A Method to Determine an E-score Threshold to Identify Potential Cross Reactive Allergens

Andre Silvanovich Ph. D.
Gary Bannon Ph. D.
Scott McClain Ph. D.

Monsanto Company
Regulatory Product Characterization Center
St. Louis, Missouri, USA
63167

The Sliding Window FASTA Search

- A means to identify potential cross reacting allergens
- FAO/WHO recommend a 35% or greater identity in 80 amino acids or greater threshold for a sliding window search
- By design, the FASTA algorithm reports and assesses the significance of alignments on the basis of shared global similarity
- The E-score is a composite measure of similarity between proteins that are aligned by FASTA is
- Pearson, the developer of FASTA states:
  - "The E() is the first number you should look at when deciding whether to analyze further a high-ranking sequence"

Objectives

- To propose a method to determine an E-score threshold
  - Model allergens and non-allergens
  - Identify a statistically relevant threshold
  - Apply a test case for a family of known allergens

E-score is an assessment of structural similarity

- What is an E-score?
  - The probability that an alignment was the product of chance
  - An alignment yielding an E-score of 1 is comparable to searching the database with a random sequence, i.e, no significance attached to this value
  - An E-score of 0.01 indicates that there is a 1 in 100 probability that the alignment is the product of chance
- The E-score reflects not only amino acid identity, it includes amino acid similarity, the length of the alignment and the size of the database being searched

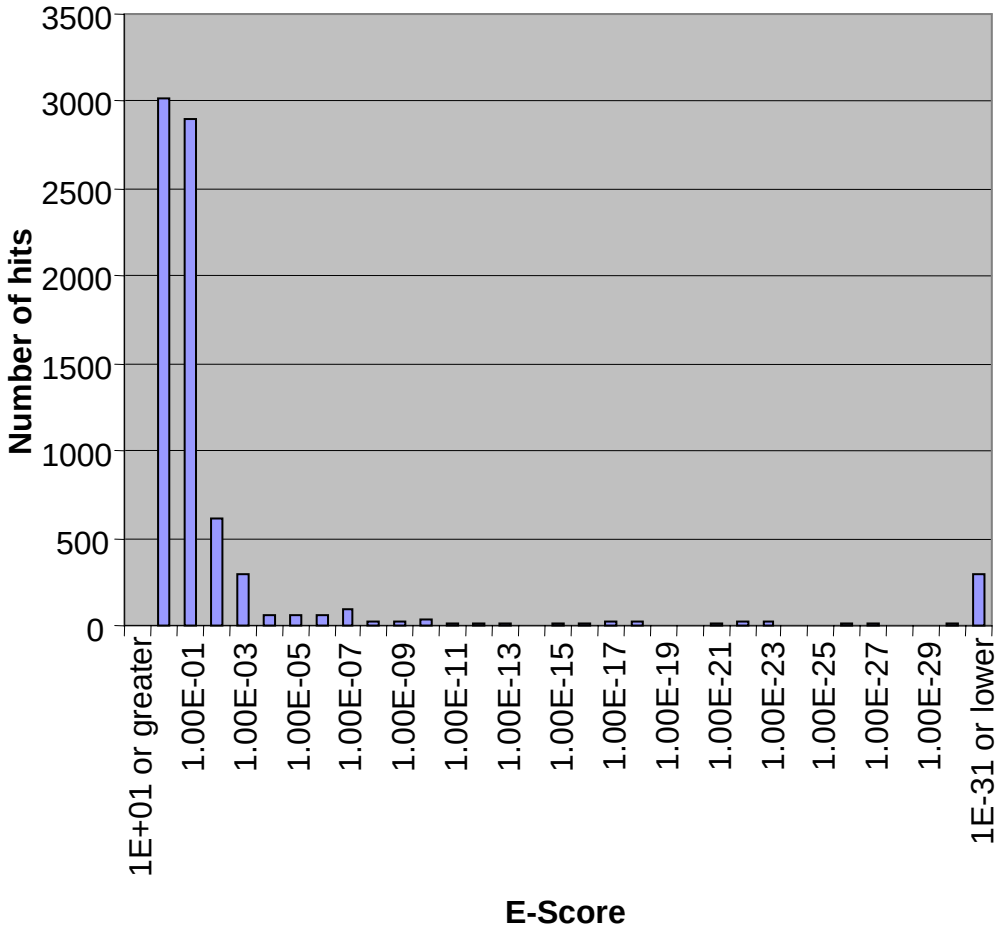
Random sequences and E-scores

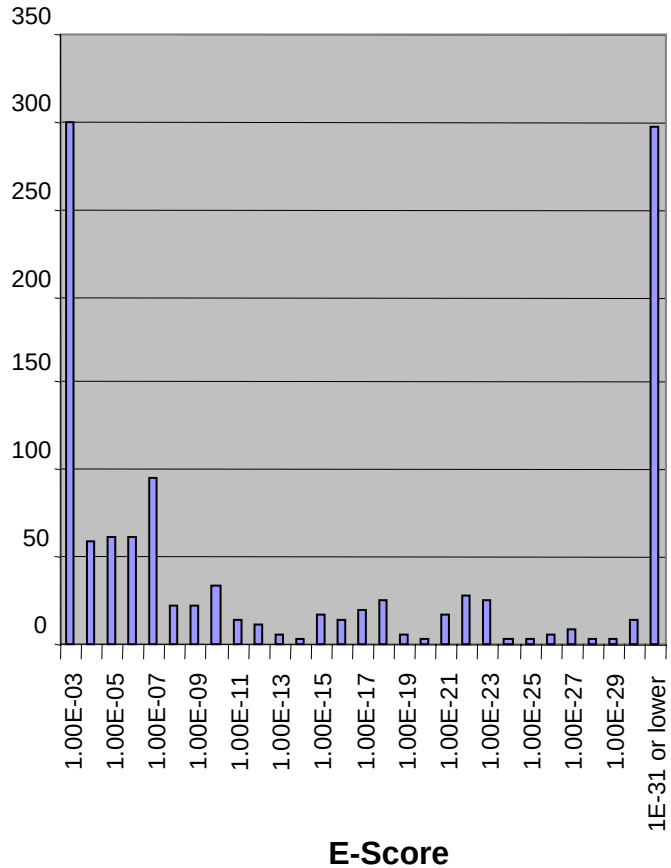
		Likely non-allergen	Potential Allergen	Allergen	Allergen
Unshuffled Query E-score		1.0E+00	3.40E-07	4.10E-75	3.40E-154
Shuffled sequence E-scores	> 10	3	0	2	8
	>= 1.0 <10.0	522	489	525	543
	>= 0.1 <1.0	407	423	411	384
	>= 0.01 <0.1	63	80	59	61
	>= 0.001 <0.01	4	5	3	9
	>= 0.0001 <0.001	1	3	0	1
	>=1E-05 <0.0001	0	0	0	0
	>=1E-06 <1E-05	0	0	0	0
	>=1E-07 <1E-06	0	0	0	0
Number of unique database sequences yielding a best match		403	421	453	425
Number of 35% over 80 aa matches with shuffled queries		1	3	1	6

Establishing an E-Score Threshold

- What is the relationship between chance and an alignment that may reflect potential cross reactivity?
- A query dataset of 7695 protein sequences derived from corn was evaluated
- The query dataset contains known allergens
  - However, the overwhelming majority of the query dataset should not be allergens
  - Corn has a low incidence of human allergy
- FASTA analysis was performed using each of the 7695 query sequences and the E-score distribution for the top alignment for each query was plotted
- Sliding window searches were performed with the 7695 query sequences
  - The FARRP allergen database and FASTA version 3.4t26 were used to conduct the analyses

The E-score Distribution for 7695 Query Sequences





Threshold Modeling Estimates

E-score threshold	Rank out of 7695 alignments	False negative rate	Potential false positive rate
0.99	4683	0 %	99.29 %
0.099	1786	0 %	98.15 %
0.0099	1176	0 %	97.19 %
0.00099	876	0 %	96.23 %
9.90E-05	818	0 %	95.97 %
9.90E-06	757	0 %	95.64 %
9.90E-07	696	0 %	95.26 %
4.20E-07	660	0 %	95.00 %
5.00E-11	519	0 %	93.64 %
2.4E-31	295	0 %	88.81 %
1.5E-55	165	58 %	80.00 %

$$\frac{4683 - 33}{4683} \times 100\%$$

33 of the 7695 queries displayed 100% identity to known allergens

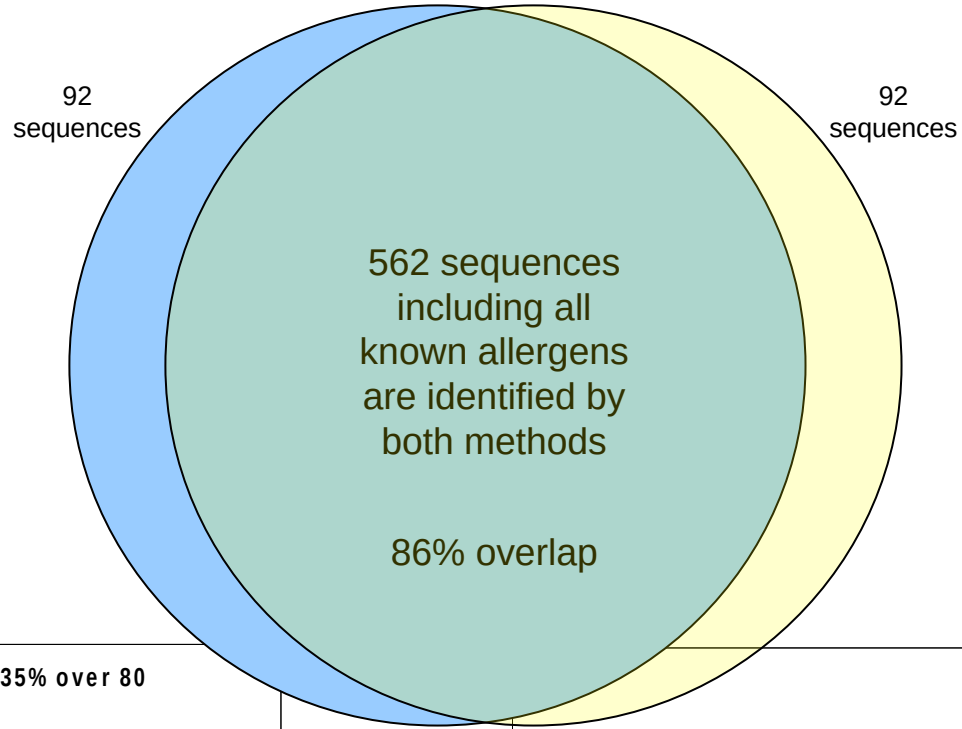
Threshold Modeling Estimates Derived from the 35% over 80 aa Search Results

E-score threshold	Rank out of 7695 alignments	False negative rate	Potential false positive rate
0.0043[€]	1073	0 %	96.92 %
3.90E-07 [¶]	654	0 %	94.95 %

[€] This value is based upon the 35% over 80 amino acid window threshold using a sliding window yielding 1073 matches with the query dataset.

[¶] This value is based upon the 35% over 80 amino acid window threshold using a full length query yielding 654 matches with the query dataset.

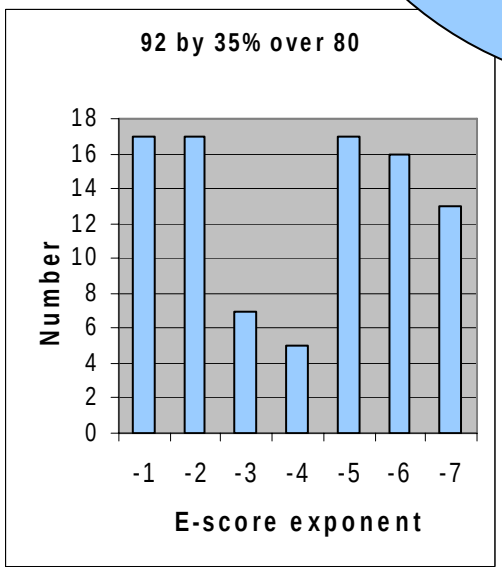
The relationship between a 35% over 80 amino acid window threshold and an E-score based threshold

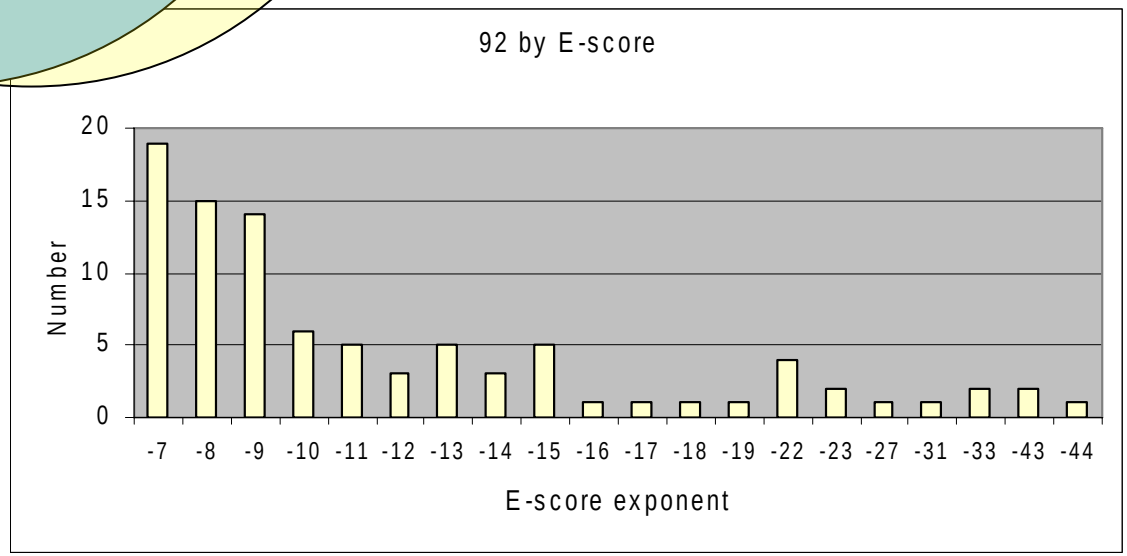


Identified only by 3.9E-7 or lower

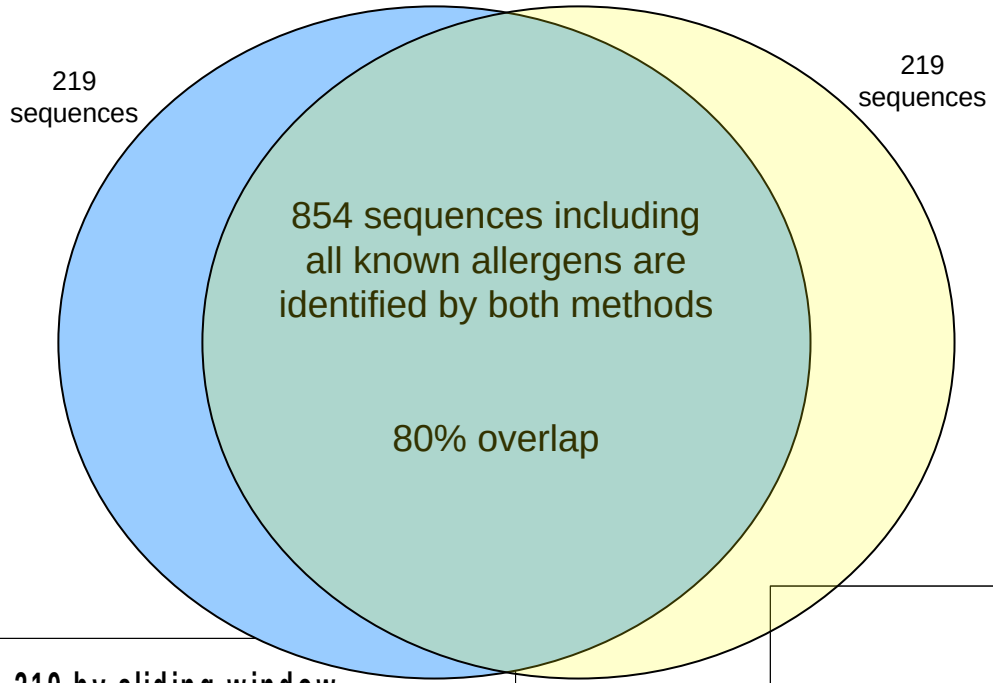
Identified only by 35% / 80 aa

Identified by both methods





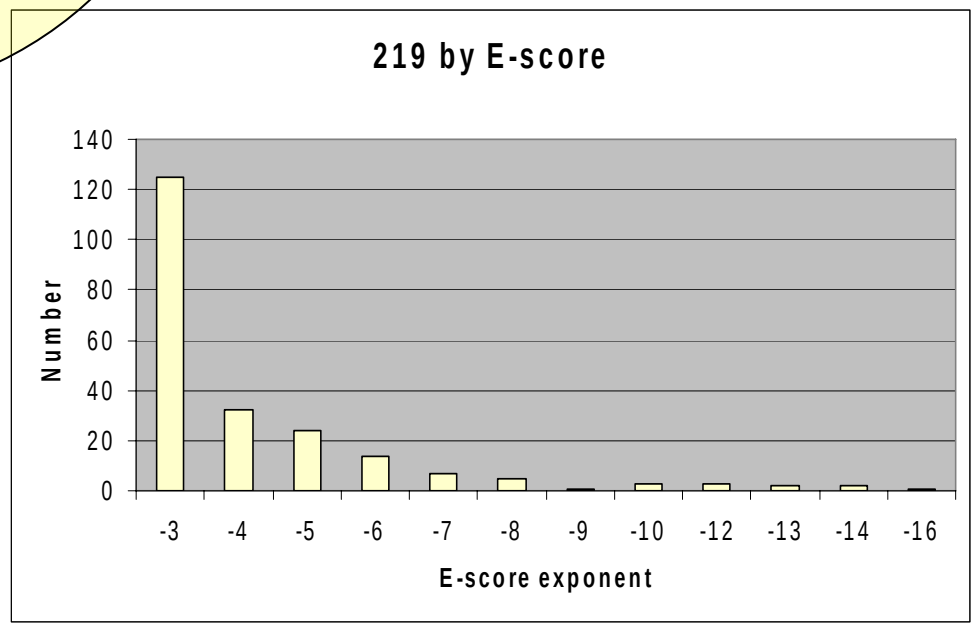
The relationship between a sliding 35% over 80 amino acid window threshold and an E-score based threshold



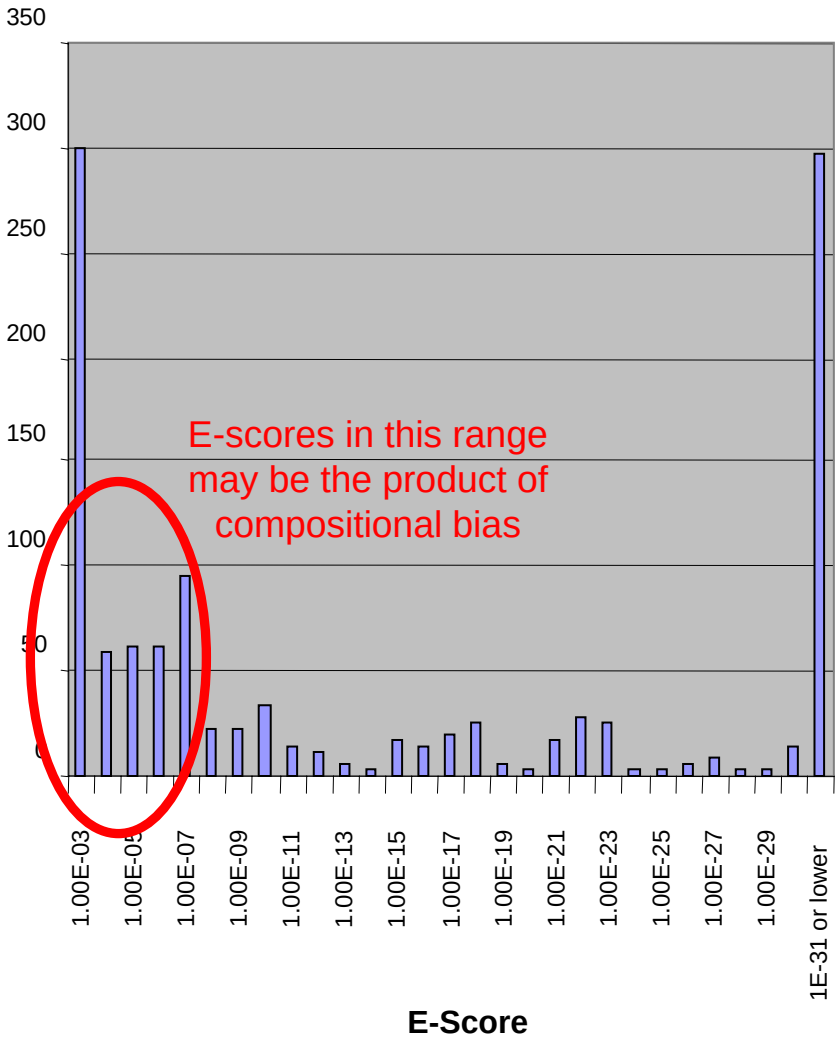
Identified only by 4.3E-3 or lower

Identified only by a sliding 35%
/ 80 aa window

Identified by both methods



The Impact of Amino Acid Composition Bias on E-scores



Protein amino acid composition may result in E-scores that are significant but that do not represent a meaningful alignment

Searches of allergen databases are subject to compositional bias created by glutamine-rich proteins that align with gliadins and glutenins

Composition bias can be uncovered using shuffled query sequences

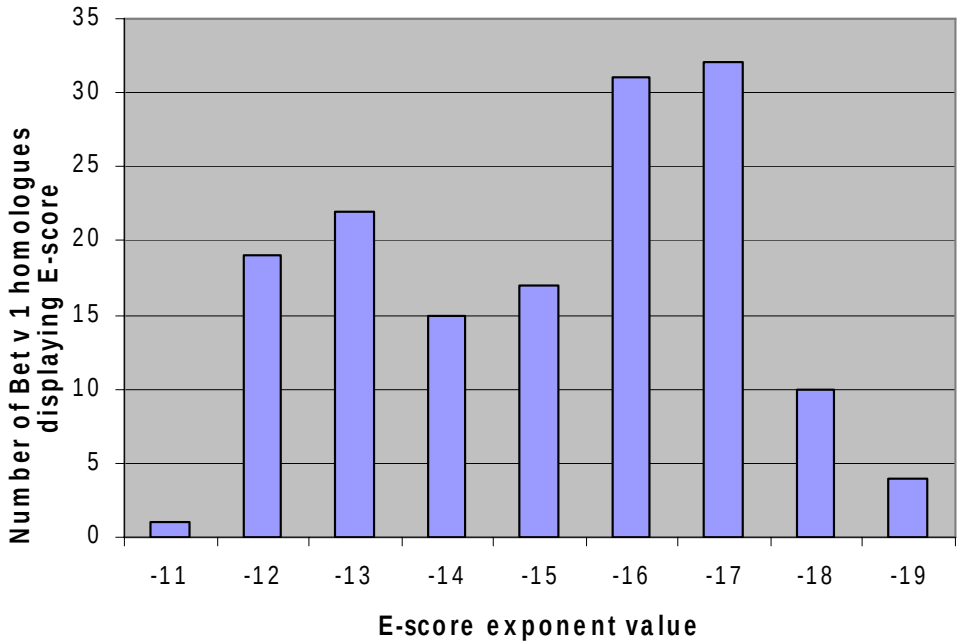
Assessing E-score and Amino Acid Composition Bias Using Shuffled Sequences

		Likely non-allergen	Potential Allergen	Allergen	Allergen	Glutamine-rich
Unshuffled Query E-score		1.0E+00	3.40E-07	4.10E-75	3.40E-154	3.70E-06
Shuffled sequence E-scores	> 10	3	0	2	8	1
	>= 1.0 <10.0	522	489	525	543	71
	>= 0.1 <1.0	407	423	411	384	351
	>= 0.01 <0.1	63	80	59	61	339
	>= 0.001 <0.01	4	5	3	9	162
	>= 0.0001 <0.001	1	3	0	1	57
	>=1E-05 <0.0001	0	0	0	0	16
	>=1E-06 <1E-05	0	0	0	0	2
	>=1E-07 <1E-06	0	0	0	0	1
Number of unique database sequences yielding a best match		403	421	453	425	114[Δ]
Number of 35% over 80 aa matches with shuffled queries		1	3	1	6	7

[Δ] Of the 114 unique sequences identified by the shuffled queries, 26 of the unique proteins were described as being gliadins and glutenins. These 26 gliadins and glutenins accounted for 844 of the 1000 top alignments for the shuffled query.

Test Case: What is the E-score Threshold for Identifying Bet V 1 Homologues?

- 151 Bet v 1 homologues of 120 amino acids or greater were extracted from the FARRP allergen database
- Each homologue was used as a query for a FASTA search of the allergen database
- The lowest E-score between the query and a Bet V 1 homologue of equal or greater length was tabulated
- The least significant E-score obtained was 5.00E-11; well below the observed 4.2E-7 level where a 95% false positive rate was identified



Conclusions

- A robust modeling procedure to determine an E-score threshold is proposed
- Using a threshold that is comparable to the 35% identity in 80 aa criteria yields an E-score of 4.2E-7
- We know its conservative based on
  - Statistics; 95% false positive, 0% false negative
  - The threshold is 100% effective in identifying know corn allergens and full length Bet V 1 homologues
  - The threshold is also of sufficient selectivity that it will result in fewer false positive sequence identifications due to composition biased alignments