Drug Discovery Today: Disease Models
Vol. xxx, No. xx 2016

In silico tools for exploring potential human allergy to proteins

Maria Hayes[1,*], Pierre Rougé[2], Annick Barre[2,3], Corinne Herouet-Guicheney[4], Erwin L. Roggen[5]

[1]Teagasc, The Irish Agricultural and Food Development Authority, Food BioSciences Department, Ashtown, Dublin 15, Dublin, Ireland
[2]Université de Toulouse, UPS, IRD, UMR 152 PharmaDev, Université Toulouse 3, Faculté des Sciences Pharmaceutiques, 31062 Toulouse cedex 09, France
[3]Paul Sabatier University – Toulouse II, Toulouse, France
[4]Bayer SAS, Human and Animal Safety Assessment – Seeds, 355 rue Dostoievski, 06903 Sophia Antipolis, France
[5]3Rs Management and Consulting ApS, Denmark

Bioinformatics can help scientists to develop hypotheses about proteins that may need to be tested further for risks of causing allergy. *In silico* methodologies and tools like databases and comparison software, play an important role in the assessment of protein allergenicity and allergenicity mechanisms. They can identify whether a novel protein is an existing allergen and/or has the potential to cross-react with an existing allergen. They cannot identify whether a novel protein will 'become' an allergen. AllergenOnline is the tool currently used for the safety assessment of novel proteins, but other tools are also available including the Structural Database of Allergenic Proteins (SDAP) and AllerTOP. Information concerning PeptideRanker, as well as the Hydrophobic Cluster Analysis (HCA) method used for identifying IgE-binding epitopes in food allergens is discussed.

Section editor:
Michelle Epstein, MD, FRCPC, Medical University of Vienna, Department of Dermatology, DIAID, Experimental Allergy, Waehringer Guertel 18-20, Room 4P9.02, A1090, Vienna, Austria.

Introduction

Allergenicity is the potential of any material to cause sensitization and allergic reaction and is frequently associated with the IgE antibody [1]. An existing allergy/allergen is a real and immediate risk [2,3]. Allergens represent a small fraction of the proteins that humans are routinely exposed to. The reason why these proteins can cause T- and B-cell responses remains largely unanswered. Furthermore, a sensitized individual may respond to proteins that share certain structural features with the protein that elicited the initial immune reaction – a phenomenon known as cross-reactivity.

In silico methodologies can identify whether a novel protein is an existing allergen or whether the novel protein has potential to cross-react with an existing allergen. However, they cannot identify whether a novel protein will 'become' an allergen [2]. Data produced from the use of *in silico* methodologies may be used to make a decision about whether additional *in vitro* and *in vivo* testing is required, by serum screening, as recommended by Codex Alimentarius Commission (2009)

*Corresponding author: M. Hayes (maria.hayes@teagasc.ie)

 http://dx.doi.org/10.1016/j.ddmod.2016.06.001

Please cite this article in press as: Hayes M, et al. *In silico* tools for exploring potential human allergy to proteins, Drug Discov Today: Dis Model (2016), http://dx.doi.org/10.1016/j.ddmod.2016.06.001

and Goodman, 2008 [2,4]. In practice, several *in silico* methodologies for determination of protein allergenicity compare amino acid sequences from a novel, trait protein to known food, contact, and respiratory allergenic proteins found in allergen databases [3].

State of the art – methods and tools currently used for allergenicity assessment

According to the most recent guidelines on the allergenicity evaluation of proteins, a novel protein should have a minimum of 35% sequence identity over a window of 80 amino acids when compared with known allergens to be considered a potential allergen [4,5]. This is a very conservative approach when we take into account the high degree of sequence identity that is needed for actual cross-reactivity which is often in excess of 50–60% sequence similarity, over significant spans of the target protein [6].

AllergenOnline (www.allergenonline.org) focuses on sequence identity matches. It provides a detailed description of accepted bioinformatics comparisons on the website. Previously, Siruguri *et al*. and Moran *et al*. have used AllergenOnline for regulatory comparisons [7,11,12]. AllergenOnline provides access to a peer reviewed allergen list and a sequence searchable database (FASTA) [7]. It is used for the identification of proteins that may present a potential risk of allergenic cross-reactivity. AllergenOnline is used currently by industry for the risk assessment of genetically modified food including proteins. The robust allergen database is updated annually by a panel of independent scientists and clinicians.

Real health risks come from inclusion of proteins in a new food, that are allergens from another source or highly likely to be cross-reactive. A much lower risk is presented by the likelihood that a protein will become an allergen *de novo*, or sensitize *de novo* and lead to allergic sensitization [15]. This may be indicated by stability in pepsin, abundance and thermal stability, but these factors could be important in elicitation not sensitization. (Where does sensitization occur? Gut, skin, mouth, airway). In using sequence comparisons, if the protein is found to have been described previously as an allergen (100% or nearly 100% identity), that is a significant risk (weight). If a protein has high sequence identity (50–70+%, it suggests the risk of probable cross-reactivity and would require serum IgE tests with properly targeted allergic human sera. If >35% identity over 80 or more amino acids between a novel and existing protein is found, that is considered a potential allergy risk by Codex [4] and should be evaluated further by serum IgE testing if a proper set of serum donors can be identified (which can be challenging for rarely reported allergenic sources).

In the past, several researchers also used step-wise contiguous identical amino acid segment searches (i.e. 6- and later 8-mer searches), as described in the FAO/WHO guidelines [5,8] to predict human allergenicity to proteins, based on the idea that these segments represented both a theoretical B-cell epitope as well as a minimum size for a conserved T-cell epitope. For instance, Stadler and Stadler [14] reported that a 6-mer match resulted in more than two-thirds of all proteins in SwissProt being predicted to be allergens, and >40% of the human genome being predicted as such. This was confirmed in other studies and as such, this approach was not seen as a reliable criterion for predicting allergenic potential [10–12]. In the past, immunologists have tried to correlate 'known' and 'predicted' B cell and T cell epitopes with allergens, compared to non-allergens or weak-allergens, and failed to be able to develop solid predictions or clusters for allergy. Unfortunately, the ideas outlined by Ladics [4], have not come to fruition.

Overall, this comparison methodology of 35% identity over at least 80 amino acids is considered to be useful for the prediction of potential cross-reactivity with known allergens, but also produces a number of false positive results. The predictive value of sequence similarity searches for allergenicity potential should be carefully deliberated using a weight of evidence approach as no single method can be fully predictive [18]. Moreover, a relatively high degree of identity at the amino acid sequence level, as commonly seen between IgE cross-reactive proteins, cannot guarantee that the protein is a cross-reactive allergen [9,13]. In other words, no perfect correlation exists between these *in silico* results and food allergenicity.

Protein families containing known allergens

The databases used in assessment of potential protein allergenicity or cross reactivity should be composed of protein sequences based on key criteria like the recognition of allergens by IgE (food allergenicity marker), which involves binding to linear or conformational epitopes on allergen surfaces, and should be proven by clinical data in humans. These protein sequence databases should be updated regularly as new allergens are discovered every year.

Ideally, the molecular basis of protein allergenicity should also be studied through analysis of its sequence, structure and B- or T-cell epitopes where they relate to allergenicity [4] but these data are often missing for most of the known allergen databases. Furthermore, B and T-cell epitope search tools may not be able to distinguish between immunogenicity and allergenicity.

Important allergenic protein families include the non-specific lipid transfer proteins (nsLTPs), the 2S albumins, and the cupin superfamily containing the 11S and 7S globulins [19]. The nsLTP proteins account for severe allergic reactions and are found in fruits from the Rosaceae family (peaches and apples), pollen, tree nuts, vegetables and peanuts [20]. Pepsin stability of proteins may be due to secondary and tertiary structural features. For instance, the presence of disulfide bridges is known to stabilize the protein structure. This is

Please cite this article in press as: Hayes M, et al. *In silico* tools for exploring potential human allergy to proteins, Drug Discov Today: Dis Model (2016), http://dx.doi.org/10.1016/j.ddmod.2016.06.001

the case for the 2S albumin family, which has a 3D structure containing four disulphide bridges. Furthermore, the abundance should be taken into account as most of the plant allergens are seed storage proteins, like the 7/8S and 11S globulins that are major components in seeds from dicotyledonous species [20]; However no clear criteria exists to define how much is too much. There is little sequence similarity between cupins and globulins although they share a similar fold, thus assessment of cross-reactivity is sometimes limited [20]. Caseins, parvalbumins and tropomyosins are found in dairy products, fish and crustaceans, molluscs and meat respectively [20]. There are many proteins in these families that have never been associated with allergy and this could be due to: (1) the broadness of the family designations, (2) there has been little or no exposure to these proteins and (3) the overall structure and sequence similarities are not sufficiently definitive in a biological sense [17].

New methodologies, new perspectives

Figure 1 illustrates these and links them to potential human allergenicity and immunogenicity to protein prediction steps. Assigning proteins as allergens may involve assessment of their amino acid and dipeptide composition using support vector machines (SVMs) [21,22]. Other methods could include motif-based techniques using the software MEME/MAST and comparison algorithms with 'Allergen Representative Proteins' (ARPs) [23]. *In silico* methods for identification of B-cell epitopes could include hydrophilicity scans, amino acid property assessment and combinations of both methods. Computational prediction methods for prediction of peptide binding to human leucocyte antigen (HLA), which is a prerequisite for T-cell recognition, are based on binding motifs, quantitative matrices or artificial intelligence methods and can reduce the number of experiments required to identify relevant T-cell epitopes [24,25]. However, to date, there has not been any demonstration that these new models outperform a FASTA sequence comparison with a well-developed allergen database using criteria of >35–40% identity over 80 amino acids. For the most part, the value of predictions made using these databases depends upon the dataset.

Hydrophobic Cluster Analysis (HCA) as a relevant tool for predicting the IgE-binding epitope regions in food allergens

The amino acid residues forming the IgE-binding epitopes exposed on the surface of allergenic proteins usually share a set of physico-chemical characteristics that can be used for predicting the potential immunogenicity and allergenicity of food proteins. These characteristics mainly consist of (1) the hydrophilicity, due to the occurrence of polar residues (Asn/N, Gln/Q, His/H, Ser/S, Thr/T, Tyr/Y residues), (2) the electronegative (Asp/D and Glu/E residues) and/or electropositive characteristics (Arg/R and Lys/K residues) of residues and (3) the flexibility of residues (Gly/G, Ser/S, Thr/T residues) [26]. Owing to the combination of the physico-chemical characteristics of their building residues, most of these epitopes coincide with loops, which often protrude from the surface of the allergenic proteins. However, other secondary structural features like strands of β-sheet or α-helix, can be readily exposed on the surface and thus, participate in the IgE-binding of food allergens.

Recently, researchers used hydropathic profiles based on different scales of hydrophilicity/hydrophobicity, flexibility and solvent exposure to predict the linear IgE-binding epitopes of allergenic proteins, either coupled with an epitope-mapping approach or structural analysis. However, hydropathic profiles suffer from inherent limitations with respect to structural information which render them unsuitable for the structural characterization of the predicted epitopes on the surface of the allergens. In this respect, HCA offers an efficient tool [26], allowing association of the predicted epitopes to structural features. The prediction of IgE-binding epitopes with HCA was successfully applied to Pru p 3 and Mal d 3, the nsLTPs from peach and apple fruits [26].

HCA was also recently applied to Sal s 1, the salmon (*Salmo salar*) parvalbumin allergen, Jug r1, the English walnut (*Juglans regia*) 2S albumin allergen, Pru p 3, the peach (*Prunus persica*) lipid transfer protein and Pis v 1, the pistachio (*Pistacia vera*) 2S albumin allergen. YASARA [27] was used to build the three-dimensional models of the proteins. The three-dimensional structure of Pru p 3 (PDB code 2ALG) was used. The IgE-binding epitopes identified on Sal s 1, Jug r 1, Pru p 3, and Pis s 1 were mapped on the molecular surface of the corresponding allergens. Molecular surface cartoons were drawn with Chimera. The HCA profiles of Sal s 1, Jug r 1, Pru p 3, and Pis v 1, were drawn from the drawhca server (http://bioserv.rpbs.univ-paris-diderot.fr/services/HCA/). Segments of the HCA profiles were predicted as putative continuous IgE-binding epitopes when they fulfilled at least three out of the four following criteria: (1) exposure to the solvent, (2) flexibility (Gly, Ser, Thr, His residues), (3) prevalence of hydrophilic residues (Asn, Gln, His, Ser, Thr, Tyr), and (4) occurrence of electropositive (Arg, Lys) and/or electronegative (Asp, Glu) residues. As shown in Fig. 2 most of the linear IgE-binding epitopes identified on Sal s 1, Jug r 1, Pru p 3 and Pis v 1, were correctly predicted on the HCA profiles of the corresponding allergens. Both the predicted and identified epitopic stretches overlapped significantly. However, some discrepancies were found, which related to (1) the extent of the IgE-binding epitopic stretch, which is often under-estimated, and (2) the prediction of extra-epitopes, which have no counterparts among the IgE-binding epitopes immunochemically identified on the molecular surface of the allergens. This is the case for the HCA profiles of Sal s 1 and Pis v 1 allergens, which exhibit an additional epitopic stretch at the C-terminal end of the sequence. In spite of these discrepancies, the critical analysis of the HCA profiles provides a

Please cite this article in press as: Hayes M, et al. *In silico* tools for exploring potential human allergy to proteins, Drug Discov Today: Dis Model (2016), http://dx.doi.org/10.1016/j.ddmod.2016.06.001

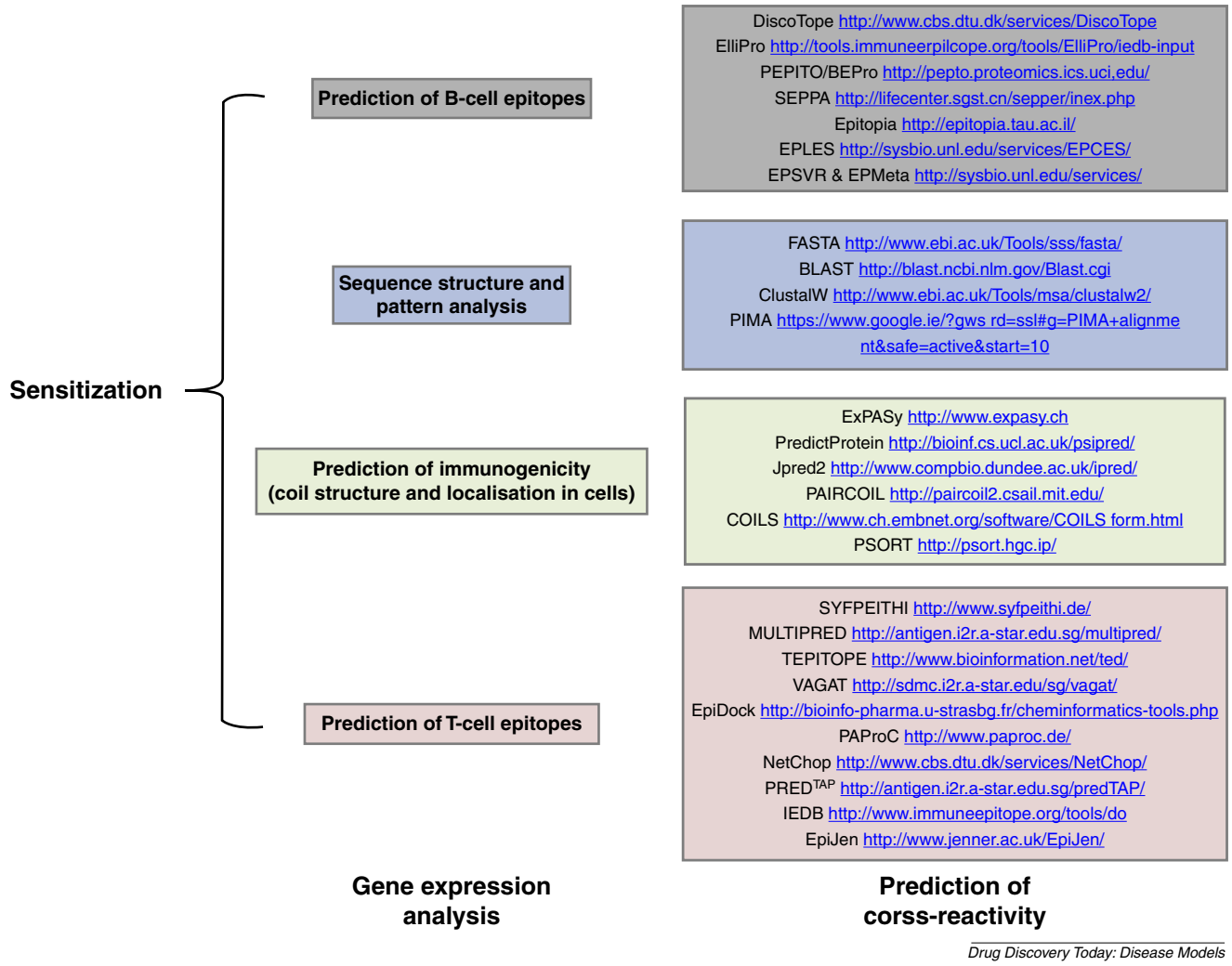

Figure 1. *In silico* tools that may help predict cross-reactivity potential.

rather accurate tool for the prediction of the IgE-binding epitopes of the food allergenic proteins, since the regions in which they occur have been rather correctly predicted.

Three-dimensional (3-D) structure of allergens

The allergenicity potential of proteins may also be identified by using 3-D structure when conformational epitopes are engaged in the allergenic reaction. Linear epitopes can be identified with FASTA and BLAST. Sequence identity using FASTA/BLAST is useful for predicting potential cross-reactivity (depending on the cut-off) as a 3-D structural prediction. In fact, most structural predictions for proteins that have not been tested by crystallography have had to have high FASTA or BLAST alignments to ensure predictions that were accurate. For risk assessment, the suggested program and link is interesting http://scanmail.trustwave.com/?c=6600&d=mcja11FmSSkm7qfkpOzDr5P9z6uTfrk8vKFVfMEU2w&s=61&u=http%3a%2f%2fwww-bionet%2esscc%2eru%2fpsd%2fcgi-bin%2fprograms%2fAllergen%2fallergen%2ecgi. However, a general structural feature of allergens that causes allergenicity has not been described up to now. Allergenicity prediction methods require information about the 3-D structure of query protein; thereby considerably restricting analysis to only those proteins whose 3-D structure is known. As a consequence, many proteins with unknown structure could be overlooked. A new method for allergenicity prediction was developed using information on protein 3-D structure [28]. Three-dimensional structures of known allergenic proteins were used for representing protein surface as patches designated as discontinuous peptides. Allergenicity was predicted by searching for these peptides in query protein sequences. It was demonstrated that the information on the discontinuous peptides may help to predict more accurately potential human allergenicity to protein. The method is available at http://www-bionet.sscc.ru/psd/cgi-bin/programs/Allergen/allergen.cgi [28].

Many freely accessible websites offer comparison tools associated with allergen databases (Table 1).

Please cite this article in press as: Hayes M, et al. *In silico* tools for exploring potential human allergy to proteins, Drug Discov Today: Dis Model (2016), http://dx.doi.org/10.1016/j.ddmod.2016.06.001

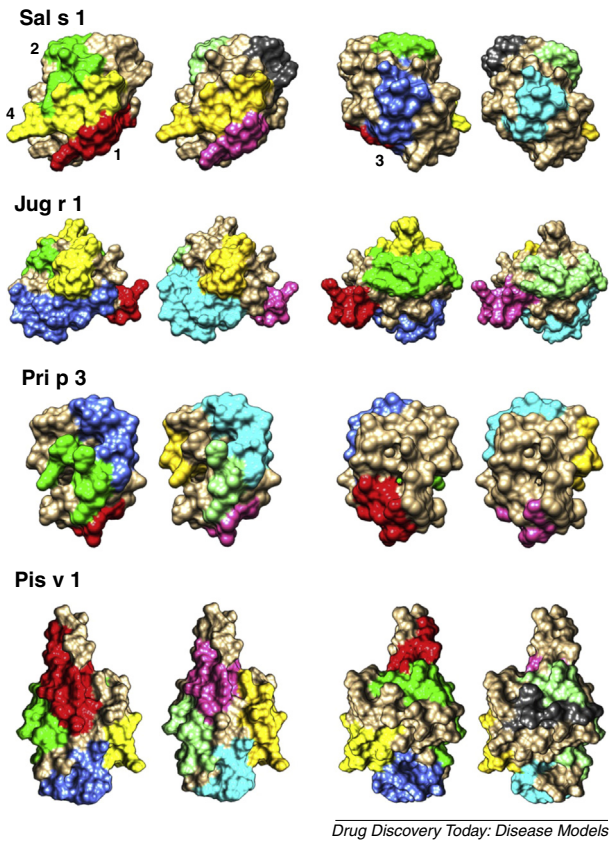

Figure 2. Identified linear IgE-binding epitopes identified on Sal s 1, Jug r 1, Pru p 3 and Pis v 1 predicted using HCA profiles of the corresponding allergens.

PREAL: prediction of allergenic protein by maximum relevance minimum redundancy feature selection

PREAL (http://gmobl.sjtu.edu.cn/PREAL/index.php) predicts potential human allergenicity to protein by integrating various protein properties, including the physicochemical and subcellular locations, using the Maximum Relevance Minimum Redundancy (mRMR) and Incremental Feature Selection (IFS) procedures [29]. The mRMR method was developed to rank each feature according to its relevance to the target and redundancy with other features [30]. IFS procedures were adopted to perform feature selection for analysing the key properties of allergenicity.

Similarities are studied by using NCBI-BLAST software. SSpro/ACCpro 4.03 [31] is used to predict secondary structures of proteins. Solubility is predicted by using the Protein Structure and Structural Feature Prediction Server (SCRATCH; http://download.igb.uci.edu/). The physicochemical properties based on (1) amino acid composition (2) molecular weight (3) hydrophobicity (4) polarizability (5) normalized van der Waals volume and (6) polarity are determined for each protein. The molecular weight of each protein also is also considered. The subcellular location description for proteins also is also incorporated into the SVM.

The PREAL method uses 1176 distinct allergenic proteins from the Swiss-Prot Allergen Index, IUIS Allergen Nomenclature, SDAP and the Allergen Database for Food Safety (ADFS) for building the positive allergen dataset. For building the negative dataset, previously reported methods by Bjorklund [32], Stadler [14] and Barrio and colleagues [33] are integrated and sequence entries removed where identify similarities are greater than 30% to known allergens [29]. In addition, sequences less than 50 amino acids are removed. Using this methodology, the subcellular locations (particularly extracellular/cell surface and vacuole) and amino acid composition were identified as the major markers for allergenicity for specific wheat and soybean proteins previously [30].

AlgPred: prediction of allergenic potential of proteins and IgE epitope mapping

AlgPred (http://www.imtech.res.in/raghava/algpred/) uses an allergen representative peptide (ARP) strategy to try to predict allergenic properties of allergens [23]. Allergens are predicted by (1) MEME/MAST motif searches; (2) SVM-based classification of allergens and non-allergens by single amino acid composition and by dipeptide composition; and (3) BLAST searches against allergen representative peptides. However, to date, PREAL and Algpred have not been demonstrated to outperform FAST or BLAST, depending on the criteria and dataset used.

AllerTOP1.0: prediction of allergenic potential of proteins

AllerTOP (http://www.pharmfac.net/allertop) attempts to predict allergenic potential of proteins by applying auto cross-covariance (ACC) pre-processing to build a dataset of known allergens, developing alignment-independent models for allergen recognition based on the main physico-chemical properties of proteins [34]. It uses five machine learning methods for classification of proteins including discriminant analysis by partial least square (DA-PLS), logistic regression (LR), decision tree (DT), naïve Bayes (NB) and *k* nearest neighbors (*k*NN). AllerTOP also try to identify the most probable route of exposure. In comparison to other models for allergen prediction, AllerTOP out-performs them with 94% sensitivity [35].

Allergen databases

On top of AllergenOnline, several databases exist for example BIOPEP. Although not fully curated and regularly updated, these databases can provide some insight on allergenicity potential of allergens. They include the Allergome (http://www.allergome.org/script/about.php), which has been designed to supply information on IgE-mediated allergens and associated clinical data. However, the use of Allergome is limited as it does not have a searchable function.

BIOPEP (http://www.uwm.edu.pl/biochemia/index.php/en/biopep) is a database of biologically active peptide

Please cite this article in press as: Hayes M, et al. *In silico* tools for exploring potential human allergy to proteins, Drug Discov Today: Dis Model (2016), http://dx.doi.org/10.1016/j.ddmod.2016.06.001

Table 1. *In silico* prediction tools for prediction of potential allergenicity of proteins or for supporting explanatory work.

Web tool	Web tool access address	Advantages of method	Reference
AllergenOnline	http://www.allergenonline.org/	• Methodology currently used for the allergenicity assessment of novel proteins • Peer reviewed allergen list (by independent scientists and clinicians) and sequence searchable tool (FASTA, exact match searches, yearly, curated and updated. • Intended for the identification of proteins that may present a potential risk of allergenic cross-reactivity • Also celiac disease protein database risk assessment tool • Hosted in the University of Nebraska, USA	[40]
AllerHunter	http://tiger.dbs.nus.edu.sg/AllerHunter	• Cross reactive allergen prediction program that uses a combination of SVM and pairwise sequence similarity • Hosted in the University of Singapore, Singapore	[30]
PREAL	http://gmobl.sjtu.edu.cn/PREAL/index.php	• Built on a combination of Support Vector Machine and protein features • Uses AllFam, UIS and Allergome allergen databases and ProAP webtool • Integrates protein biochemical and physicochemical properties (molecular weight, secondary structure propensity, hydrophobicity, polarizability, solvent accessibility, normalized van der Waals volume, polarity, and length) • Integrates sequential features and subcellular locations • mRMR and IFS used to identify allergenicity features • Hosted in the Shanghai Jiao Tong University, China	[6]
AllerTOP 1.0	http://www.pharmfac.net/allertop/	• Based on physicochemical protein properties • Uses a protein sequence mining method (autocross covariance transformation of protein sequences into uniform equal-length vectors). The proteins are classified by *k*-nearest neighbor algorithm (*k*NN, $k = 3$) based on training set containing 2210 known allergens from different species and 2210 non-allergens from the same species. • Hosted in the Sofia University, BulgariaBulgaria	[41,42]
SDAP	http://fermi.utmb.edu/SDAP/	• Investigation of the cross-reactivity between known allergens and in predicting the IgE-binding potential of food proteins • 3-D searches • Possibility to retrieve information related to an allergen from the most common protein sequence and structure databases (SwissProt, PIR, NCBI, PDB), to find sequence and structural neighbors for an allergen, and to search for the presence of an epitope other the whole collection of allergens • Various computational tools that can assist structural biology studies related to allergens • Hosted in the University of Texas, USA	
AlgPred	http://www.imtech.res.in/raghava/algpred/	• Allows prediction of allergens (and its position) based on similarity with known IgE epitopes • Uses several tools (SVM, MEM/MAST, BLASTBLAST, 2890 allergen-representative peptides) and combined approaches • Hosted in the Bioinformatics centre at CSIR-Institute of microbial technology, India	[23]
BIOPEP	http://www.uwm.edu.pl/biochemia	• Contains data on allergenic proteins including names, sequence, sequences of experimental/predicted epitopes • Includes AllFam allergen family and epitopes • Hosted in the University of Warmia and Mazury, Poland	[39]
Pôle Bioinformatique Lyonnais (PBIL)	http://pbil.univ-lyon1.fr/	• Presents information concerning peptide sequence bioactivities on predicted and known allergenic proteins	[43]

Please cite this article in press as: Hayes M, et al. *In silico* tools for exploring potential human allergy to proteins, Drug Discov Today: Dis Model (2016), http://dx.doi.org/10.1016/j.ddmod.2016.06.001

Table 1 (*Continued*)

Web tool	Web tool access address	Advantages of method	Reference
FLAPs	ulfh@slv.se	• Structure prediction of proteins • Filter-length adjusted allergen peptides (FLAPS) database	[44]
BcePred	http://www.imtech.res.in/raghava/bcepred/	• Evaluates the performance of existing linear B-cell epitope prediction methods. 1029 B-cell epitopes • Based on physico-chemical properties (hydrophilicity, flexibility/mobility, accessibility, polarity, exposed surface and/or turns) on a non-redundant dataset from Swiss-Prot • Hosted in the Bioinformatics centre at CSIR-Institute of microbial technology, India	[31]
BepiPred 1.0	http://www.cbs.dtu.dk/services/BepiPred/	• Predicts the location of linear B-cell epitopes using both a hidden Markov model and a propensity scale method • Hosted in the Technical University of Denmark, Denmark	[45]
ABCpred	http://omictools.com/abcpred-s6519.html	• Predicts B cell epitopes in an antigen sequence, using artificial neural network. • Ils able to predict epitopes with 65.93% accuracy using recurrent neural network • Hosted in the Bioinformatics centre at CSIR-Institute of microbial technology, India	[23]
Bpredictor	https://code.google.com/p/my-project-bpredictor/	• Prediction of conformational B-cell epitopes from 3-D structures by random forests with a distance-based feature. • Limited update: last update in 2011	[46]
Epitopia	http://epitopia.tau.ac.il/	• Detection of immunogenic regions in protein structures or sequences (PDB and FASTA) • Machine learning scheme (i.e. Naive Bayes classifier) to rank individual amino acids in the protein, according to their potential of eliciting a humoral immune response • Identify B-cell epitopes (physico-chemical and structural-geometrical properties) • Hosted in Tel Aviv University, Israel	[47]

sequences associated with a program enabling the construction of profiles of the potential biological activity of protein fragments, calculation of quantitative descriptors as measures of the value of proteins as potential precursors of bioactive peptides, and prediction of bonds susceptible to hydrolysis by endopeptidases in a protein chain as well as allergenicity potential. It contains a small number of proteins (i.e. 135) but also allergenic epitopes [36]. Most of the epitopes used are registered in the Immune Epitope Database (IEDB) [37]. Secondary peptide structures are predicted using GOR V program [38]. BIOPEP is a database of peptides that contains recently identified allergenic peptides. Recently, sixty sequences of epitopes from the BIOPEP database attributed to tropomyosin from the shrimp *Farfantepenaeus aztecus* (Pen a 1.0102) were used as query sequences [39]. Vertebrate tropomyosins (e.g. from vertebrates used as food resources) contain fragments containing between 10 and 15 amino acid residues revealing 100% identity with epitopes from allergen Pen a 1.0102. Fragments identical to epitopes from Pen a 1.0102 are common in sequences of invertebrate tropomyosins, including those annotated in the Allergome database. Common epitopes are a probable molecular basis for cross-reactivity between food and non-food invertebrates. Some epitopes, especially rare penta-peptides containing the DEERM sequence, are present in sequences of proteins not sharing homology with tropomyosins. This fragment was found to be present in several proteins, from edible plants and animals as well as pathogenic microorganisms.

Conclusion

This paper reviews current *in silico* tools for assessing potential human allergenicity to proteins. These methods use a number of physico-chemical features (mainly amino acid searches) of proteins that can be predicted, but a strict, structural correlation between these features and allergenicity does not exist. Use of future innovative *in silico* methods for the prediction of allergenicity will be largely influenced by the choice of databases and algorithms that will be developed, standardized and most importantly empirically validated. Prediction of potential allergy is not proof of allergy. Further biochemical testing (IgE blotting) and biological tests including Basophil, skin prick tests, or *in vivo* challenge tests

Please cite this article in press as: Hayes M, et al. *In silico* tools for exploring potential human allergy to proteins, Drug Discov Today: Dis Model (2016), http://dx.doi.org/10.1016/j.ddmod.2016.06.001

with allergenic subjects are needed to validate allergy to protein predictions.

Conflict of interest

The authors declare that there are no conflicts of interest.

Acknowledgements

This work was supported by the EU COST Action ImpARAS FA1402. The opinions expressed herein and the conclusions of this publication are those of the authors.

References

[1] Verhoeckx KCM, Vissers YM, Baumert JL, Faludi R, Feys M, Flanagan S, et al. Food processing and allergenicity. Food Chem Toxicol 2015;80:223–40.
[2] Goodman RE. J Huazhong Agric Univ 2014;33(6):85–113.
[3] Goodman RE, Vieths S, Sampson HA, Hill D, Ebisawa M, Taylor SL, et al. Allergenicity assessment of genetically modified crops—what makes sense? Nat Biotechnol 2008;26(January (1)):73–81. http://dx.doi.org/10.1038/nbt1343. Review. Erratum in: Nat Biotechnol. 2008 February;26(2):241. PMID: 18183024.
[4] Codex Alimentarius Commission guidelines. Foods derived from modern biotechnology (2003 reprinted in the second edition in 2009). World Health Organization and Agricultural Organization of the United Nations; 2009.
[5] Mirsky HP, Cressman RJ, Ladics GS. Comparative assessment of multiple criteria for the in silico prediction of cross-reactivity of proteins to known allergens. Regul Toxicol Pharmacol 2013;67(2):232–9.
[6] Wang J, Yu Y, Zhao Y, Zhang D, Li J. Evaluation and integration of existing methods for computational prediction of allergens. BMC Bioinform 2013;14:S1. http://dx.doi.org/10.1186/1476-2105-14-S4-s1.
[7] Ladics GS. Current Codex guidelines for assessment of potential protein allergenicity. Food Chem Toxicol 2008;46:S20–3.
[8] FAO/WHO. Evaluation of allergenicity of genetically modified foods. Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods derived from biotechnology, 22nd–25th January, 2001. FAO/WHO; 2001. 15.
[9] Aalberese RC, Crameri R. IgE-binding epitopes: a reappraisal. Allergy 2011;66:1261–74.
[10] Goodman RE, Ebisawa M, Ferreira F, Sampson HA, van Ree R, Vieths S, et al. AllergenOnline: a peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. Mol Nutr Food Res 2016. http://dx.doi.org/10.1002/mnfr.201500769.
[11] Siruguri V, Bharatraj DK, Vankudavath RN, Mendu VV, Gupta V, Goodman RE. Evaluation of Bar, Barnase and Barstar recombinant proteins expressed in genetically engineered *Brassica juncea* (Indian mustard) for potential risks of food allergy using bioinformatics and literature searches. Food Chem Toxicol 2015;83:93–102.
[12] Moran DL, Tetteh AO, Goodman RE, Underwood MY. Safety assessment of the calcium-binding protein, apoaequorin, expressed by *Escherichia coli*. Regul Toxicol Pharmacol 2014;69(2):243–9.
[13] Puumalainen TJ, Poikonen S, Kotoyuori A, Vaali K, Kalkkinen N. Napins, 2 S albumins, are major allergens in oilseed rape and turnip rape. J Allergy Clin Immunol 2006;117:426–32.
[14] Stadler MB, Stadler BM. Allergenicity prediction by protein sequence. FASEB J 2003;17(9):1141–3.
[15] Ladics GS, Fry J, Goodman R, Herouet-Guicheney C, Hoffmann-Sommergruber K, Madsen CB, et al. Allergic sensitization: screening methods. Clin Transl Allergy 2014;4:13.
[17] Mittag D, Batori V, Neudecker P, Wiche R, Friis EP, Ballmer-Weber BK, et al. A novel approach for investigation of specific and cross reactive IgE epitopes on Bet v 1 and homologous food allergens in individual patients. Mol Immunol 2006;43:268–78.
[18] Bragin AO, Demenkov PS, Kolshonov NA, Ivanisenko VA. Accuracy of protein allergenicity prediction can be improved by taking into account data on allergenic protein discontinuous peptides. J Biomol Struct Dyn 2013;1:59–64.
[19] Radauer C, Bublin M, Wagner S, Mari A, Breiteneder H. Allergens are distributed into a few protein families and possess a restricted number of biochemical functions. J Allergy Clin Immunol 2008;121(4):847–52.
[20] Shewry PR, Jenkins JA, Beaudoin F, Mills ENC. The classification, functions and evolutionary relationships of plant proteins in relation to food allergies. In: Mills CEN, Shewry PR, editors. Plant food allergens. Blackwell Sciences; 2004. p. 24–86.
[21] Kumar KK, Sheloker PS. An SVM method using evolutionary information for the identification of allergenic proteins. Bioinformation 2008;2(6):253–6.
[22] 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. http://dx.doi.org/10.1371/journal.pone.0005861.
[23] Saha S, Raghava SPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acid Res 2006;34. http://dx.doi.org/10.1093/nar/gkL343.
[24] Tomar N, De RK. Immunoinformatics: an integrated scenario. Immunology 2010;131(2):153–68.
[25] Lundegaard C, Lund O, Nielsen M. Predictions versus high-throughput experiments in T-cell epitope discovery: competition and synergy? Expert Rev Vaccines 2012;11(1):3–54.
[26] Borges J-P, Barre A, Culerrier R, Archimbaud N, Didier A, Rougé P. How reliable is the structural prediction of IgE-binding epitopes of allergens? The case study of plant lipid transfer proteins. Biochemie 2007;89(1):83–97.
[27] Krieger E, Koraimann G, Vriend G. Increasing the precision of comparative models with YASARA NOVA – a self-parameterizing force field. Proteins 2002;47(3):393–402.
[28] Bragin AO, Demenkov DS, Kolchanov NA, Ivanisenko VA. Accuracy of protein allergenicity prediction can be improved by taking into account data on allergenic protein discontinuous peptides. J Biomol Struct Dyn 2013;13(1):59–64.
[29] Wang J, Zhang D, Li J. PREAL: prediction of allergenicity protein by maximum relevance minimum redundancy (mRMR) feature selection. BMC Syst Biol 2013;7(5):59.
[30] Liu y, Gu W, Zhang W, Wang J. Predict and analyse protein glycation sites with the mRMR and IFS methods. BioMed Res Int 2015. http://dx.doi.org/10.1155/2015/561549.
[31] Magnan CN, Balidi P. Sspro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profile, machine learning and structural similarity. Bioinformatics 2014;30(18):2592–7.
[32] Bjorklund AK, Soeria-Atmadja D, Zorzet A, Hammerling U, Gustafsson MG. Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. Bioinformatics 2005;21(1):39–50.
[33] Barrio AM. EVALLER: a web server for *in silico* assessment of potential protein allergenicity. Nucleic Acids Res 2007;35(Suppl 2):W694–700.
[34] Dimitrov I, Bangor I, Flower DR. AllerTop v 2 – a server for in silico prediction of allergens. J Mol Model 2014;20(6):2278.
[35] Dimitrov I, Flower DR, Doytchinova I. AllerTop – a server for in silico prediction of allergens. BMC Bioinform 2013;14(6):S4.
[36] Minkiewicz P, Dziuba J, Iwaniak A, Dziuba M, Darewicz M. BIOPEP database and other programs for processing bioactive peptide sequences. J AOAC Int 2008;91:965–80.
[37] Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, et al. The Immune Epitope Database 2.0. Nucleic Acids Res 2010;38:D854–62.
[38] Sen TZ, Jernigan RL, Garnier J, Kloczkowski A. GOR V server for protein secondary structure prediction. Bioinformatics 2005;21(11):2787–8.
[39] Minkiewicz P, Sokolowska J, Darewicz M. The occurrence of sequence identical with epitopes from the allergen pen a 1.0102 among food and non-food proteins. Pol J Food Nutr Sci 2015;65(1):21–9.
[40] Goodman RE, Silvanovich A, Hileman RE, Bannon GA, Rice EA, Astwood JD. Bioinformatic methods for identifying known or potential allergens in the safety assessment of genetically modified crops. Comments Toxicol 2002;8(3):251–69.

Please cite this article in press as: Hayes M, et al. *In silico* tools for exploring potential human allergy to proteins, Drug Discov Today: Dis Model (2016), http://dx.doi.org/10.1016/j.ddmod.2016.06.001

[41] Wold S, Jonsson J, Sjostrom M, Sandberg M, Rännar S. DNA and peptide sequences and chemical processes multi-variably modelled by principal components analysis and partial least square projections to latent structures. Anal Chim Acta 1993;277:239–53.
[42] Dimitrov I, Flower DR, Doytchinova I. AllerTOP – a server for *in silico* prediction of allergens. BMC Bioinform 2013;14:S4.
[43] Pierriere G, Combet C, Penel S, Blanchet C, Thioulouse J. Integrated databanks access and sequence/structure analysis services at PBIL. Nucleic Acids Res 2003;31(13):3393–9.
[44] Soeria-Atmadja D, Lundell T, Gustafsson MG, Hammerling U. Computational detection of allergenic proteins attains a new level of accuracy with in silico variable-length peptide extraction and machine learning. Nucleic Acids Res 2006;34(13):3779–93.
[45] Larsen JE, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Res 2006;2:2. http://dx.doi.org/10.1186/1745-7580-2-2.
[46] Zhang W, Xiang Y, Zhao M, Zou H, Ye X, Liu J. Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature. BMC Bioinform 2011;17. http://dx.doi.org/10.1186/1471-210512-341.
[47] Rubinstein ND, Mayrose I, Pupko T. A machine-learning approach for predicting B cell epitopes. Mol Immunol 2008;46:840–7.

Please cite this article in press as: Hayes M, et al. *In silico* tools for exploring potential human allergy to proteins, Drug Discov Today: Dis Model (2016), http://dx.doi.org/10.1016/j.ddmod.2016.06.001