.

AlgPred

AlgPred (Saha and Raghava 2006) integrates several different methods to identify allergens. Its test set and training set include 578 strands of allergen sequences and 700 strands of nonallergen sequences. These sequences were divided into five clusters by BLAST similarity search (E value $8E-4$, identity 26 %), where no similarity exists between different clusters and each cluster contains a nearly equal number of sequences. A classifier is then designed using a corresponding machine learning approach. By grouping the sequences into clusters, the similarity within the training set can be reduced without reducing the total number of sequences, so that accuracy of classification prediction can be improved.

Allerdicator

Allerdicator (Dang and Lawrence 2014) proposes a sequence-based method for allergenicity prediction. The basic idea of this program is to represent a protein by its k -mer feature. Due to the feature construction nature and adoption of a linear SVM model, linear prediction can be achieved. Allerdicator uses both naïve Bayes (NB) and SVM classifiers. The SVM model performs better than NB when dealing with highly similar sequences between allergens and nonallergens. The length of k -mer peptide is the most important parameter of the Allerdicator model. The performance of Allerdicator varies when using different k values. A k value of 5 or 6 shows the highest performance, with a value of $k = 5$ indicating a near-perfect false positive rate (FP), while a value of $k = 6$ shows higher sensitivity while maintaining a low FP rate.

SORTALLER

SORTALLER (Zhang et al. 2012) is an online allergen classifier based on an allergen family featured peptide (AFFP) dataset and normalized BLAST E values, which established the featured vectors for the support vector machine (SVM) model. AFFPs are allergen-specific peptides panned from irredundant allergens and harbor perfect information with noise fragments eliminated because of their similarity to nonallergens. SORTALLER performed significantly better than other existing software and reached a perfect balance with high specificity (98.4 %) and sensitivity (98.6 %) for discriminating allergenic proteins from several independent datasets of protein sequences of diverse sources, also highlighted by a Matthews correlation coefficient (MCC) as high as 0.970, a fast running speed, and rapid predicting of a batch of amino acid sequences with a single click.

12.4 Conclusions

With the continuous development of genetically modified crops and products, the allergenicity problem is becoming increasingly severe. Bioinformatics studies have made much progress in predicting the allergenicity of highly conserved sequences with breakthroughs in sequence acquisition, processing, storage, distribution, analysis, interpretation, etc. However, precise distinction between potential allergens and nonallergens is still far from perfect. Thus, assessment of allergenicity and cross-reactivity of proteins by the comprehensive application of mathematics, computer science, and biology tools will be the direction of this area of research in the coming years.

References

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389–402.
- Dang HX, Lawrence CB. Allerdicator: fast allergen prediction using text classification techniques. *Bioinformatics.* 2014.
- FAO/WHO. Evaluation of allergenicity of genetically modified foods. Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology. 2001;22–5.
- Fiers MW, Kleter GA, Nijland H, Peijnenburg AA, Nap JP, van Ham RC. Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinf.* 2004;5:133.
- Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30(9):1236–40.
- Martinez Barrio A, Soeria-Atmadja D, Nister A, Gustafsson MG, Hammerling U, Bongcam-Rudloff E. EVALLER: a web server for in silico assessment of potential protein allergenicity. *Nucleic Acids Res.* 2007;35(Web Server issue):694–700.

- Muh HC, Tong JC, Tammi MT. AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins. *PLoS ONE*. 2009;4(6):e5861.
- Nevado B, Perez-Enciso M. Pipeliner: software to evaluate the performance of bioinformatics pipelines for next-generation resequencing. *Molecular ecology resources*. 2015;15(1):99–106.
- Pearson WR. FASTA: Improved tools for biological sequence comparison. *Proc Natl Acad Sci*. 1988;85(8):2444–8.
- Popov I, Nenov A, Petrov P, Vassilev D. Bioinformatics in Proteomics: A Review on Methods and Algorithms. *Biotechnol Biotechnol Equip*. 2014;23(1):1115–20.
- Riaz T, Hor HL, Krishnan A, Tang F, Li KB. Web Allergen: a web server for predicting allergenic proteins. *Bioinformatics*. 2005;21(10):2570–1.
- Saha S, Raghava GP. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res*. 2006;34(Web Server issue):W202–9.
- Srinivasan SM, Vural S, King BR, Guda C. Mining for class-specific motifs in protein sequence classification. *BMC Bioinformatics*. 2013;14:96.
- Whelan FJ, Yap NV, Surette MG, Golding GB, Bowdish DM. A guide to bioinformatics for immunologists. *Front Immunol*. 2013;4:416.
- Wu FX, Resson H, Dunn MJ. Focus on bioinformatics in proteomics. *Proteomics*. 2013;13(2):219–20.
- Yu Guoxian, Rangwala H, Domeniconi C, Zhang Guoji, Zhang Zili. Predicting Protein Function Using Multiple Kernels. *IEEE/ACM Trans Comput Biol Bioinform*. 2015;12(1):219–33.
- Zhang L, Huang Y, Zou Z, He Y, Chen X, Tao A. SORTALLER: predicting allergens using substantially optimized algorithm on allergen family featured peptides. *Bioinformatics*. 2012;28(16):2178–9.

Author's Biography



Yuyi Huang is a research associate working at Guangdong Provincial Key Laboratory of Allergy and Clinical Immunology, the State Key Laboratory of Respiratory Disease, the State Key Clinical Specialty in Allergy of the Second Affiliated Hospital of Guangzhou Medical University. Email: chvipdata@126.com.

He got his master degree of Immunology and Molecular Biology from Guangzhou Medical University in 2010. He is currently working at Guangzhou Medical University. He has experience of alleviation of the allergenic potential of allergens. His main research is focused on allergen bioinformatics. He has developed an allergen database ALLERGENIA and an allergen predicting software SORTALLER.



Ailin Tao is a Professor of Guangzhou Medical University, Director of Guangdong Provincial Key Laboratory of Allergy and Clinical Immunology, Principal Investigator of the State Key Laboratory of Respiratory Disease, Deputy Director of the State Key Clinical Specialty in Allergy of the Second Affiliated Hospital of Guangzhou Medical University, Member of the State Committee for Transgenic Safety Assessment, Standing Committee Member of Allergy Branch of Guangdong Medical Association, Member of Guangdong Provincial Committee for Transgenic Safety Assessment, and Master Tutor.

Prof. Ailin TAO earned his doctorate degree from the State Key Laboratory of Crop Genetic Improvement of Huazhong Agricultural University in 2002, followed by a postdoctoral training at Postdoctoral Station of Basic Medicine in Shantou

University Medical College, majoring in allergen proteins. Recent studies of his lab mainly focus on allergy bioinformatics, allergy and clinical immunology, diseases models such as allergic asthma, allergic rhinitis, infection and inflammation induced by allergy, and inflammatory and protracted diseases caused by antigen or superantigen. He has accumulated some experience in the allergology field including the mechanism on immune tolerance, allergy triggering factors and chronic inflammation pathways, and allergenicity evaluation and modification for food and drugs. He proposed some new concepts such as "Representative Major Allergens," "Allergenicity Attenuation" of immunotoxin and allergens, and "broad-spectrum immunomodulator." He put forward the theoretical hypothesis of "Balanced Stimulation by Whole Antigens." He focuses on the diagnosis of allergic disease, and the allergenicity medical evaluation and modification for food and drugs. He constructed the method system for prediction, quantitative assessment, and modification in parallel on allergenicity epitopes, which has been applied for more than 20 cases of allergens. He has developed a bioinformatics software for allergen discriminating, SORTALLER (<http://sortaller.gzhmu.edu.cn>), which performed significantly better than other existing software and reached a perfect balance with high specificity (98.4 %) and sensitivity (98.6 %) for discriminating allergenic proteins from several independent datasets of protein sequences of diverse sources, also highlighting with the Matthews correlation coefficients as high as 0.970, fast running speed and rapidly predicting a batch of amino acid sequences with a single click. The software has been frequently used by researchers from lots of institutions in more than 20 Chinese cities and over 30 countries worldwide, thus becoming the TOP 1 in this field. He has set up an allergen database ALLERGENIA (<http://allergenlia.gzhmu.edu.cn>), which has obvious international advantages in five aspects: wide allergen coverage, nonredundancy, astringency, accuracy, and friendly and usable analysis functions.