Sequence Analysis for Assessing Potential Allergenicity

STEVEN M. GENDEL

Biotechnology Studies Branch, National Center for Food Safety and Technology, Summit-Argo, Illinois 60501, USA

ABSTRACT: Sequence analysis plays an important role in assessing the potential allergenicity of proteins used in transgenic foods, particularly for proteins that have not previously been part of the food supply. Sequence comparisons are used to indicate potential unexpected cross reactivity to existing allergens and to assess the potential for developing new sensitivities. Although the concept of using sequence analysis is straightforward, implementing a bioinformatic analysis that is accurate and complete can be complex. Several factors need to be considered, including the design and content of the sequence database, the analysis strategy, and the criteria for evaluating the results.

KEYWORDS: bioinformatic sequence analysis; allergenicity; allergen sequence database

INTRODUCTION

Bioinformatic sequence analysis plays an important role in assessing the potential allergenicity of proteins used in transgenic foods, particularly for proteins that have not previously been part of the food supply. Sequence comparisons are used to indicate potential unexpected cross reactivity to existing allergens and to assess the potential for developing new sensitivities. For example, Metcalfe *et al.*[1] explicitly recommend the use of sequence analysis for proteins derived from sources that are not known to be allergenic (FIG. 1), as have other expert panels.[2] Although the concept of using sequence analysis is straightforward, implementing a bioinformatic analysis that is accurate and complete can be complex.

Four major issues should be considered when using sequence analysis for assessing potential allergenicity: whether to compare amino acid or nucleic

Address for correspondence: Steven M. Gendel, Ph.D., Biotechnology Studies Branch, National Center for Food Safety and Technology, U.S. Food and Drug Administration, 6502 S. Archer Road, Summit-Argo, IL 60501. Voice: 708-728-4110; fax 708-728-4177.
sgendel@cfsan.fda.gov

Ann. N.Y. Acad. Sci. 964: 87–98 (2002). **© 2002 New York Academy of Sciences.**

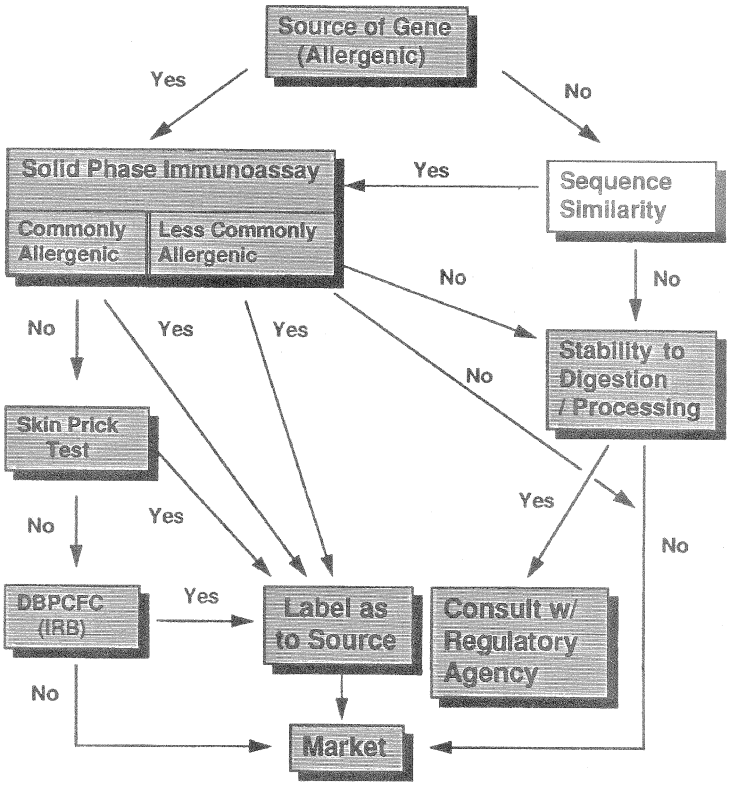


FIGURE 1. The allergenicity assessment flowchart from Metcalfe *et al.*[1] showing the importance of sequence analysis for assessing the potential allergenicity of proteins from a nonallergenic source in this scheme. (Used with permission.)

acid sequences, the composition and design of the allergen sequence database, the algorithms and strategies used, and the criteria for evaluating the results.

AMINO ACID VERSUS NUCLEIC ACID COMPARISONS

Virtually all sequence data obtained in the last 10 years have been nucleic acid sequences; the corresponding amino acid sequences have been derived by translation of open reading frames.[3,4] Therefore, allergenicity assessment could be carried out using either the primary nucleic acid sequences or the derived amino acid sequences. Although many sophisticated tools are available for nucleic acid sequence analysis, there are several reasons for using amino acid sequences in an allergenicity assessment. First, use of the amino acid se-

quences avoids problems with codon degeneracy that might obscure relationships. Two sequences, especially short peptides, might have a high degree of similarity at the amino acid sequence level but differ significantly at the nucleic acid sequence level because of codon degeneracy and differences in codon preferences in the host organisms. Second, the use of amino acid sequences eliminates the need to consider the potential presence of introns and regulatory sequences. Nucleic acid sequences are generally determined either from genomic clones or cDNA. In either case, the resulting nucleic acid sequence will contain more than the coding sequence. The presence of these extraneous sequences makes it more difficult to assess similarities within coding regions, particularly when introns interrupt regions of similarity or occur at different sites within homologous genes. Third, amino acid comparisons simplify consideration of amino acid substitutions. Protein similarities often involve evolutionarily or biochemically conservative substitutions that cannot be directly recognized at the nucleic acid sequence level (see below). Overall, analysis of amino acid sequence similarities provides the most relevant and direct information for assessing potential allergenicity.

DATABASE COMPOSITION AND DESIGN

Sequence similarity analysis involves comparison of a query sequence to each member sequence in a database containing known allergen sequences (this comparison is also known as an alignment). Therefore, the design and composition of the database used is critical in determining the accuracy and value of the analysis. Two types of databases can be used: large, general purpose, primary databases and smaller, specialized, secondary databases. Examples of primary protein sequence databases include GenPept (derived by translation from GenBank),[5] the SwissProt database,[6] and the Protein Information Resource (PIR) database.[7]

Several difficulties are inherent in using large primary resources for allergenicity assessment. The structure and annotation of these databases make it impossible to locate all relevant sequences and to avoid irrelevant sequences. Keyword filters or searches are incomplete because many allergen sequences (especially food allergens) are not identified as such in the annotation and imprecise because the annotations of other proteins contain references to allergen recognition, processing, and so forth. An example of the first problem is the milk protein allergen α-lactalbumin, which is not identified as an allergen in any of the large databases (FIG. 2). The second problem is exemplified by the large number of immunoglobulin sequences from different species that mention allergy in the annotation because of their role in the immune response (FIG. 3).

A second problem arising from the use of large databases is that inconsistencies exist that make it difficult to identify identical or related sequences in

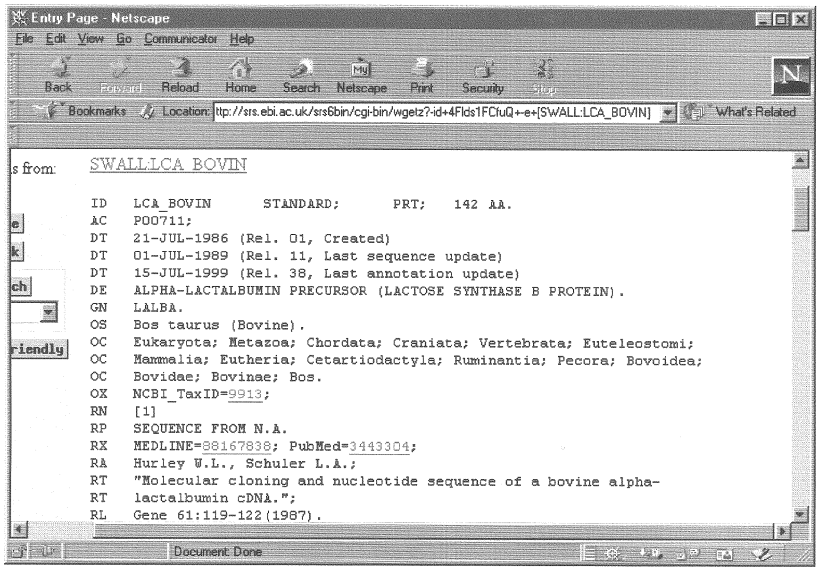


FIGURE 2. An example from the SwissProt database of the annotation accompanying the sequence of a known allergen (milk α-lactalbumin) that does not identify the protein as an allergen.

multiple databases. Databases differ in the use of accession numbers and/or entry identifiers, the amount and type of information present in the annotation, and the treatment of overlapping or contiguous sequences. Further, although most primary databases provide cross-references to related entries in other databases, there are numerous errors in these pointers.

In addition, the comprehensive nature of these large databases and the fact that they were primarily designed for evolutionary analysis can create problems when they are used for allergenicity assessments, particularly when the query sequence is a member of a protein family. For example, a query of the SwissProt database with the Cry1a sequence resulted in 120 "hits." Of these, 112 were other Cry proteins, and the highest ranked non-Cry protein on the list was 108 on the list. Given that the purpose of an allergenicity assessment is to find unexpected sequence similarities with known allergens, not obvious evolutionary relationships, it would be very difficult to use the results of this query.

Finally, one problem for which there is no solution at this time is the fact that not all sequence information in the databases is searchable! It is common to find information on minor forms, variant sequences, or other inconsistencies in the annotation (FIG. 4). The inability to access these "hidden sequences" with currently available software could prevent recognition of immunologically significant short sequence matches.

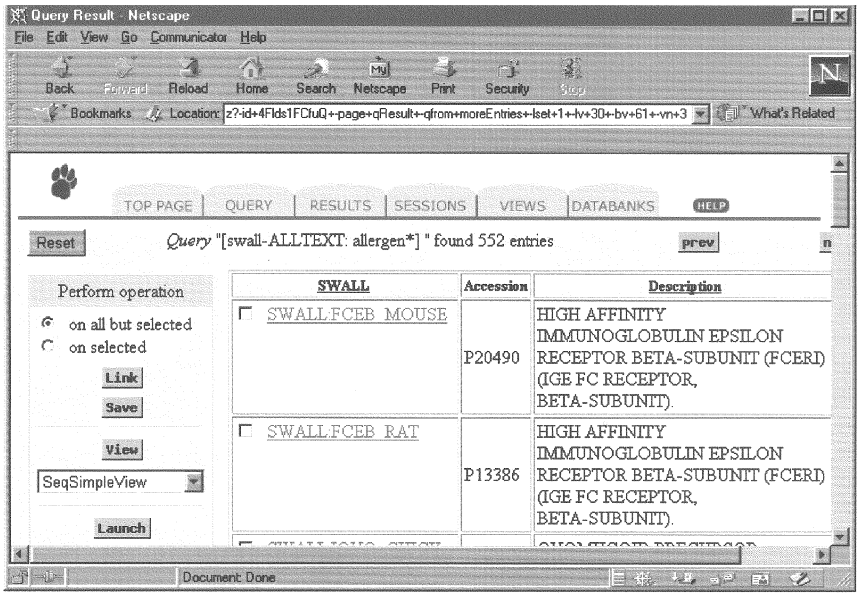
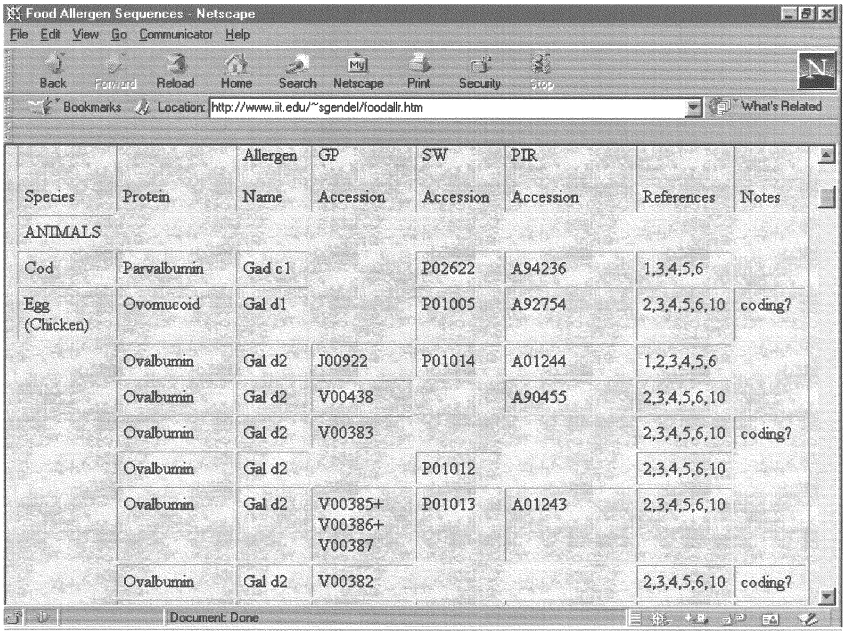


FIGURE 3. An example from the SwissProt database of some of the irrelevant immune system sequences that are retrieved using the keyword "allergen."

```
FT   VARIANT     115    115       E -> V (IN MINOR FORM).
FT   VARIANT     172    172       K -> A (IN MINOR FORM).
FT   VARIANT     185    185       K -> I (IN MINOR FORM).
FT   VARIANT     235    236       LG -> RA (IN MINOR FORM).
SQ   SEQUENCE   273 AA;  29325 MW;  05A0B1A8FAC7B159 CRC64;
     MKPFCVFLTF FLLLAASSKK VDSAETVSFN FNSFSEGNPA INFQGDVTVL SNGNIQLTNL
     NKVNSVGRVL YAMPVRIWSS ATGNVASFLT SFSFEMKDIK DYDPADGIIF FIAPEDTQIP
     AGSIGGGTLG VSDTKGAGHF VGVEFDTYSN SEYNDPPTDH VGIDVNSVDS VKTVPWNSVS
     GAVVKVTVIY DSSTKTLSVA VTNDNGDITT IAQVVDLKAK LPERVKFGFS ASGSLGGRQI
     HLIRSWSFTS TLITTTRRSI DNNEKKIMNM ASA
```

FIGURE 4. An example from the SwissProt database of sequence variant information that cannot be searched because it is present only in the annotation.

Therefore, it is preferable to use a specialized, secondary database of allergen protein sequences for allergenicity assessment. A secondary database can be designed and constructed to eliminate many of the problems discussed above. Specialized databases can contain data from multiple sources, and the efforts of the database administrator or curator ensure uniform data quality and presentation. Although several proprietary allergen databases exist in the biotech industry, only two allergen sequence databases have been published. The first, published in Metcalfe *et al.*,[1] does not exist in electronic form and has not been updated since the original publication. The second is the Protein



Species	Protein	Allergen Name	GP Accession	SW Accession	PIR Accession	References	Notes
ANIMALS							
Cod	Parvalbumin	Gad c1		P02622	A94236	1,3,4,5,6	
Egg (Chicken)	Ovomucoid	Gal d1		P01005	A92754	2,3,4,5,6,10	coding?
	Ovalbumin	Gal d2	J00922	P01014	A01244	1,2,3,4,5,6	
	Ovalbumin	Gal d2	V00438		A90455	2,3,4,5,6,10	
	Ovalbumin	Gal d2	V00383			2,3,4,5,6,10	coding?
	Ovalbumin	Gal d2		P01012		2,3,4,5,6,10	
	Ovalbumin	Gal d2	V00385+ V00386+ V00387	P01013	A01243	2,3,4,5,6,10	
	Ovalbumin	Gal d2	V00382			2,3,4,5,6,10	coding?

FIGURE 5. An example of the data contained in the PAD, including the source of the protein, the name of the protein, the allergen designation, accession numbers for each reference database, citation information, and notations regarding any questions or problems with the sequences. All accessions on a single line have exactly the same sequence, accessions for the same protein on different lines contain sequences that differ in either amino acid sequence or length or both. All citations and explanations are given at the bottom of the database table.

Allergen Database (PAD), which is part of the Biotechnology Information for Food Safety Database project at the National Center for Food Safety and Technology (NCFST).[8] The PAD is available through the NCFST web site (http://www.ncfst.iit.edu) and is continually being updated and expanded. It contains two main sections: food allergen sequences and nonfood allergen sequences. TABLE 1 summarizes the composition of the current release of the PAD. Each entry has been validated, and accessions from GenPept, Swiss-Prot, and the PIR with the same coding sequences have been identified. The PAD also differentiates multiple sequences for the same protein (FIG. 5).

COMPARISON ALGORITHMS

A variety of algorithms and implementations can be used when comparing a query sequence to an allergen sequence database. Although the ease of use

TABLE 1. Summary of the composition of the current release of the protein allergen database

	Food allergen database	Nonfood allergen database
Unique sequences	140	342
Proteins	48	228
Species	19	81
GenPept accessions	99	224
SwissProt accessions	63	184
PIR accessions	100	190

SOURCE: From Gendel.[8] Used with permission.

and structure of the output can vary depending on the implementation, any local alignment program is appropriate for use in an allergenicity assessment. The two most popular local alignment programs are BLAST[9] and the various versions of FASTA.[10] Although BLAST is more efficient for searching large databases, FASTA is preferred for allergenicity assessment, because it can be more sensitive in locating very short matches and it is easier to use with local databases.

FASTA finds potential sequence similarities by locating short, exact matches (usually two amino acids) between the query sequence and each member of the database. A "score" for the match is calculated on the basis of a scoring matrix that assigns a value (either positive or negative) to every possible amino acid pairing that can be produced in an alignment. Each initial exact match is extended by progressively considering the adjacent amino acid pairings in both directions, and a new total score is calculated as each additional amino acid pair is considered. This extension continues as long as the total score remains above a predefined threshold value. The final length of the match and the "score" used to calculate the significance of the match are defined by the furthest amino acid pairs to make a positive contribution to the final score. A match region may contain exact amino acid pairings, mismatches that have positive scores, and mismatches that have negative scores. Therefore, the overall length of the final similarity region and the degree of mismatch allowed depend on the scoring matrix used. Thus, it is important to understand the effects of different scoring systems when evaluating the use of FASTA and other local alignment algorithms for allergenicity assessment.

Most scoring systems, as instantiated in particular scoring matrices, assign a positive value to an exact amino acid match.[11,12] Many also assign positive values of varying magnitude to those mismatches that are considered to be conservative replacements. The actual values used are generally derived by

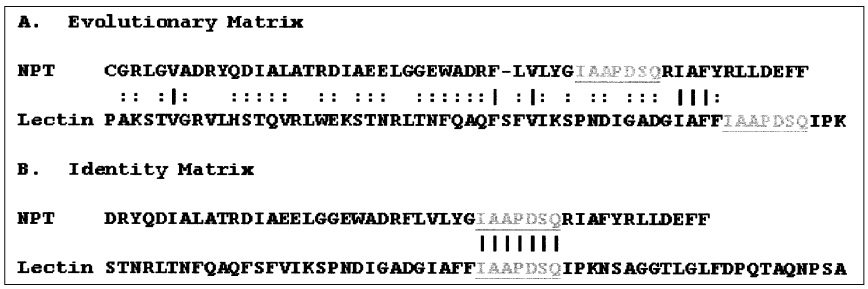


FIGURE 6. An example of the effect of using different scoring systems on the alignment produced between two protein sequences. **(A)** Alignment of the neomycin phosphotransferase protein with peanut lectin using an evolutionary matrix. Exact amino acid matches are indicated with a line (|), whereas evolutionarily conservative substitutions are indicated with a colon (:). **(B)** Alignment of the same sequences using an identity matrix. This analysis found the seven-amino-acid exact match that was not seen using the evolutionary matrix.

evolutionary analysis of families of related proteins and can therefore be considered to allow evolutionarily conservative mismatches. Scoring systems that allow biochemically or physically conservative mismatches can also be used, as can a system that only allows exact identity. The default scoring system used by all available local alignment algorithms is an evolutionary matrix.

The use of an evolutionary scoring matrix leads to situations in which the partial scores generated by a large number of evolutionarily conservative mismatches may obscure short immunologically relevant exact matches in the same region with lower total scores. For example, alignment of the amino acid sequence for the selectable marker neomycin phosphotransferase with a peanut lectin using an evolutionary scoring matrix produces the alignment shown in FIGURE 6A, with a large number of conservative replacements and no exact matches longer than three contiguous amino acids. However, alignment of the same sequences using an identity matrix shows a seven-amino-acid exact match in the same sequence regions (FIG. 6B). Because allergenicity assessments are specifically intended to identify short immunologically relevant matches in sequences that are known to be evolutionarily unrelated, it appears that the use of an identity matrix rather than the normal default matrix is preferable, at least for an initial analysis.

It is also important to note that some implementations of the various alignment algorithms use an "expectation value" to determine which scores are considered significant and are therefore reported in the program output. Expectation values are useful when searching large databases because they can eliminate reporting of irrelevant matches from dispersed repetitive elements (such as Alu sequences). However, immunologically relevant short matches

TABLE 2. Examples of minimal Ara h 2 IgE binding epitopes

Epitope	AA sequence	Ara h 2 position
1	**HASARQQWEL**	15–24
5	**KIQRDEDS**	49–58
6	**RDPYSP**	57–66
8	**LQGRQQ**	115–124
9	**KRELRN**	127–136
10	**QRCDLDVE**	143–152

SOURCE: Adapted from Stanley *et al.*[14]

may also be eliminated. In many cases, the user can set the threshold e-value as a program parameter. When this is the case, several values should be tested to ensure that no significant results are lost. If possible, a software implementation that does not use e-values should be used, particularly with smaller, specialized allergen databases and stringent scoring systems.

CRITERIA FOR EVALUATING RESULTS

Developing criteria for evaluating the results of database queries for assessing potential allergenicity remains a difficult problem. It is important to recognize, however, that questions about the most appropriate criteria for identifying immunologically significant matches are independent of database design or query method.

The only specific published criterion of potential allergenic significance is that of Metcalfe *et al.*,[1] which states that "an immunologically significant sequence identity requires a match of at least eight contiguous identical amino acids." Recently, several *in vitro* studies have characterized the epitopes of food allergen proteins in great detail. For example, Stanley *et al.*[13] identified minimal IgE-binding epitopes in the peanut allergen Ara h 2 ranging from 6 to 10 amino acids in length (TABLE 2). Further, Shin *et al.*[14] used *in vitro* amino acid replacements to show that the amino acids essential for antibody binding are not necessarily contiguous (TABLE 3) and that amino acid substitutions within an epitope could increase, decrease, or not affect antibody binding (FIG. 7).

If, as the data of Stanley *et al.*[13] and Shin *et al.*[14] suggest, the criterion of Metcalfe *et al.*[1] is too stringent, the use of a single scoring matrix for assessing potential allergenicity may not be efficacious. Instead, a two-step procedure could be used that would allow recognition of short (not necessarily contiguous) regions with high levels of identity as well as internal and adjacent conservative replacements.

TABLE 3. Amino acids essential for IgE binding to several Ara h 1 epitopes (underlined and in bold)

Epitope	Amino acid sequence
1	AKS**PY**Q**K**KT
2	QEP**DDL**KQKA
3	LE**YDP**RL**VY**D
4	GE**R**TR**GRQ**PG
5	PGDYDD**D**RRQ

SOURCE: Adapted from Shin *et al.*[15]

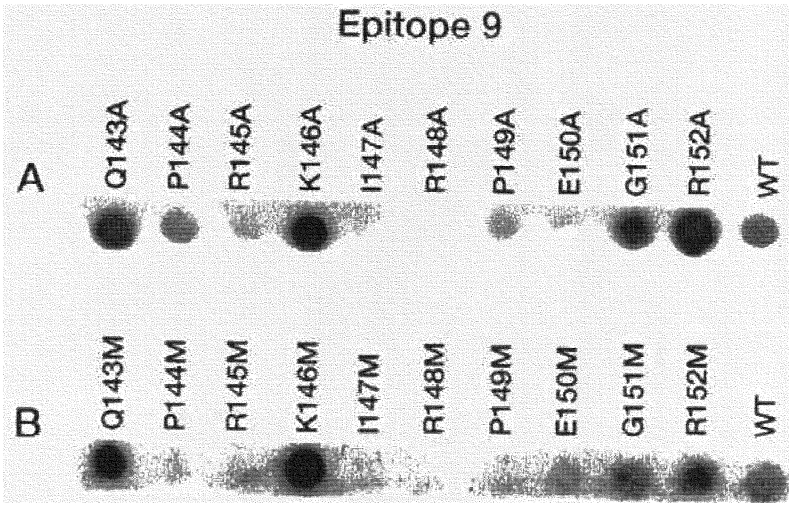


FIGURE 7. An example of the different effects of amino acid substitutions on antibody binding to peptide fragments derived from epitope 9 from Ara h 2. **(A)** Effect of substituting A for each naturally occurring amino acid. **(B)** Effect of substituting M for each naturally occurring amino acid. WT is the unmodified wild-type peptide. Some substitutions eliminate binding (for example at amino acid 148), while others increase binding (for example at amino acid 146). From Shin *et al.*[14] Used with permission.

An example of one approach to a two-step assessment is shown in FIGURE 8. In this case, a FASTA analysis was carried out with the Cry3A protein as the query sequence against the PAD. In the first step, an identity matrix was used as described above to locate regions with high levels of sequence identity. One region of relatively high identity (7 of 10 amino acids) was found with β-lactoglobulin (FIG. 8). In the second step, this region was realigned using an evolutionary matrix (PAM 250) to locate evolutionarily conservative replacements within this region. After this second alignment, 10 of 10 amino

```
ALIGNMENT USING IDENTITY SCORING

Cry3A               QLPPETTDEPLEKGYS
                       ||  || |||       7/10
β-lactoglobulin     VRTPEVDDEALEKFDK

REALIGNMENT USING SIMILARITY SCORING

Cry3A               QLPPETTDEPLEKGYS
                       ||::||:|||       10/10
β-lactoglobulin     VRTPEVDDEALEKFDK
```

FIGURE 8. An example of a two-step alignment between the Cry3A protein sequence and bovine β-lactoglobulin. The top alignment, produced using an identity matrix, shows a region with a high level of identity (7 of 10 amino acids). Realignment of this region using an evolutionary matrix shows that the entire region can be considered highly conserved.

acids were found to be either identical or conservative replacements, suggesting that the potential immunologic significance of this similarity should be evaluated further.

CONCLUSIONS

Bioinformatics and protein sequence analysis play an important role in assessing the potential allergenicity of new food proteins, particularly for proteins that have not previously been part of the food supply. The construction and use of specialized databases (such as the PAD) will continue to play an important role in ensuring that allergenicity assessments are complete and accurate. Efforts to update these databases will continue, as will the development of standards for database design that allow data interchange between different allergen databases. In the future, both structural and immunologic information (such as epitope sequences) will be integrated into these databases and will be used to improve allergenicity assessments. There is also a need for ongoing characterization and comparison of analysis strategies and data-reporting procedures to ensure that the results of these analyses can be compared between databases and readily understood by others.

ACKNOWLEDGMENTS

This work was partially supported by Cooperative Agreement FD-000431 between the U.S. Food and Drug Administration and the National Center for Food Safety and Technology.

REFERENCES

1. METCALFE, D., J. ASTWOOD, R. TOWNSEND, *et al.* 1996. Assessment of the allergenic potential of food derived from genetically engineered crop plants. Crit. Rev. Food Sci. Nutr. **36S:** S165–S186.
2. FOOD AND DRUG ADMINISTRATION. 1994. Summary: Conference on Scientific Issues Related to Potential Allergenicity in Transgenic Food Crops. FDA. Washington, DC.
3. BAXEVANIS, A. 2001. The molecular biology database collection: an updated compilation of biological database resources. Nucleic Acid Res. **29:** 1–10.
4. WHEELER, D., D. CHURCH, A. LASH, *et al*. 2001. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. **29:** 11–16.
5. BENSON, D., I. KARSCH-MIZRACHI, D. LIPMAN, *et al*. 2000. GenBank. Nucleic Acids Res. **28:** 15–18.
6. BAIROCH, A. & R. APWEILER. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28:** 45–48.
7. BARKER, W., J. GARAVELLI, Z. HOU, *et al*. 2001. Protein information resource: a community resource for expert annotation of protein data. Nucleic Acids Res. **29:** 22–28.
8. GENDEL, S. 1998. Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. Adv. Food Nutr. Res. **42:** 63–92.
9. ALTSCHUL, S., T. MADDEN, A. SCHAFFER, *et al*. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.
10. PEARSON, W. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. **183:** 63–98.
11. ALTSCHUL, S. 1991. Amino acid substitution matricies from an information theoretic perspective. J. Mol. Biol. **219:** 555–565.
12. HENIKOFF, S. & J. HENIKOFF. 1992. Amino acid substitution matricies from protein blocks. Proc. Natl. Acad. Sci. USA **89:** 10915–10919.
13. STANLEY, J., N. KING, A. BURKS, *et al*. 1997. Identification and mutational analysis of the immunodominant IgE binding epitopes of the major peanut allergen Ara h2. Arch. Biochem. Biophys. **342:** 244–253.
14. SHIN, D., C. COMPADRE, S. MALEKI, *et al*. 1998. Biochemical and structural analysis of the IgE binding sites on Ara h1, an abundant and highly allergenic peanut protein. J. Biol. Chem. **273:** 13753–13759.