2mEPSPS PROTEIN

EPITOPE HOMOLOGY AND

N-GLYCOSYLATION SEARCHES

DATA REQUIREMENT
No applicable guidelines

***IN SILICO* STUDY**

AUTHOR: A. CAPT

TESTING FACILITY:

Bayer CropScience
355, rue Dostoïevski
BP 153
06903 Sophia Antipolis Cedex
France

SPONSOR:

Bayer AG
Bayer CropScience
Alfred Nobel Str. 50
40789 Monheim
Germany

STUDY COMPLETED ON: NOVEMBER 27, 2008

STATEMENT OF NO DATA CONFIDENTIALITY CLAIM

No claim of confidentiality is made for any information contained in this study on the basis of its falling within the scope of FIFRA § 10 (d) (1) (A), (B) or (C).

Company Name:

Company Agent:

Title:

Signature:

Date: _____

These data are the property of Bayer CropScience, and as such, are considered to be confidential for all purposes other than compliance with FIFRA § 10. Submission of these data in compliance with FIFRA does not constitute a waiver of any right to confidentiality which may exist under any other statute or in any other country.

STATEMENT CONCERNING GOOD LABORATORY PRACTICE

The *in silico* search was conducted in the spirit of Good Laboratory Practices (GLP). The final report was not audited by the Quality Assurance Unit.

Company: Bayer CropScience

Submitter:

Author:

Date: November 27, 2008

A. CAPT

Sponsor Representative:

Date: NOVEMBER 27, 2008

C. HERQUET-GUICHENEY

FLAGGING STATEMENTS

This page is reserved for flagging statements as may be required by US EPA.

SIGNATURE

I, the undersigned, hereby declare that the work was performed under my supervision according to the procedures described and that this report provides a correct and faithful record of the results obtained.

There were no circumstances which affected the quality and integrity of the data.

Author:

Date: November 27, 2008

A. CAPT

TABLE OF CONTENTS

SUMMARY

This *in silico* study evaluated the potential amino acid similarity of the double mutant maize 5-enol pyruvylshikimate-3-phosphate synthase (2mEPSPS) protein with all theorical linear epitopes found on known allergens. The purpose was to identify any short sequences of amino acids that might represent an isolated shared allergenic epitope that may not be detected during the overall homology analysis.

In addition, this study considered the potential N-glycosylation sites by searching their known consensus sequence as they may be found in allergenic proteins.

The epitope homology study was carried out by comparing the amino acid sequence of the 2mEPSPS protein, subdivided into 8 amino acid blocks, with all known allergens present in the public allergen database AllergenOnline (www.allergenonline.com; release 8.0, 1313 sequences). The algorithm used was FindPatterns (GCG package) and the criterion indicating potential allergenicity was a 100 % identity on a window of 8 amino acids with an allergenic protein.

No significant similarities were found between the 8 linearly contiguous amino acid blocks, which compose the 2mEPSPS protein, and known allergens from the AllergenOnline database. Therefore, the 2mEPSPS protein is not expected to cross-react with known allergens.

In addition, two potential N-glycosylation sites were identified on the 2mEPSPS amino acid sequence by using the N - X~(P) - [S,T] consensus sequence.

INTRODUCTION

One of the key endpoints evaluated for assessing the safety of a novel protein is the amino acid (aa) sequence similarity with known allergens. For this reason, this comparison is performed as part of current food safety evaluation strategies (FAO/WHO, 2001; CAC 2003).

First, an overall amino acid sequence homology search was performed to evaluate the potential amino acid sequence similarity with known toxins and allergens. However, a single short stretch of homology might not have been retained within the overall homology search, e.g. if the overall homology between the compared proteins is very low. Therefore, a more detailed analysis is required.

The objective of this epitope sequence homology search was to identify any short sequences of amino acids representing isolated shared allergenic epitopes that may have not been picked up by the overall homology analysis. This *in silico* approach was conducted in order to identify the potential amino acid sequence similarity of the 2mEPSPS protein with epitopes (8 linearly contiguous amino acids) belonging to known allergens. This report presents the results of this detailed search.

Moreover, although N-glycosylation can be found in various proteins with no allergenic potential, it is also a frequent post-translational modification found on allergenic proteins. Therefore, the search of potential N-glycosylation sites was also performed.

The study time schedule was as follows:

Study initiation date	November 24, 2008
Date of search	November 24, 2008

MATERIAL AND METHODS

1 - AMINO ACID QUERY SEQUENCE

The amino acid query sequence was coded using the one-letter code adopted by the Commission on Biochemical Nomenclature of the IUPAC-IUB (1984) (see Appendix A).

As described in the document number M-234186-01-1 (De Beuckeleer, 2003), the query sequence corresponding to the double mutant maize 5-enol pyruvylshikimate-3-phosphate synthase (2mEPSPS) protein is as follows:

```
MAGAEEIVLQP IKEISGTVKL PGSKSLSNRI LLLAALSEGT TVVDNLLNSE DVHYMLGALR        61
 TLGLSVEADK AAKRAVVVGC GGKFPVEDAK EEVQLFLGNA GIAMRSLTAA VTAAGGNATY       121
 VLDGVPRMRE RPIGDLVVGL KQLGADVDCF LGTDCPPVRV NGIGGLPGGK VKLSGSISSQ       181
 YLSALLMAAP LALGDVEIEI IDKLISIPYV EMTLRLMERF GVKAEHSDSW DRFYIKGGQK       241
 YKSPKNAYVE GDASSASYFL AGAAITGGTV TVEGCGTTSL QGDVKFAEVL EMMGAKVTWT       301
 ETSVTVTGPP REPFGRKHLK AIDVNMNKMP DVAMTLAVVA LFADGPTAIR DVASWRVKET       361
 ERMVAIRTEL TKLGASVEEG PDYCIITPPE KLNVTAIDTY DDHRMAMAFS LAACAEVPVT       421
 IRDPGCTRKT FPDYFDVLST FVKN                                              445
```

2 - EPITOPE HOMOLOGY SEARCH METHOD

2.1 Search algorithm

An algorithm was developed, using the FindPatterns program from Genetic Computer Group (GCG), to search the potential common epitopes between the query protein and allergenic sequences present in the allergen database.

The FindPatterns program is a sequence comparison algorithm that is used to identify sequences that contain short patterns. FindPatterns can recognize the patterns ambiguously and allow mismatches (but not gaps).

In this study, the exact process used to search potential epitopes in the query sequence was as follows:

1. The complete amino acid sequence of the 2mEPSPS protein was subdivided into overlapping blocks of 8 amino acids (8 amino acid peptides) starting from position 1 with a sliding window of 1 amino acid:

   e.g. Query LYSSDWLIYKTT
   Block1 LYSSDWLIYK
   Block2 YSSDWLIYKT
   Block3 SSDWLIYKTT etc.

2. Each block was compared with entire sequences of the Allergen Database by using the FindPatterns algorithm. For each block, if a 100 % match was found between the block and an allergen sequence, the block size was extended by 1 amino acid. This new block was then compared with the allergen sequences.

The same process continued until no exact match was found between the block and the each allergen sequence.

3. At the end of the process, if an exact match was found between a block (size ≥ 8 amino acids) and an allergen sequence, the result was recorded. This block represents a potential epitope.

2.2 Allergen database

The allergen database used was AllergenOnline, version 8.0, 2008. AllergenOnline allergen database (www.allergenonline.com) is a free, publicly available, archived resource list of known and putative allergens and sequences. The database is updated annually by searching NCBI and IUIS annotated sequences and by evaluating the candidate entries for evidence of protein allergenicity (i.e. IgE binding test) and food allergy (e.g. clinical test). A peer review panel of food allergy experts from academia is in charge of this curation. The exact list of experts is reported on the website. They identify whether proteins are allergens, putative allergens or unlikely to be allergenic based on predefined criteria, which are described on the website. Version 8.0 of the database includes 1 313 unique sequences that are clustered into 484 allergen groups based on species (n=230) and sequence identities (Thomas *et al*., 2008).

2.3 Matching criterion

Although the distinction between allergenic and non-allergenic epitopes remains unclear, a search for any 6 or more contiguous amino acids that are identical to any segment of any known allergen (food, inhalant or contact allergen) has been recommended in the 2001 FAO/WHO report (FAO/WHO, 2001).

However, the use of short amino acid matches must be viewed with caution as it is not clear whether searching for matches of less than 8 amino acids is scientifically justified. Experimental data showed that large numbers of non-allergens have matched sequences of 6 or 7 with known allergens (Hileman *et al*., 2002; Kleter and Peijnenburg, 2002; Stadler and Stadler, 2003), hence any such false positives cannot be interpreted as indication of an allergenic potential. More recently, the ILSI workshop participants (Thomas *et al*., 2005) agreed that the 6-mer or 7-mer sliding amino acid window searches do not provide any value for the bioinformatics evaluation of novel proteins and that further research is required to define a useful window size for predicting allergenicity. The ILSI approach also supports to follow the *Codex Alimentarius ad hoc* expert panel recommendations, *i.e.* "the size of the contiguous amino acid search should be based on a scientifically justified rationale" (CAC, 2003, page 18). Therefore, although this may also overestimate the number of potentially allergenic proteins (Thomas *et al*., 2008), only the matches of 8 contiguous and identical amino acids and above may have some biological relevance.

Thus, the matching criterion selected for identification of significant similarity to an allergen was a 100 % identity over a linear contiguous 8 amino acid segment.

2.4 Expression of results

If matches were identified, the results of each individual pattern (matching sequence of ≥ 8 amino acids with 100% identity) were reported as follows:

1. Description of the hit protein,
2. Accession number of the hit protein,
3. Position of the pattern on the query sequence,
4. Position of the pattern on the hit protein sequence,
5. Length of the pattern,
6. Sequence of the pattern, with its flanking sequence (amino acids at each border).

If no matches were identified between the query sequence blocks and the entire allergen database sequences, a computer message 'No hit found' appeared on the computer screen.

3 - N-GLYCOSYLATION SEARCH METHOD

This *in silico* approach provides the possibility to search for potential N-glycosylation sites in any protein sequence.

The best studied mode of glycosylation is the formation of an N-glycosidic linkage to Asparagin in the polypeptide chain. The necessary (but not sufficient) criterion for protein N-glycosylation is the presence of the sequence N-X~(P)-S/T, where N = Asparagin, X~(P) = any amino acid except Proline (P), S = Serin and T = Threonin, in the query sequence. Although rare, the sequence motif N-X-C can also be an acceptor site (where N = Asparagin, X = any amino acid and C = Cystein).

Therefore, the consensus sequences searched were of the following type:
N - X~(P) - [S,T] or N - X - C.

Expression of results:

Each potential N-glycosylation site was reported with its amino acid sequence, its flanking sequence (amino acids at each border) and its location in the query sequence.

If there were no matches with the consensus sequences, a computer message 'No hit found' appeared on the computer screen.

4 - DATA STORAGE

All raw data, supporting documents as well as final report are maintained in the document archive room. All of the above will be archived for at least 10 years in the designated areas at:

Bayer CropScience
355, rue Dostoïevski
BP 153
06903 Sophia Antipolis Cedex
France

RESULTS

No significant similarities were found between the 8 linearly contiguous amino acid blocks, which compose the 2mEPSPS protein, and known allergens from the AllergenOnline database.

In addition, the identification of potential N-glycosylation sites were carried out by searching their known consensus sequences. Two potential N-glycosylation sites were identified on the 2mEPSPS amino acid sequence by using the N - X~(P) - [S,T] consensus sequence (Table 1).

CONCLUSION

The lack of any significant amino acid sequence homology leads to the conclusion that the 2mEPSPS protein is not expected to cross-react with known an allergens.

This corroborates the results from the overall homology search, which are described by Capt (2008).

Two potential N-glycosylation sites have been identified in this *in silico* study. Only an experimental approach could confirm an effective N-glycosylation modification on the protein.

REFERENCES

DART Numbers	Author(s), year, title, source, edition, pages, web page
	Capt, A. 2008. 2mEPSPS protein - Overall amino acid sequence homology search with known toxin and allergens. *In Silico* Study. Bayer Cropscience. 526 pages.
M-276947-01-1	*Codex Alimentarius* commission (CAC). 2003. *Codex* principles and guidelines on food derived from biotechnology. *Codex Alimentarius* Commission 2003. CAC/GL44-2003 and CAC/GL45-2003.
M-234186-01-1	De Beuckeleer; M. 2003. Description of the amino acid sequence of the double mutated maize 5-enol pyruvylshikimate-3-phosphate synthase (2mEPSPS). Study number EPSPSaas/01. Bayer CropScience. June 23, 2003. 5 pages.
M-215805-01-1	FAO/WHO. 2001. Report of a joint Food and Agriculture Organization of the United Nations (FAO)/World Health Organization (WHO) expert consultation on foods derived from biotechnology. Rome, Italy, 22-25 January, 2001. 35 pages.
M-218356-01-1	Hileman R.E., Silvanovich A., Goodman R.E., Rice E.A., Holleschak G., Astwood J.D., Hefle S.L. 2002. Bioinformatic methods for allergenicicty assessment using a comprehensive allergen database. *Int. Arch. Allergy Immunol.* 128: 280-291.
M-264607-01-1	IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). 1984. Nomenclature and symbolism for amino acids and peptides. Recommendations 1983. *Eur. J. Biochem*. 138:9-37. http://www.chem.qmw.ac.uk/iupac/AminoAcid/
M-276626-01-1	Kleter G.A., Peijnenber A.C.M. 2002. Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE-binding linear epitopes of allergens. *Biomed. Central.* 2:1-11.
M-256322-01-1	Stadler, M.B., Stadler, B.M. 2003. Allergenicity prediction by protein sequence. *FASEB J*. 17: 1141-1143.
M-263625-01-1	Thomas, K., Bannon, G., Hefle, S., Herouet, C., Holsapple, M., Ladics, G., MacIntosh, S., Privalle, L., 2005. *In silico* methods for evaluating human allergenicity to novel proteins: International bioinformatics workshop meeting report, 23-24 February 2005. *Tox. Sc.* 88:307-310.
M-307889-01-1	Thomas, K., Herouet-Guicheney, C., Ladics, G., McClain, S., Macintosh, S., Privalle, L., and Woolhiser, M. 2008. Current and future methods for

evaluating the allergenic potential of proteins: International workshop report 23-25 October 2007. *Food Chem. Toxicol.* 46:3219-3225.

ABBREVIATIONS AND ACRONYMS

%	Percent(age)
2mEPSPS	Double mutant maize 5-enol pyruvylshikimate-3-phosphate synthase
λ	Lambda
aa	Amino acid(s)
DNA	Desoxyribo nucleic acid
cDNA	Complementary DNA
E or E-Value.....	Expect(ed) value
EPA	Environmental Protection Agency
EPSPS	5-enol pyruvylshikimate-3-phosphate synthase
FAO.....	Food and Agriculture Organization
GCG	Genetic Computer Group
IgE.....	Immunoglobulin E
IUPAC.....	International Union of Pure and Applied Chemistry
IUB.....	International Union of Biochemistry
JCBN.....	Joint Commission on Biochemical Nomenclature
K.....	Statistical parameter for calculating BLAST scores
L	Gap opening penalty
Ln	Natural logarithm
Log	Common logarithm
N°	Number
NCBI.....	National Center for Biotechnology Information
S	Score or raw score
S'.....	Normalized score or bit score
US or USA	United States of America
WHO	World Health Organization

GLOSSARY

Alignment

The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

Bit score

The value S' is derived from the raw alignment score S in which the statistical properties of the scoring system used have been taken into account. Since bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches.

If S is the raw score for a local alignment, the normalized score S' (in bits) is calculated by the formula $S'=(\lambda S-\ln K)/\ln 2$. A normalized "bit score", S' with E value = E, is statistically significant if it exceeds log N/E where N is the size of the search space.

Description

General descriptive information about the protein sequence stored in the database. This information is generally sufficient to identify precisely the protein.

Epitope

Immunoreactive sequence of an allergen.

E-value

Expect(ed) value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E-value, the more significant the score.

The E-value corresponding to a given bit score is: $E=Kmne^{-\lambda S}$.

Gap

A space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment.

Homology

Similarity attributed to descent from a common ancestor.

Identity

The extent to which two (nucleotide or amino acid) sequences are invariant. The identity percentage is the minimal percentage of identical residues between query sequence and sequence database hit.

Local Alignment

The alignment of some portion of two protein (or nucleic acid) sequences.

Motif

A short conserved region in a protein sequence. Motifs are frequently highly conserved parts of proteins.

Optimal Alignment

An alignment of two sequences with the highest possible score.

Positive

The Positive Percentage is the number of positive residues, dependent on the substitution matrix used.

Query

The input sequence with which all of the entries in a database are to be compared.

Similarity

The extent to which nucleotide or protein sequences are related. The extent of similarity between two sequences can be based on percent sequence identity and/or conservation.

TABLE

TABLE 1 – POTENTIAL N-GLYCOSYLATION SITES

CONSENSUS : N - X~(P) - [S,T] TOTAL FINDS : 2

POSITION ON QUERY	POTENTIAL SITE
118	TAAGG **NAT** YVLDG
394	PPEKL **NVT** AIDTY

APPENDIX

APPENDIX A - AMINO ACID CODES

One-letter codes	Three-letter codes	Amino-acid names
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamic acid or Glutamine
X	Xaa	Any amino acid

FINAL REPORT AMENDMENT

There is no final report amendment at this time.

This page has been left bank intentionally.