

新規統合型アレルゲンデータベース Allergen Database for Food Safety (ADFS) の構築

中村亮介・手島玲子[#]・高木加代子・澤田純一

Development of Allergen Database for Food Safety (ADFS):
an integrated database to search allergens and predict allergenicity

Ryosuke Nakamura, Reiko Teshima[#], Kayoko Takagi and Jun-ichi Sawada

Allergy has been one of the most common chronic health problems in recent years, and the introduction of recombinant proteins into foods and other products has raised public concern about the induction of allergy. Prediction of food allergenicity is very important but still unsatisfactory. By using an enormous amount of data produced by genomic, functional, and structural studies, bioinformatics can provide useful insights into allergenicity.

We have developed a web server database system which is comprised of allergenic proteins for food safety and homology search tools, "Allergen Database for Food Safety (ADFS)"¹⁾. Since ADFS includes the sequences of known allergens and B-cell epitopes, a potential allergenicity of a novel protein in food should be detected by homology search.

The database contains allergens classified into 8 categories (pollen, mite, animal, fungus, insect, food, latex, and others), together with the public database accession numbers of their genes, and their epitope and 3D-structure information. Users can easily search allergens with keywords and amino acid sequences through the graphical interfaces.

1) URL : <http://allergen.nihs.go.jp/ADFS/>

Key Words: allergen, database, food safety, allergenicity prediction, epitope

(Received May 31, 2005)

1. はじめに

今日の生物学において、世界中の研究者が容易にアクセスできる大規模データベースと洗練された解析ツールは必須なものとなっている。遺伝子やタンパク質の一次配列に関するこの種のデータベースについては、日米欧において満足のいく機能を持つものが複数構築されている。しかし、アレルゲンに関しては、小規模なデータベースは数種あるものの²⁾、その分類学的情報や一次配列等、様々な情報を集め解析する有用なデータベースは、未だ確立されていない。そこで本研究では、新開発食品のアレルゲン性を予測する目的でアレルゲンに関する既存の各種データベースや一次文献の情報を集積し、アレルゲン名・カテゴリ（花粉・ダニ・動物・カビ・昆虫・食物・ラテックス・その他）・キーワード（動物種・一般名等）およびアミノ酸配列等により検索可能な新規統合型アレルゲンデータベース（Allergen Database for

Food Safety; ADFS）を構築した¹⁾。さらに、新規産生タンパク質と既知アレルゲンとの交差反応を考える上で非常に重要となるB細胞エピトープ（抗原決定基）情報および立体構造情報（PDBエントリ）をここに付加した。特に、エピトープ情報に関しては独自に新たな文献検索を行ない、データ整備に努めた。また、このデータベースは、FAO/WHOが提唱するタンパク質のアレルゲン性予測インターフェイスを有しており、アレルゲン性予測ツールとしても利用することができる。

2. 方 法

データは原則としてすべて2004年3月の時点で収集・解析した。なお、エピトープ情報については、2005年3月現在のデータを入力した。アレルゲンの一次配列データは、List of Allergens in Swiss-Prot³⁾、Allergen Nomenclature⁴⁾、The Biotechnology Information for Food Safety Database⁵⁾、SDAP-Structural Database of Allergenic Proteins⁶⁾より収集し、IDが重複するものを除いた。IDの優先順位は、SwissProt、Pir、TrEMBL、GenBank/GenPeptの順とした。登録されている配列デー

[#]To whom correspondence should be addressed: Reiko Teshima; Kamiyoga 1-18-1, Setagaya, Tokyo 158-8501, Japan; Tel: 03-3700-1141 ext.243; Fax: 03-3707-6950; E-mail: rteshima@nihs.go.jp

タがcDNAの場合、TrEMBLにより対応するアミノ酸配列を取得した。

各アレルゲンのアノテーション情報（動物種・一般名・注釈等）はSDAPから収集し、同サイトが提供する「Source」属性を、8種のカテゴリ（花粉・ダニ・動物・カビ・昆虫・食物・ラテックス・その他）に再編成した。また、下記に示す11のキーワードによりEntrez PubMed内を検索し、エピトープ情報を含む文献を抽出し、精読した後SDAPのエピトープ情報に追加した。

IgE・Binding, Epitope, Identification, Immunoglobulin E, Epitope Mapping, Sequence, Analysis, Peptide, Recognition, IgE・epitopes, Linear

システムの構成としては、OSとしてSolaris 9を、サーバーエンジンとしてTomcat 5.0.28を、WebサーバーとしてApache 2.0.53を、データベースエンジンとしてMySQL 4.0を、SRSシステムとしてSRS 7.1.3にUniProtデータベースを導入したものを用いた。アミノ酸配列に基づくアレルゲンタンパク質の検索（Protein Search）にはprotein・protein BLAST（blastp 2.2.10）を、エピトープ配列内の検索（Epitope Search）にはBLAST Search for short, nearly exact matchesを用いた。FAO/WHOに提唱されたタンパク質のアレルゲン性予測法⁷⁾としては、Hilemanらの方法⁸⁾を一部改変したものをを用いた。すなわち、FASTAアラインメントプログラム（Ver. 33t08d4）によりクエリ（ユーザにより問われている）タンパク質の全長を既知アレルゲンと比較し、得られたアラインメント中のオーバーラップアミノ酸長（およびその一致率）と連続一致アミノ酸長をそれぞれ異なる閾値で判定するというものである。

3. 結 果

3.1 アレルゲン検索

重複のないアレルゲンの一次配列データとして730種を選出し、データ精査の後、ウェブブラウザで検索できるデータベース（Allergen Database for Food Safety; ADFS）として公開した¹⁾。Fig.1aはその検索画面である。様々な検索手法を視覚的にレイアウトし、初見ユーザにも戸惑うことなく操作できるよう配慮した。個々のアレルゲンに関する詳細な情報は、別ウインドウに表示される。Fig.1bに、Bos d 6（ウシ血清アルブミン）の例を示した。詳細データ表示画面からは、UniProt、GenBank、PDB、Entrez PubMedへのリンクが適宜貼っており、必要に応じて外部のデータベースを参照することを可能にした。

我々が行った今回の文献調査では、エピトープ情報に関する32報の論文が見つかり、32種のアレルゲン（Ana o 1.0101, Ana o 1.0102, Ana o 2, Asp f 1, Asp f 3, Ber e 1, Bla g 7, Blo t 10, Blo t 11, Bos d 6, Bos d 8,

Cry j 2, Fag e 1, G2a, Gal d 1, Gal d 2, Hel a 2, Hev b 6.02, Jug r 1, Jun a 1, Tri a 19, Par j 1, Par j 2, Pen a 1, Pen ch 13, Pen ch 18, Pen n 18, Phl p 13, Phl p 5.0102, Pru av 1, Pru p 3, およびSes i 2）に対し、計201種のエピトープ情報を追加することができた。これらの情報をSDAPによる既知エピトープ情報に加えることにより、エピトープが一つ以上明らかになった登録アレルゲン（立体エピトープを含む）の数は43種になった。また、立体構造が分かっているアレルゲンは157種存在した。これらエピトープおよび立体構造が分かっているアレルゲンの情報は、ADFSのKeyword Searchより知ることができるようにした。

ユーザは、興味あるアミノ酸配列と相同性のあるアレルゲンをBLASTアルゴリズムにより高速に検索することが可能となり（Protein Search）、E・value（閾値として利用する期待値）やMatrix（アミノ酸置換スコアの行列）などのオプションはユーザが任意に設定できるようにした。また、特にエピトープ配列に限った検索は、Epitope Searchにより行なうこととした。Epitope SearchではE・valueおよびMatrixの初期値を短いアミノ酸配列に最適化してあり、4残基からなるエピトープも検索することができるようにした。この機能を用いると、エピトープ既知のアレルゲンの場合、そのエピトープ配列をADFS内のエピトープ配列または全配列、あるいはUniProt内の全配列に対して検索し、交差反応性の疑いがあるタンパク質を検索することが可能となった。たとえば、大豆のアレルゲンG2aのエピトープである²¹⁹SGFAPEFLKEAFGVN²³³をEpitope Searchにかけると、G2a自身の他に、ピーナッツのアレルゲンAra h 3が検索結果として出力される。これらはいずれもグリシニンというタンパク質で、実際両者の間には交差反応性があることが報告されている⁹⁾。

3.2 アレルゲン性予測

タンパク質のアレルゲン性予測は、現在FAO/WHOの方法（2001年）に準じたものが可能である。2001年FAO/WHO法⁷⁾とは、1)シグナル配列を除いたクエリタンパク質をN末端側から80残基（またはそれ以上）のアミノ酸ウインドウで順に区切ってゆき、FASTAアラインメントプログラムにより既知アレルゲンとの比較を行ない、35%以上のアミノ酸が一致する場合や、あるいは2)クエリタンパク質の6～8残基の連続するアミノ酸が既知アレルゲンと完全一致する場合にアレルゲン性が疑われる、とするものである。しかし、上記方法はウインドウ単位に細分化されたクエリ配列を大量に処理する必要があり、計算速度の遅延を招く。そこでADFSにおいては、FAO/WHO法の改変法を用いることにした⁸⁾。すなわち、まずクエリ配列の全長をFASTAアラインメ

a)

Top
Logout
Allergen Search
Name
Category
Keyword
Sequence Search
Protein
Epitope
Allergenicity Prediction
FAO/WHO
Motif-based
Site Map

Allergen Search: Keyword Search

Result : 39 Page 1 of 3 << prev 1 2 3 next >>

No.	Allergen-Scientific Name	Taxonomy Name	Species Common	Category	Epitope	Structure	Description
11	Alt a 1	Alternaria alternata	-	fungus	2	-	
52	Ara h 1	Arachis hypogaea	peanut	food	21	1	vicilin
73	Ara h 2	Arachis hypogaea	peanut	food	10	-	conglutin
83	Ara h 3	Arachis hypogaea	peanut	food	4	-	glycinin
98	Asp f 1	Aspergillus fumigatus	-	fungus	13	7	mitogillin
111	Asp f 2	Aspergillus fumigatus	-	fungus	9	-	
120	Asp f 3	Aspergillus fumigatus	-	fungus	7	1	peroxisomal protein
136	Asp f 13	Aspergillus fumigatus	-	fungus	5	-	alkaline serine protease
235	Bio t 10	Blomia tropicalis	mite	mite	1	-	tropomyosin
250	Bos d 6	Bos domesticus	domestic cattle	food	2	1	serum albumin
252	Bos d 8	Bos domesticus	domestic cattle	food	18	-	caseins

b)

Search Result : Detail Information

- Allergen-Scientific Name: Bos d 6
- Taxonomy Name: Bos domesticus
- Species Common Name: domestic cattle
- Category: food
- Allergen Description: serum albumin
- Sequence: GenBank: GeneID: 2190337, UniProt: accNo. P02769
- Epitope: LSLINRLC, HFEYAVSVLL
- Structure: PDB ID: 1Z7B
- Reference (PMID: 12054661): Title: Some human B and T cell epitopes of bovine serum albumin, the major beef allergen. Author: Tanabe S, Kobayashi Y, Takahata Y, Morimatsu F, Shibata R, Nishimura T. Journal: Biochem Biophys Res Commun. 2002 May 24;293(5):1348-53.

正式名
由来生物の学名
由来生物の一般名
カテゴリ
注釈
一次配列
エピトープ
立体構造
文献

Fig.1 Web pages of the search window and the detailed information window of ADFS

a) 検索画面。キーワード検索によりエピトープ情報を持つアレルゲンを表示させた例。左側には様々な検索・解析ツールが視覚的にレイアウトされている。b) 詳細データ表示画面。例としてaの「Bos d 6」をクリックするとこのような別ウインドウが開く。アレルゲンの動物種や一般名、カテゴリ、エピトープ配列などが表示されるとともに、一次配列 (GenBank/UniProt)、立体構造 (PDB/HSSP)、文献 (PubMed) へのリンクが示される。

ントにより既知アレルゲンとの相同性比較を行ない、1) 両者においてオーバーラップしているとみなされたアミノ酸長が80残基以上に達し、かつその35%以上のアミノ酸が一致する場合、あるいは2) 連続して完全一致した最大アミノ酸長が6~8残基以上に達した場合に「陽性」と判定される。いずれの手法においても、1)はタンパク質の比較的大きな構造の類似性を調べるもので、2)は局所的な一致をみるものである。ユーザはこれらのパラメータおよびE-valueを任意に変更し、クエリタンパク質のアレルゲン性を予測するとともに、類似する既知

アレルゲンに関する情報を容易に得ることが可能になった。

以上のADFSの機能をまとめると、Fig.2のようになる。

4. 考 察

大規模な公的データベースであるSwissProt, Pir, GenPept/GenBankにはそれぞれ593種, 348種, 668種の互いに重複するアレルゲンが登録されていたが、ADFSはこれらを統合し、730種のエントリとしてまとめた。これは、アミノ酸配列比較に基づくアレルゲン性

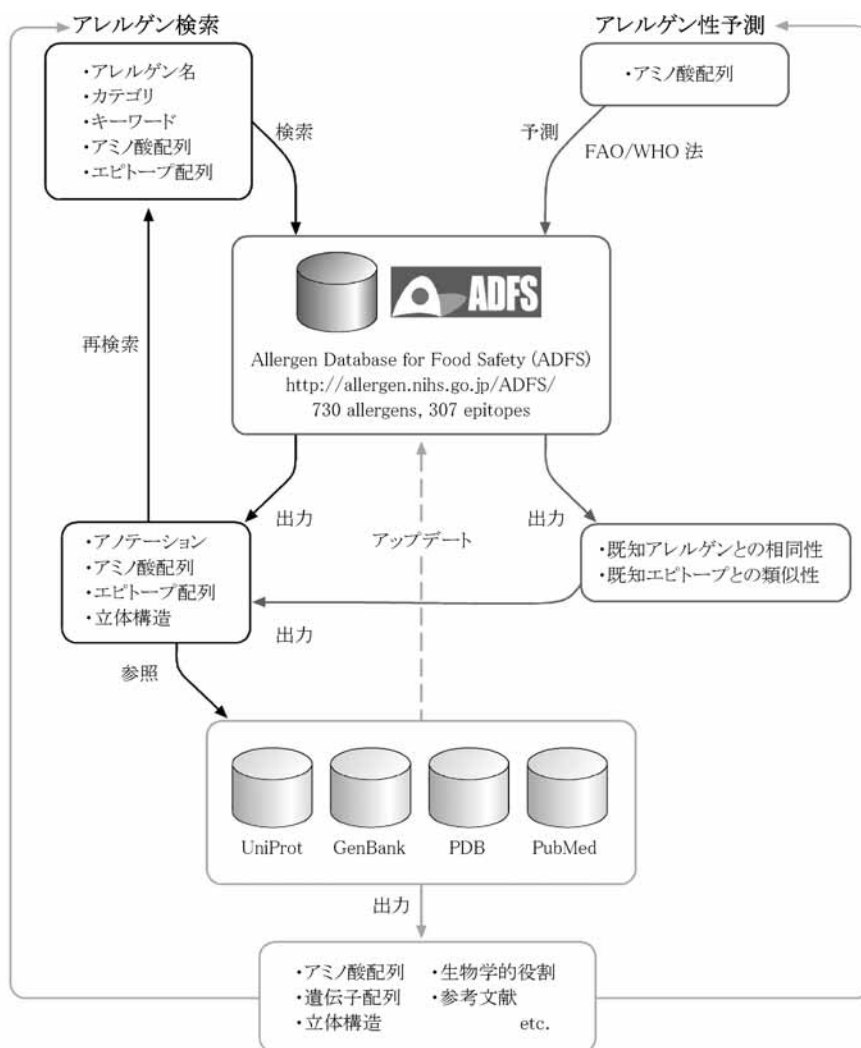


Fig.2 Schematics of ADFS

本データベースは、キーワードやアミノ酸配列など様々なクエリによりアレルゲンを検索する機能と、FAO/WHOの改変法に基づくアレルゲン性予測機能からなる。得られた出力結果は、再度ADFS内を検索したりリンクにより外部データベースを参照したりすることができる。

の予測に関し、非常に有用な情報を提供できるデータベースが構築できたと考えられる。また、既存のアレルゲンデータベースは一般にキーワード検索やアミノ酸配列検索における柔軟性が低く、また検索結果の表示が複雑で分かりにくいなどといった問題があったが、ADFSにおいてはこれらのインターフェースを改善し、より直感的な操作を可能にした。特にBLASTアルゴリズムによる配列検索は有用で、4残基からなる短いエピトープ配列も検索することができた。現在のアルゴリズムでは3残基からなるエピトープについては検索できないが、このような極端に短いエピトープはAsp f 2の¹³⁸HWR¹⁴⁰のみであり、全体としての検索のパフォーマンスは非常に高いと考えられる。また、本データベースの特徴の一つであるエピトープ情報の整備に関しては、今回新たに32種のアレルゲンに対してエピトープ情報を付加し、

ADFSに登録されたエピトープ(立体エピトープを含む)既知のアレルゲンは43種、エピトープの配列数は307種に達した。これは、我々が知るかぎり、アレルゲンエピトープに関するデータベースとしては現時点(2005年6月)で世界最大のものである。エピトープ情報は、あるタンパク質と既知アレルゲンとの交差反応性を予測する上で極めて重要であり、バイオテクノロジー応用食品に含まれる新規タンパク質等のアレルゲン性を予測する上で、ADFSは有用な情報を提供することが可能であると思われる。そのため、今後もEntrez PubMedを通じて定期的にエピトープ情報の収集に努め、データのアップデートを行なっていきたいと考えている。

アレルゲン性予測に関しては、2001年FAO/WHO法を改変し、高いパフォーマンスの予測を可能とした。2001年FAO/WHO法と本法ではアミノ酸のウインドウ

に関する取扱いが異なるため、まれに結果が相違する場合がある。多くのタンパク質は分子内に機能的・構造的にまとまったドメイン構造を持つが、ドメインの大きさは様々であり、一概には決められない。FASTA アライメントにより類似性の高い領域を自動的に抽出する本方法は、ウインドウサイズを初めから80残基等に固定してアライメントを行なうFAO/WHOの方法に比べ、より機能的・構造的に意味がある構造の類似性を調べることが可能であると思われる。

しかし、さらに予測率の高いアレルゲン性予測手法の開発が望まれていることもまた事実である。2003年に発表されたStadlerらによるMotif-based法¹⁰⁾は、従来の予測法に比べパフォーマンスの向上が認められており、ADFSにおいては同法に基づくアレルゲン性予測のインターフェイスを現在開発中である。また、アレルゲンデータのアップデートが非常に重要となるため、年間最低1回以上のアップデートを計画している。定期的なデータのアップデートとMotif-based予測インターフェイスの確立を遂行することにより、ADFSは世界で最も洗練されたアレルゲンデータベースの一つとなることが期待される。

謝 辞

本研究は、厚生労働科学研究費の支援を受けて行なわれたものである。

文 献

- 1) URL : <http://allergen.nihs.go.jp/ADFS/> (June 2005)
- 2) Gendel, S.M.: *J. AOAC Int.*, **87**, 1417-1422 (2004)
- 3) URL : <http://www.expasy.org/cgi-bin/lists?allergen.txt> (June 2005)
- 4) URL : <http://www.allergen.org/> (June 2005)
- 5) URL : <http://www.iit.edu/~sgendel/fa.htm> (June 2005)
- 6) URL : <http://fermi.utmb.edu/SDAP/> (June 2005)
- 7) *Joint FAO/WHO Expert Consultation on Foods Derived from Biotechnology*, (2001)
URL : <http://www.fao.org/es/ESN/food/pdf/allergygm.pdf> (June 2005)
- 8) Hileman, R.E., Silvanovich, A., Goodman, R.E., Rice, E.A., Holleschak, G., Astwood, J.D., Hefle, S.L.: *Int. Arch. Allergy Immunol.*, **128**, 280-291 (2002)
- 9) Xiang, P., Beardslee, T.A., Zeece, M.G., Markwell, J., Sarath, G.: *Arch. Biochem. Biophys.*, **408**, 51-57 (2002)
- 10) Stadler, M.B., Stadler, B.M.: *FASEB J.*, **17**, 1141-1143 (2003)