Review

Practical and predictive bioinformatics methods for the identification of potentially cross-reactive protein matches

Richard E. Goodman

Food Allergy Research and Resource Program, Department of Food Science & Technology, University of Nebraska, Lincoln, Nebraska, USA

A bioinformatics comparison of proteins introduced into food crops through genetic engineering provides a mechanism to identify those proteins that may present an increased risk of allergic reactions for individuals with existing allergies. The goal is to identify proteins that are known to be allergens or are so similar to an allergen that they may induce allergic cross-reactions. Three comparative approaches have traditionally been used, or considered for safety evaluations. One identifies any short (6 – 8) amino acid segment of the protein that exactly matches a known allergen sequence. The second is an overall primary sequence comparison using Basic Local Alignment Search Tool (BLAST) or FASTA to find matches of greater than 35% identity over 80 amino acids. The third is based on 3-D prediction programs to identify 3-D similarities that might predict potential cross-reactivity. The utility of each of these approaches was debated in the bioinformatics workshop. The consensus agreement from the expert workshop participants was that the short-segment match (*e.g.*, 6 – 8 amino acids) provides an unacceptably high rate of false positive matches and an uncertain rate of true positive matches, and was not particularly useful for an allergenicity evaluation performed in the context of comprehensive safety evaluation. There was no consensus regarding the most appropriate bioinformatics method, an acceptable scoring criteria for triggering closer examination subsequent to a positive match, or an acceptable scoring mechanism for ranking the utility of the various 3-D approaches that were discussed during the workshop. However, the general consensus was that the most practical approach at this time is to evaluate primary sequence identities to known allergens using either FASTA or BLAST. While there was good agreement that identities of greater than 35% over 80 or more amino acids (recommended by Codex in 2003) is quite conservative, the conclusion was that additional data or studies would be needed to justify changing this criterion as there is some evidence that some individuals sensitized to proteins in evolutionarily conserved protein families may experience cross-reactions to proteins sharing approximately 40% identity.

Keywords: Allergen / Conformational epitope / Discontinuous epitope / Epitope / IgE

Received: June 29, 2005; revised: February 11, 2006; accepted: April 8, 2006

1 Introduction

Genetic engineering (GE) and food processing methods are both being used to introduce specific beneficial proteins into foods and consumer products. Regulatory bodies in various countries differ in their requirements for evaluating the safety or potential risks of products that include these newly introduced proteins. The US, EU, and many other countries have now implemented requirements for evaluating the potential allergenicity of such products [1]. The primary focus, either elaborated or implied, is on preventing the transfer of a known allergen, or a protein sufficiently similar to a known allergen that it may trigger allergic cross-reactions. Specific methods are used to reduce the possibility of transferring an allergen or cross-reactive protein. One of the most informative tests is the use of computer programs to compare the amino acid sequence of the

Correspondence: Professor Richard E. Goodman, Food Allergy Research and Resource Program, Department of Food Science & Technology, University of Nebraska, Lincoln, Nebraska, USA
E-mail: rgoodman2@unlnotes.unl.edu
Fax: +1-402-472-1693

Abbreviations: BLAST, Basic Local Alignment Search Tool; **GE,** genetic engineering; **FAO/WHO,** Food and Agriculture Organization/World Health Organization introduced protein with those of known allergens. Any significant resulting matches to a known allergen are used to identify a group of potentially at-risk individuals who could be tested to evaluate potential clinical risks by serum testing, skin prick testing, or even food challenge. The use of an effective system of comparison is important for the success of the allergenicity assessment.

The ideal situation for the allergenicity prediction would be to determine whether all biologically important IgE-binding epitopes were known for all of the major allergens, and if computer programs could predict biologically important similarities between a new protein and the epitopes of an allergen. Unfortunately, very few epitopes have been thoroughly mapped for even a few allergens using sera from representative allergic populations. Furthermore, while IgE epitopes may consist of short-sequential amino acid segments, in some cases they are formed by 3-D structures produced by the arrangement amino acids that are spatially close due to protein folding, and are therefore called discontinuous epitopes [2]. Even for some of the most highly studied allergens, close evaluation of the sequences and 3-D structures have demonstrated variability between epitopes recognized by individuals or between apparent cross-reactive structures that are not perfectly conserved or predictable [3, 4]. What we do know about many important allergens and even minor allergens is the primary (amino acid) sequence of the protein. Therefore, the evaluation of protein sequence and structure have focused on the use of general local sequence alignment algorithms such as FASTA or Basic Local Alignment Search Tool (BLASTP) that are frequently used in academic research to efficiently identify sequences from related species that are likely to be homologous. Empirical results demonstrate that proteins that are closely matched in sequence have similar structure and the most highly similar protein matches found with these programs often correspond to antibody cross-reactivity and clinical reactivity [5].

On the basis of scientific data available in the early 1990s, the first widely published comprehensive recommendation for evaluating the potential allergenicity of genetically engineered crops suggested performing a local alignment by FASTA or BLASTP to identify probable homologs and then identify any exact matches of eight contiguous amino acids shared between the query sequence (GE protein) and any allergen [6]. Scoring parameters (gap penalties, mismatch penalties, *etc.*) of both FASTA and BLAST programs can be modified so that results can differ markedly. While there are generally accepted default criteria used by either program to identify probable homologs, those criteria were not specified by Metcalfe *et al.* [6]. Furthermore, there has not been a generally recognized level of identity between two proteins that is considered “significant” in regard to the potential for cross-reactivity. As explained in the original publication, the criterion of an eight amino acid match was meant to identify potentially shared IgE or T-cell epitopes [6]. However, questions have been raised by many authors about the possibility of missing important matches using this criterion [7, 8]. The Food and Agriculture Organization/World Health Organization (FAO/WHO) 2001 scientific panel recommended a dual test, one looking for any match of six contiguous amino acids, the second looking for identity matches above 35% over any 80 amino acid segment of the query protein compared to any known allergen. The predictive value of an eight amino acid match, or even smaller matching segments such as six, however, had not been tested in published studies until around 2002. The FAO/WHO panel report stimulated a number of efforts as reported in three studies that demonstrated that more appropriate criteria are needed. Hileman *et al.* [9] found that roughly 80% of randomly chosen protein sequences of maize match an allergen if a six amino acid match is used, while comparisons for eight amino acid matches or those with >35% identity over 80 amino acids were more accurate in identifying proteins with overall FASTA alignments that would indicate an increased potential for shared sequential and conformational epitopes. Kleter and Peijnenburg [8] found also that six amino acid matches also identify matches unlikely to cause cross-reactions, and they investigated using a subsequent evaluation for antigenicity prediction for any matched proteins.

There have also been suggestions that structural comparisons or motif recognition patterns would provide better predictions for evaluating the potential allergenicity of novel proteins [10, 11]. But to date, these methods have generally not been used to evaluate a wide range of proteins in the context of the allergenicity assessment.

As the sequence comparison, or bioinformatics, is a very important part of the safety assessment process for genetically engineered crops, and significant questions have been raised about the best methods to perform such tests, the International Life Sciences Institute–Health and Environmental Sciences Institute sponsored a scientific workshop in Mallorca, Spain in February 2005 to address these questions. Participating scientists included a broad spectrum of experts in bioinformatics, food safety, and allergy/allergenicity.

This chapter will provide a brief description of bioinformatics methods that have been used, or proposed for regulatory studies evaluating proteins from genetically engineered crops that were discussed at the meeting. Key points of the discussion will be reviewed, and the points of consensus or majority opinions voiced by the experts will be presented.

2 Allergen databases

Allergen specific sequence databases [12] are very useful for improving the efficiency of the computerized sequence comparisons to identify potentially cross-reactive allergens, in contrast to searches using a more generalized sequence database such as NCBI or Swiss-Prot. A general database screen would require significantly more manual data analysis of identified matches in order to evaluate the allergenicity of the matched sequences. In 1996, public allergen databases were not available on the internet. Researchers could compile their own list from GenBank or Swiss-Prot by searching protein or cDNA sequences with query terms such as "allergen." Now several databases of allergens (*e.g.*, Allergenonline: http://allergenonline.com) are available on the internet and can be queried with the amino acid sequence of any protein. Additionally, 3-D or structural data are available for a few of the most studied allergens. The focus of this paper is on the methods that can be used to compare a sequence to the database to evaluate the similarity of any protein to known or putative allergens.

Two cautions are in order regarding "allergens" in any database. First, even the most carefully curated database will have a number of proteins included for which there is little objective evidence for allergenicity, and conversely, all will miss a few allergens. Second, the range of potencies, or incidence of allergy associated with various proven allergens is wide, as is the potential reactivity if the protein is presented in food, or *via* the airway. These factors should be kept in mind when considering the potential risk of allergy from an introduced protein having a sequence that matches any allergen.

3 Short contiguous amino acid matches

The practice of searching for identical short amino acid matches between the introduced GE protein and sequences from known allergens had not been evaluated in terms of efficiency until recently [9, 10]. However, short peptide sequence matches have been used to evaluate most, if not all, of the proteins introduced into commercially available GE crops that have been reviewed by US, Japan and EU regulators. Summary safety data for a number of GE products are available online (http://usbiotechreg.nbii.gov/database_pub.asp; http://www.agbios.com/dbase.php). The original recommendation [6] suggested first aligning the query protein (introduced GE protein) with any known allergen be performed using FASTA or BLAST and the resulting alignments of presumably homologous sequences be evaluated for any identical eight amino acid match. Such a match was thought to indicate that the GE protein might cause cross-reactions in those allergic to the matched allergen as proteins having an overall alignment which suggests overall structural similarity and evolutionary divergence, which also has significant short sequence matches, may have a higher probability of sharing IgE epitopes. However, in practice, these searches have been performed using simple algorithms that were developed to find identical strings, or words, much like the search function of a word processing program [9]. Each individual possible contiguous amino acid sequence segment of the protein was used to search a selected allergen database for an exact match. The first search cycle would compare amino acids one to eight of the query to the database, the second cycle would query amino acids two to nine. The process would continue until the final eight amino acid segment is tested. Currently, there are at least two websites (Allermatch: http://allermatch.org and the Structural Database of Allergenic Proteins (SDAP: http://fermi.utmb.edu/SDAP/) that provide short-segment (6 or 8 amino acids) matching algorithms. However, since these procedures do not first screen for potential homology, the results may be significantly different than anticipated by the original authors [6]. When direct string matches have been performed, it is clear from the high number of matches using common protein sequences that this method greatly overestimates probable allergenic matches [9]. The data demonstrate that the high false positive rate observed with six or seven amino acid matches is unacceptable and that the eight amino acid matches may be irrelevant. There are no clear data to demonstrate that this approach adds value over a properly performed local sequence alignment.

4 Local alignments with FASTA or BLASTP

FASTA [13] and BLAST [14] are computer algorithms that were written to provide efficient computer comparisons of nucleic acid or protein sequences derived from distantly related organisms. The algorithms scan each sequence in a database using the query sequence to identify short "seed" matches. If a minimal match is found, the sequence match is optimized to provide a best overall match for that sequence. As the entire database is scanned, a list is formed with the best matches at the top. Available scoring matrices that may be selected for use by the algorithm (BLOSUM 50, PAM 62, PAM 250, *etc.*) were developed with biases toward more (or less) conservation of structure based on structurally dominant amino acids (*e.g.*, proline, tyrosine, phenylalanine) and frequencies of occurrence. Additional scoring factors such as gap extension penalties (corrections for the insertion of an artificial gap in a sequence, to optimize the alignment) will alter the alignment and scores. The default settings for FASTA run on Allergenonline are BLOSUM 50, with ktup (seed word size) of 2, gap penalty of –10, and gap extension penalty of –2. The resulting Z score is converted into an E score that includes a correction for database size and complexity. The smaller the E score the more similar two sequences are, and the greater the likelihood they share overall structure and evolutionary heritage. In general, an *E* score value less than 0.02 indicates probable homology [15]. However, based on experience, *E* scores between 10^{-2} and 10^{-7} are relatively common between protein sequences that are evolutionarily related, but do not share histories of allergic cross-reactivity [9]. Since *E* scores will vary with the size of the database, and scoring matrices (*e.g.*, vicilin of walnut, GI 6580762 matches lentil allergen Len c 1, GI 29539109 with an *E* score of 4.3×10^{-11} using FASTA with Allergenonline version 5.0, but the *E* score using BLAST with NCBI results in an *E* score of 9×10^{-82}), while percentage identities are constant (39% identity over 410 amino acid alignment for the same alignment of walnut and lentil proteins), it is reasonable to simply evaluate the *E* score for any match to decide whether the query protein sequence represents a close homologue of any allergen.

What value of shared identity might represent a realistic risk of cross-reactivity? Pearson [15], who developed the FASTA algorithm, observed that proteins that are apparently quite distantly related through evolution may share as little as 20–25% identity over the majority of the length of the encoded polypeptides and still be considered homologs. These proteins will usually share common structural folds and either similar or related functions. However, Pearson [15] also noted that two quite distinct proteins that do not share overall structure or function may by chance share 50% identity over a segment of 20–40 amino acids.

Aalberse [16] observed that it is rare to find true clinical cross-reactivity in a single patient if proteins share less than 50% identity over their full-lengths, while proteins sharing greater than 70% identity are commonly cross-reactive. At the same time, the FAO/WHO expert panel recommended that an identity greater than 35% over 80 or more amino acids should be used as a guideline to suggest possible cross-reactivity (Food and Agriculture Organization of the United Nations. FAO Corporate Document Repository. http://www.fao.org/documents/show_cdr.asp?url_file=/docrep/007/y0820e/y0820e00.htm). Some very elegant studies have been performed that evaluate the relationship between clinical reactivity, *in vitro* binding, and histamine release of homologous allergens that span a wide range of sequence identity [5]. Those studies have shown marked reduction in IgE binding of 100–1000-fold for proteins sharing only ~40% identity. However, the examples may overestimate true cross-reactivity because individuals are exposed to related allergens and may be sensitized to one or more of the related homologs [17, 18]. Few other detailed studies have been performed to evaluate the relationship between cross-reactivity and sequence or structural similarity with a quantitative measurement of binding efficacy.

While some may argue in favor of adjusting (or not) the criteria of similarity matches with 35% identity over 80 amino acids, it is important to recognize that a match of greater than 50% to 70% identity over the full-length of two proteins is much more likely to indicate potential *in vitro* cross-reactivity and/or clinical relevance. Although the direct proof of clinical cross-reactivity is nearly impossible to obtain, there are many studies that demonstrate the positive correlation between percent identity and likelihood of *in vitro* IgE binding and clinical reactivity. For example, Beyer *et al.* [19] found that while 12 of 14 hazelnut allergic individuals had serum IgE that bound to the 11S hazelnut globulin, only about one-half experienced any clinical reactivity to the relatively unrelated peanut, walnut, brazil nut, cashew nut or almond, and only one of those experienced reactions to more than two. This lack of clinical reactivity occurs despite the fact that the 11S globulins that have been sequenced from that group (hazelnut, peanut, brazil nut, cashew) share between 45 and 55% identity. Since the 11S albumins are major seed storage proteins, they would all be expected to share clinical symptoms if 45–55% identity was likely to indicate cross-reactivity. Sanchez-Monge *et al.* [20] found that 18 of 18 individuals with allergies to garden peas are also allergic to lentils. They identified two related pea proteins, vicilin and convicilin, and fragments thereof, which appear responsible for essentially all IgE reactivity in pea extract. Pea and lentil vicilins are ~90% identical, the two convicilins are over 70% identical and within each species, vicilin and convicilin share approximately 60% identity.

Based on the current evidence, as well as the general observations of Aalberse [16], matches of greater than 50% identity by FASTA or BLAST should be evaluated very closely to determine whether a protein of interest may lead to cross-reactions in individuals with existing allergies, while a match of less than 35–40% identity is likely to represent a much lower probability of cross-reactivity. Further scientific evaluation of this issue will likely require specific *in vitro* IgE-binding studies using well-characterized sera from individuals allergic to the matched allergen. The data suggests that 35% identity over 80 amino acids represents a very conservative limit for triggering further examination. In fact, this limit may be much lower than necessary to still protect most allergic individuals from the introduction of a cross-reactive protein. However, additional data from well-controlled studies would likely be required to justify raising the limit above 35% identity over 80 amino acids.

5 Structural comparisons

As the 3-D structure of more allergens have been either directly (X-ray crystallography or NMR) or indirectly (computer prediction) analyzed and compared, it is clear that many important allergens can be grouped into a small number of structural families [21]. This implies that the structural folds of proteins within the family are similar. IgE binding to many allergens is conformational, meaning the epitope may be comprised of amino acids that are not adjacent in primary sequence. Structurally similar proteins are likely to share conformational epitopes. Since primary structure (sequence) is an essential determinant for secondary structure, and secondary structure can alter the fine-positioning of the side-groups of amino acids in nonlinear epitopes, the ability of a given IgE antibody to bind to any region on a protein may be altered by the surrounding protein structure. Therefore, one could argue that the most appropriate way to evaluate potential cross-reactivity would include a comparison of 3-D structures.

There are various programs and approaches that could be used to evaluate the 3-D structure of a protein, given the primary structure [10, 11, 21]. The size, hydrophobicity, polarity, and charge of the side-groups of each amino acid contributes to the 3-D structure. Adjacent and even relatively distant amino acids in the primary sequence will interact to influence the final protein shape. One can imagine that various factors and methods are used by different programs to predict overall and fine-structures of proteins, but in all cases the primary amino acid sequence plays an important role in establishing the structure. The local environment (solvent, pH, salt) can alter the overall shape. Some structural prediction programs evaluate similarities in the overall sequences of proteins for specific secondary structures (*e.g.*, likely disulfide bonds, alpha helices, turns) that are likely to dominate secondary shape and may represent an antibody epitope. Others focus on segments of the full-sequence, predicting local structures independent of the overall structure or they predict the overall structure and plot surface exposure [22]. For selected cross-reactive pairs of allergenic proteins, one or more of these structural prediction algorithms is likely to be highly predictive for estimating potential cross-reactivity. However, there do not appear to be any uniform measurements of similarity that have been identified to predict probable cross-reactivity over a widely divergent set of proteins. No one has published a measurement analogous to a percent identity (over the full-length, or even specified segment), that is quantifiable and tested on a diverse set of proteins that are known to be cross-reactive or not cross-reactive. In the future, it may be possible to develop a unified 3–D prediction tool and criteria that would be useful for predicting potential cross-reactivity; however, further evaluation and additional data are needed to determine the feasibility of such an approach.

6 Consensus and summary

The participants at the bioinformatics workshop in Mallorca represented a wide range of scientific expertise in bioinformatics, structural biochemistry, allergy, immunology, food science, and food safety regulations. A number of participants had previously been involved in assessing the potential allergenicity of GE crops either as a reviewer, a producer, a regulator, or an expert advisor to FAO/WHO or some country. Discussions happened throughout the workshop about the predictive power of various sequence and structural comparison methods. New data were presented relative to the high incidence of randomly selected short amino acid sequences [23]. Therefore, a formal question was posed to all regarding whether there is a useful predictive value in continuing to perform a short, six, seven, or eight amino acid sequence match. There was agreement that searches for short matches are not predictive and should not be used to evaluate the potential allergenicity of proteins.

There was apparent agreement that structural comparisons may be very useful for evaluating the cross-reactivity of two proteins, and predicting changes that might alter antibody binding. However, much of these data were apparently obtained from allergens for which some of the IgE epitopes have already been mapped. It was not clear, however, whether these methods would be predictive for proteins that are not highly similar in overall sequence identity. There was a vigorous debate as to which approach would be most predictive or whether any have the ability to predict potential cross-reactivity for proteins that are not closely related to a known allergen. There was also a lack of clarity about how structural similarities might be scored to provide guidance for the safety assessment.

There was agreement among workshop participants that FASTA or BLAST algorithms comparing a query sequence to those of known allergens is an efficient way to identify proteins that should be studied further for potential cross-reactivity. There was sufficient debate to conclude that there is currently not enough data to change (*i. e.*, increase) the recommended guideline (*i. e.*, greater than 35% identity over any segment of 80 or more amino acids to any allergen requires additional testing with human serum (Food and Agriculture Organization of the United Nations. FAO Corporate Document Repository. http://www.fao.org/documents/show_cdr.asp?url_file=/docrep/007/y0820e/y0820e00.htm and Codex Alimentarius Commission, 2003. Food and Agriculture Organization of the United Nations. FAO Corporate Document Repository.http://www.fao.org/documents). While a number of scientists pointed out examples where proteins with 50% identity had not been shown to cause *in vitro* or *in vivo* cross-reactions, there were a few examples where proteins of approximately 40% identity shared some degree of *in vitro* cross-reactivity that correlated with clinical reactions to the sources of those allergens. However, it was also argued that in at least some cases, the individuals might have been sensitized to other similar proteins, skewing the data. While a few individuals expressed some concern that there may be an instance when a FASTA or BLAST search looking for >35% identity over 80 amino acids might miss a few cross-reactive matches, the general opinion seemed to be that this criteria is quite conservative and should continue to be used in the allergenicity assessment of GE proteins.

7 References

[1] Goodman, R. E., Hefle, S. L., Taylor, S. L., van Ree, R., *Int. Arch. Allergy Immunol.* 2005, *137*, 153–166.

[2] Bannon, G. A., Ogawa, T., *Mol. Nutr. Food Res.* 2006, *50*, DOI: 10.1002/mnfr.200500276.

[3] Jarvinen, K. M., Chatchatee, P., Bardina, L., Beyer, K., Sampson, H. A., *Int. Arch. Allergy Immunol.* 2001, *126*, 111–118.

[4] Mirza, O., Henricksen, A., Ipsen, H., Larsen, J. N. *et al.*, *J. Immunol.* 2000, *165*, 331–338.

[5] Scheurer, S., Son, D. Y., Boehm, M., Karamloo, F. *et al.*, *Mol. Immunol.* 1999, *36*, 155–167.

[6] Metcalfe, D. D., Astwood, J. D., Townsend, R., Sampson, H. A. *et al.*, *Crit. Rev. Food Sci. Nutr.* 1996, *36*, S165–S186.

[7] Gendel, S. M., *Adv. Food Nutr. Res.* 1998, *42*, 45–62.

[8] Kleter, G. A., Peijnenburg, A. A., *BMC Struct. Biol.* 2002, *12*, 8.

[9] Hileman, R. E., Silvanovich, A., Goodman, R. E., Rice, E. A. *et al.*, *Int. Arch. Allergy Immunol.* 2002, *128*, 280–291.

[10] Stadler, M. B., Stadler, B. M., *FASEB J.* 2003, *17*, 1141–1143.

[11] Ivanciuc, O., Schein, C. H., Braun, W., *Bioinformatics* 2002, *18*, 1358–1364.

[12] Gendel, S. M., Jenkins, J. A., *Mol. Nutr. Food Res.* 2006, *50*, DOI: 10.1002/mnfr.200500271.

[13] Pearson, W., Lipman, D., *Proc. Natl. Acad. Sci. USA* 1988, *85*, 2444–2448.

[14] Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., *J. Mol. Biol.* 1990, *215*, 403–410.

[15] Pearson, W. R., *Methods Enzymol.* 1996, *266*, 227–258.

[16] Aalberse, R. C., *J. Allergy Clin. Immunol.* 2000, *106*, 228–238.

[17] Ferreira, F., Hawranek, T., Gruber, P., Wopfner, N., Mari, A., *Allergy* 2004, *59*, 243–267.

[18] Salcedo, G., Sanchez-Monge, R., Diaz-Perales, A., Garcia-Casado, G., Barber, D., *Clin. Exp. Allergy* 2004, *34*, 1336–1341.

[19] Beyer, K., Grishina, G., Bardina, L., Grishin, A., Sampson, H. A., *J. Allergy Clin. Immunol.* 2002, *110*, 517–523.

[20] Sanchez-Monge, R., Lopez-Torrejon, G., Pascual, C. Y., Varela, J. *et al.*, *Clin. Exp. Allergy* 2004, *34*, 1747–1753.

[21] Jenkins, J. A., Griffiths-Jones, S., Shewry, P. R., Breiteneder, H., Mills, E. N. C., *J. Allergy Clin. Immunol.* 2005, *115*, 163–170.

[22] Kulkarni-Kale, U., Bhosle, S., Kolaskar, A. S., *Nucleic Acids Res.* 2005, *33*, W168–W171.

[23] Silvanovich, A., Nemeth, M. A., Song, P., Herman, R. *et al.*, *Toxicol. Sci.* 2006, *90*, 252–258.