Utility of "Sliding Window" FASTA in Predicting Cross-Reactivity with Allergenic Proteins

Bob Cressman

Pioneer Crop Genetics

The issue…

- FAO/WHO 2001 –

"Step 2: *prepare a complete set of 80-amino acid length sequences derived from the expressed protein…"*

"Cross-reactivity between the expressed protein and a known allergen (as can be found in the protein databases) has to be considered when there is:

1) more than 35 % identity in the amino acid sequence of the expressed protein (i.e. without the leader sequence, if any), using a window of 80 amino acids and a suitable gap penalty"

```
 1  MSPQTETKAS VGFKAGVKDY KLTYYTPEYE TLDTDILAAF RVSPQPGVPP  50
51  EEAGAAVAAE SSTGTWTTVW TDGLTNLDRY KGRCYHIEPV AGEENQYICY 100
```

Break sequence into all possible 80 residue fragments

```
1  MSPQTETKAS…TDGLTNLDRY  80
                   +
2  SPQTETKAS…TDGLTNLDRYK  81
                   +
3  PQTETKAS…TDGLTNLDRYKG  82
                   +
4  QTETKAS…TDGLTNLDRYKGR  83
                   ⋮
20 YKLTYTPEYE…AGEENQYIC  99
                   +
21 KLTYYTPEYE…AGEENQYICY 100
```

Number of peptides = L-79; where L is length of protein

For a 500 aa protein sequence, there are 421 different FASTA searches to perform

Align each 80 residue amino acid sequence against allergen dataset using FASTA

Review, sort alignments for > 35% identity over = 80 residues

October 24th, 2007

FASTA Analysis

- Asked to look at prevalence of potential cross reactive ORFs in maize genomic sequences…
  - identified by FGENESH software
  - No known similarities to public proteins
  - 1270 total ORFs (1102 ORFS > 80 residues; amenable to FASTA analysis)

- Used Perl scripts:
  - Break each protein into 80 residue sub-peptides
  - Run FASTA33 on each peptide against FARRP6 database
  - Collate and screen FASTA outputs for >35% over 80 aa
  - Over 1.9 million alignments processed

Results

- Seventy three ORFs were above the FAO/WHO FASTA threshold
- Represented 6.7% of the total
- SWISS-PROT index – 0.4% (Stadler and Stadler, 2003)

FASTA algorithm (Pearson and Lipman, 1988)

- Local alignment between query and database sequences
- Four steps*
  - Finds all sets of matches *k* residues or greater (default *k* for proteins = 2) between query and all database proteins
  - All matches within 16 aa are joined; regions with highest density of matches are identified
  - These regions scored with substitution matrix (default matrix for proteins = BLOSUM50). Highest scoring regions identified, joined using gap creation/extension penalties, and ranked
  - Highest scoring database matches subjected to Smith Waterman local alignment
- More sensitive, slower then BLAST (FASTA is unfiltered)

*Adapted from Mount, David. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, 2001.

FASTA Statistics

- FASTA has built in statistical analysis to evaluate the significance of an alignment
  - Uses alignments to generate statistical distribution of alignment scores (score vs. log of sequence length)
  - z score – related to degree of deviation from the distribution
  - *E* score – reflects probability of observing a greater z score within the distribution
    - Higher number - higher likelihood that alignment due to "random" alignment
    - Lower number – more significant
    - Affected by protein length, database size, FASTA parameters (scoring matrix, gap penalties)
    - Important to be consistent when performing comparisons –same parameters

Conventional FASTA

- FASTA already screens across the entire length of the protein
  - Finds highest scoring region of alignment, assesses degree of significance.

```
  1  MSPQTETKAS VGFKAGVKDY KLTYYTPEYE TLDTDILAAF RVSPQPGVPP  50
 51  EEAGAAVAAE SSTGTWTTVW TDGLTNLDRY KGRCYHIEPV AGEENQYICY 100
```

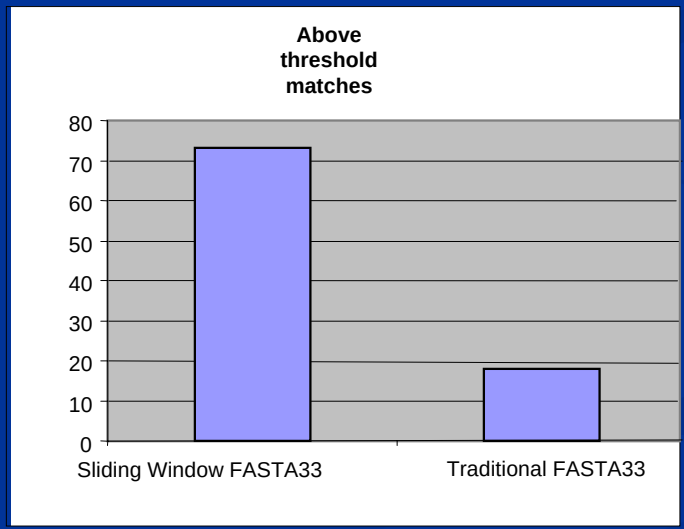
```
>>gi|113560|sp|P22284|MPA91_POAPR Pollen allergen KBG 31  (373 aa)
 initn:  59 init1:  59 opt:  65  Z-score: 91.0  bits: 23.6 E():  3.7
Smith-Waterman score: 65;  31.034% identity (58.621% similar) in 58 aa overlap
(8-65:130-187)

                                      10        20        30
TEST_P                          MSPQTETKASVGFKAGVKDYKLTYYTPEYETLDTDIL
                                       | :|||||:|     :  : :|:|: : :
gi|113 KPAPKVAAYTPAAPAGAAPKATTDEQKLIEKINVGFKAAVAAAAGVPAASKYKTFVATFG
     100       110       120       130       140       150

        40        50        60        70        80        90
TEST_P AAFRVSPQPGVPPEEAGAAVAAESSTGTWTTVWTDGLTNLDRYKGRCYHIEPVAGEENQY
       ||   :   ::  |  |||||: ::: |
gi|113 AASNKAFAEALSTEPKGAAVASSKAVLTSKLDAAYKLAYKSAEGATPEAKYDAYVATLSE
     160       170       180       190       200       210
```

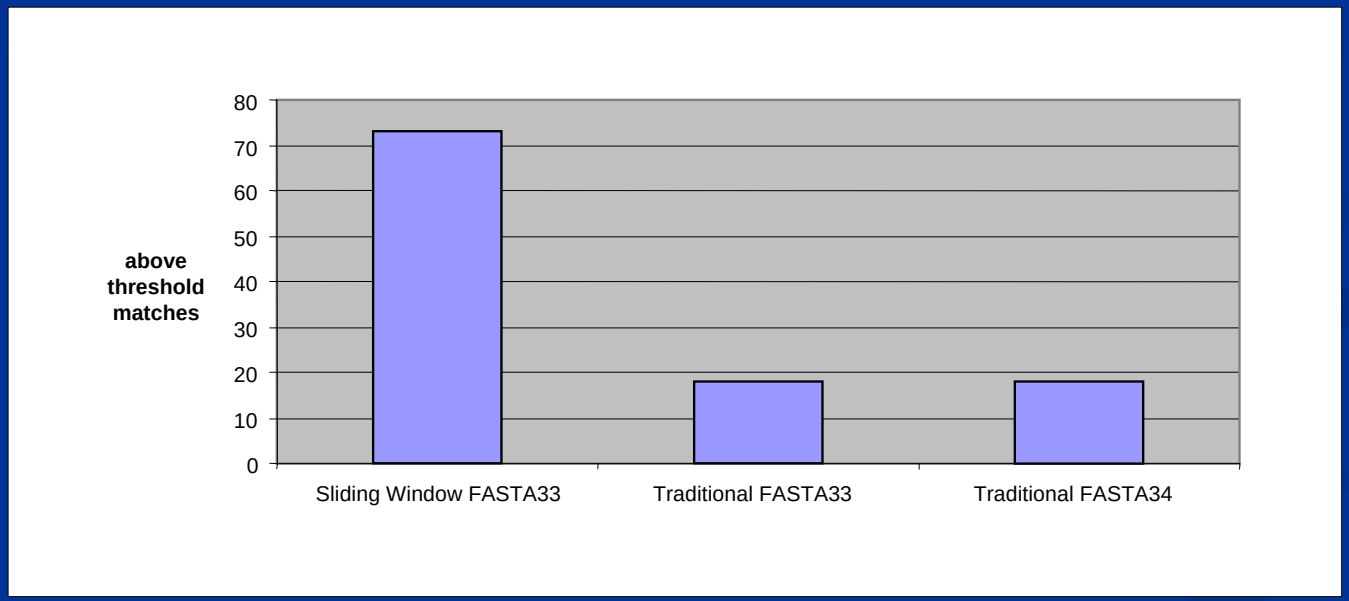
FASTA Comparison

- Decided to repeat analysis of 1102 maize ORFs using "conventional" FASTA
  - Eighteen positives out of 1102 (1.7%)
  - Five fold decrease in "positives"



FASTA33 vs. FASTA34

- "Discovered" newer version of FASTA
  - Change in gap creation penalty from version 33 (-12 to -10)
  - Does this make a difference?
  - Apparently not…



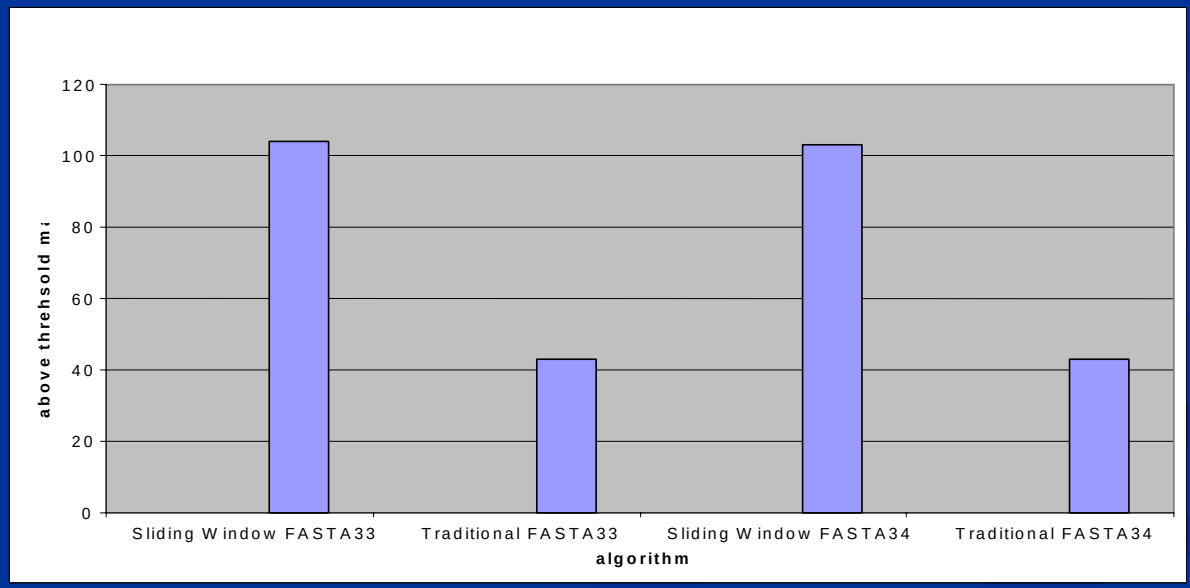
- Decided to use FASTA34 (Version 34t25) for searches

More comparisons…

- Discussion with Andre Silvanovitch, Gary Bannon (Monsanto)
  - Maize ORFs do not represent "real" proteins…
  - Andre provided table of 1000 NCBI proteins (907 proteins longer than 80 aa)
  - Randomly selected

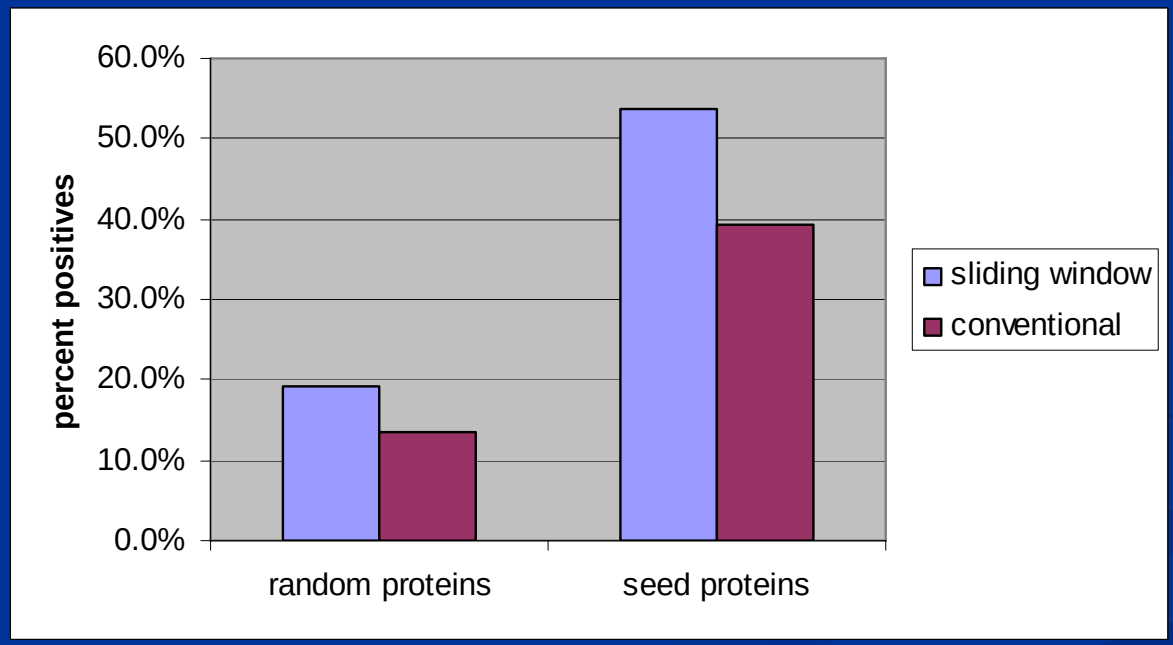
907 Random Proteins - Results

- Forty three positives (4.7%) returned from conventional FASTA searches versus 103 (11.5%) using sliding window search
- *E* scores for conventional FASTA reflect greater significance (lower scores)
- Independent of FASTA version used (33 vs. 34)



Other results

- Also compared
  - 89 random maize proteins
  - 97 seed specific maize proteins
- This work recently published:
  - Ladics et al., 2007, Mol. Nutr. Food Res. 51(8):985-998.



- Sliding window search result

35.000% identity in 80 aa overlap

```
                                10        20        30        40
Q41759            SAASPRG------RRAPVLHRALRRHPRHVRADDIRRHGRRDTVDARHLR
                        |||      :|:|  :|:  |:|     || ||: ||:   :|
gi|116 RCTKLEYDPRCVYDPRGHTGTTNQRSPPGERTRGRQPGDY--DDDRRQPRREE-GGRW--
                  70        80        90       100         110

           50        60        70        80
Q41759 EHAPAPRREGRLRHLPRVSRQDTRRPPRDTQRPRFL
         :||  || : ::  |  |:| ||| :  |:||
gi|116 --GPAGPREREREEDWRQPREDWRRPSH--QQPRKIRPEGREGEQEWGTPGSHVREETSR
           120       130       140         150       160       170
```

- Conventional search result

32.692% identity in 104 aa overlap

```
490       500       510       520       530       540
Q41759 FHDVPTTRVRNHADHPTDHHPTTSAASPRGRRAPVLHRALRRHPRHVRADDIRRHGRRDT
                     |  |  ||:  || |:|:    |:  |:|     || ||: ||:
gi|116 SRCTKLEYDPRCVYDPRGHTGTTNQRSPPGERT----RG--RQPGDY--DDDRRQPRREE
      60        70        80        90             100         110
              550       560       570       580       590       600
Q41759 VDARHLREHAPAPRREGRLRHLPRVSRQDTRRPPRDTQRPRFLRRHAHRSCRHVGEGGAV
         :|     :||  || : ::  |  |:| ||| :  |:|| :| ::::: :: |  |:
gi|116 -GGRW----GPAGPREREREEDWRQPREDWRRPSH--QQPRKIRPEGREGEQEWGTPGSH
                 120       130       140         150       160
```

False negative screen

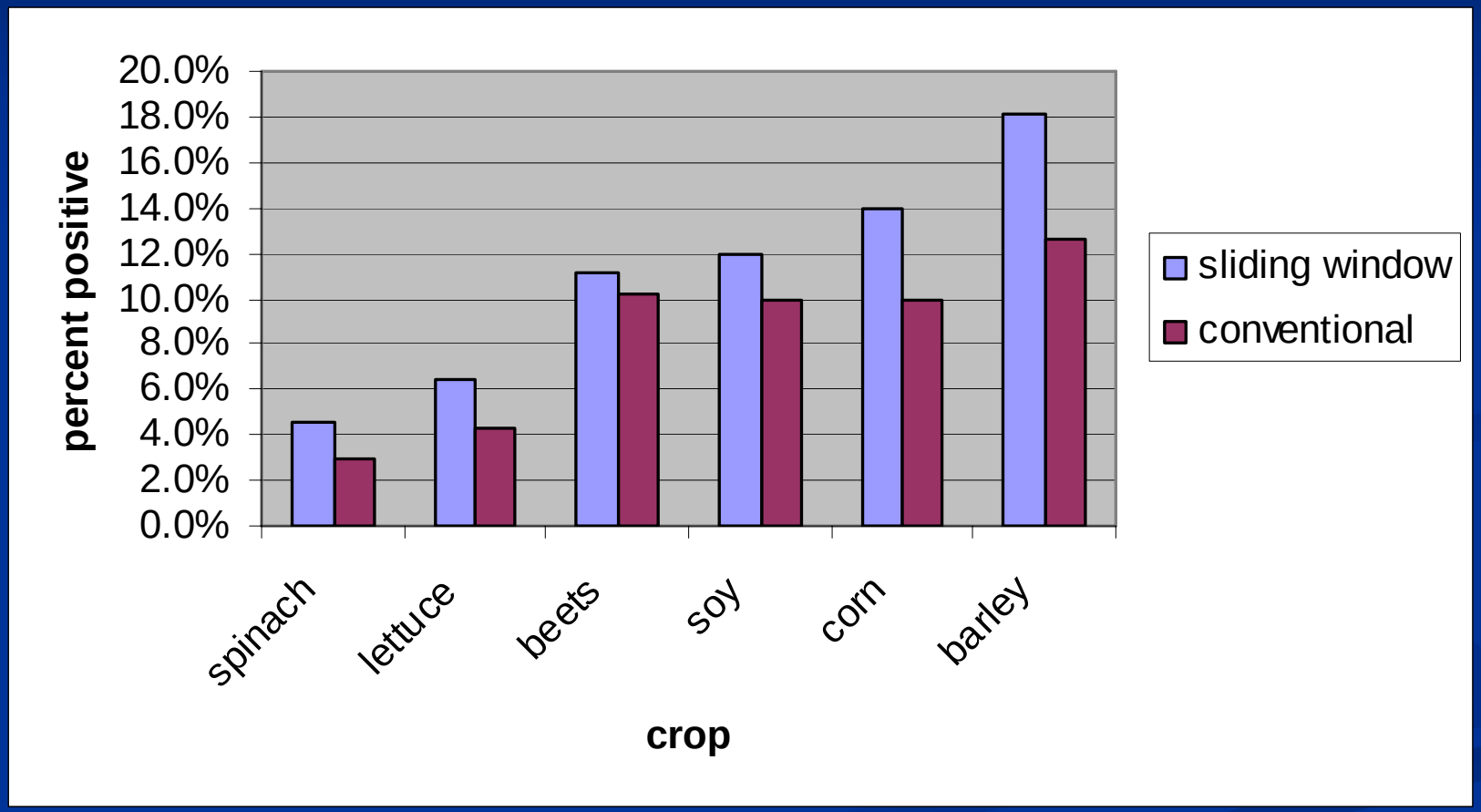
- Is conventional FASTA sensitive enough?
- How do we test this?
  - Examined similarities with Bet v 1 pollen allergen family
    - Believed to form basis for the current FASTA criteria
- All comparisons generated positives using both methods

	Sliding window			Conventional		
comparison	identity (%)	length	E score	identity (%)	length	E score
Bet v 1 vs Dau c 1	40	80	1.90E-11	38.1	155	1.90E-20
Dau c 1 vs Bet v 1	40.7	81	4.30E-10	38.1	155	2.10E-19
Bet v 1 vs Api g 1	45	80	2.10E-12	41.9	155	3.60E-23
Api g 1 vs Bet v 1	45	80	2.30E-12	41.9	155	1.70E-24
Bet v 1 vs Mal d 1	61.3	80	1.60E-19	56	159	2.70E-34
Mal d 1 vs Bet v 1	61.3	80	6.30E-24	56	159	8.50E-33
Bet v 1 vs Pyr c 1	62.5	80	3.00E-19	57.5	160	3.70E-35
Pyr c 1 vs Bet v 1	62.5	80	6.80E-24	57.5	160	5.10E-37
Bet v 1 vs Pru a 1	62.5	80	5.30E-20	59.4	160	1.80E-38
Pru a 1 vs Bet v 1	62.5	80	1.90E-23	59.4	160	2.50E-43

Recent work…

- Decided to compare positives obtained in different crops
  - Lettuce (185 proteins)
  - Spinach (200 proteins)
  - Beets (224 proteins)
  - Barley (200 proteins)
  - Soy (200 proteins)
  - Maize again (200 proteins)
- Removed all proteins < 80 aa, all hypothetical, unnamed, putative, RefSeq accessions, then randomly selected sequences for analysis
- Ran both sliding window and conventional searches…

Results


20.0%
18.0%
16.0%
14.0%
12.0%
10.0%
8.0%
6.0%
4.0%
2.0%
0.0%
percent positive
spinach
lettuce
beets
soy
corn
barley
crop
sliding window
conventional

A closer look

- Out of 1209 proteins (recent work):
  - 3 proteins were identified as positive by conventional FASTA; missed by sliding window search
  - Example:
    - gi|109238657|cystatin Hv-CPI11 [Hordeum vulgare subsp. vulgare]
    - Both techniques identify phytocystatin [Actinidia deliciosa] (GI#40807635) as top scoring alignment:

conventional FASTA				sliding window FASTA		
identity	length	E score		identity	length	E score
35.052	97	1.10E-06		36.765	68	1.60E-08

- Shorter length causes alignment to fall below threshold

A closer look (cont.)

- A total of 35 accessions were classified as positives in the sliding window searches, but not returned using conventional FASTA
- Retrieved corresponding conventional alignments
- Nine of the proteins had no corresponding alignment in the conventional FASTA search
  - Less significant (*E* scores range from 0.5 to 6.9)
  - "True" false positives?
- Two alignments were to the same allergen sequence, but to different regions:

Conventional FASTA allergen region	Sliding window FASTA allergen region
Residues 59-167	Residues 271-336

A closer look (cont.)

- When the remaining 24 were compared to corresponding conventional FASTA alignments:
  - All but one alignment are at or near the threshold of identity (35-40% identity)
  - The majority (70%) possessed more significant alignments (lower *E* scores) using the conventional FASTA searches

A closer look (cont.)

- Some accessions displayed little or no change in significance (*E* score) using either method:
  - Believed to be caused by low complexity sequences
    - Runs of repetitive amino acids
    - Celiac proteins, leader sequences

EPISQQQQQQQQQQILQQILQQQL

ILQRSGSSSSSSSEDD

- Length of low complexity regions stays the same, but represents greater portion of an 80 residue sub-peptide
- FASTA – no low complexity filtering

- Approximately ½ of the sliding window positive accessions fell below the 35% threshold when the conventional FASTA algorithm was employed…
  - Example:

GI number 154816295 - short-chain dehydrogenase/reductase protein sliding window output:

```
>>gi|85701146|sp|P0C0Y5|MTDH_CLAHE Probable NADP-depende  (267 aa)
 initn:  42 init1:  42 opt:  86  Z-score: 133.3  bits: 30.6 E(): 0.016
Smith-Waterman score: 86;  35.366% identity (60.976% similar) in
82 aa overlap (4-79:183-252)
               10        20        30        40              50
154816 ALVGLTRNLAVELAPFGIRVNCVSPFGIATPMTADFIGLE-REVFENMI----NGVAN-L
         : ::|:|| |   |: ||| :||  | | :: ||:  | :::::::|    :|:|: |
gi|857 GCIHMARSLANEWRDFA-RVNSISPGYIDTGLS-DFVPKETQQLWHSMIPMGRDGLAKEL
             190        200       210        220       230
          60        70        80
154816 KGVTHKPDDVAYAALYLASDEAKYV
       ||        ||  :|:||| : |
gi|857 KG--------AY--VYFASDASTYTTGADLLIDGGYTTR
      240                 250       260
```

- Conventional FASTA output:

```
>>gi|85701146|sp|P0C0Y5|MTDH_CLAHE Probable NADP-depende  (267 aa)
 initn: 243 init1:  74 opt: 321  Z-score: 462.6  bits: 93.3 E(): 7.3e-21
Smith-Waterman score: 321; 30.682% identity (63.636% similar) in 264 aa overlap (8-255:18-
265)
                        10        20          30        40
154816           MSIPAKRLEGKVALITGAAS--GIGECCAKLFAAHGAKVIIADVQDQLG-
                       |:|||:::|||::  |:|   |:  |  || | |: ::   |
gi|857 MPGQQATKHESLLDQLSLKGKVVVVTGASGPKGMGIEAARGCAEMGAAVAITYASRAQGA
             10        20        30        40        50        60
         50           60        70        80        90        100
154816 -QAVSE---AIGSSNSMYIHCDITNEEEVKNTIDTAVATYGKLDIMFNNAGI-ADAFKPR
        : |:|   : | : : | :|:: : |  :: :  :|| :|::| :: |||  ||:
gi|857 EENVKELEKTYGIKAKAY-KCQVDSYESCEKLVKDVVADFGQIDAFIANAGATADS---G
              70         80        90       100       110
            110       120       130       140       150         160
154816 IMDNEKKDIERVLGVNVIGTFLCMKHAARVMVPQKSGSIITTSSLTSHLGGMASH--AYS
       |:|:  :  ::|: |:: ||| | | ::: :  : :||:: |:|:::|:::: ::  :|:
gi|857 ILDGSVEAWNHVVQVDLNGTFHCAKAVGHHFKERGTGSLVITASMSGHIANFPQEQTSYN
        120       130       140       150       160       170
              170       180       190       200       210
154816 CSKHALVGLTRNLAVELAPFGIRVNCVSPFGIATPMTADFIGLE-REVFENMI----NGV
        :| : : ::|:|| |   |: ||| :||  | | :: ||:  | :::::::|    :|:
gi|857 VAKAGCIHMARSLANEWRDFA-RVNSISPGYIDTGLS-DFVPKETQQLWHSMIPMGRDGL
        180       190       200       210       220       230
          220       230       240       250       260       270
154816 AN-LKGVTHKPDDVAYAALYLASDEAKYVTAQNMLVDGGLSYCNNSFNMFKYPEEDT
       |: |||        ||  :|:||| : |:|: ::|:||| :
gi|857 AKELKG--------AY--VYFASDASTYTTGADLLIDGGYTTR
          240                 250       260
```

- Is this a significant alignment…?

- Sliding Window FASTA output:

```
>>gi|128388|sp|P19656|NLTP_MAIZE Nonspecific lipid-trans  (120 aa)
 initn: 468 init1: 468 opt: 468  Z-score: 567.1  bits: 109.7 E(): 1.1e-26
Smith-Waterman score: 468;  100.000% identity (100.000% similar) in 75 aa overlap
 (1-75:1-75)
               10        20        30        40        50        60
test.0 MARTQQLAVVATAVVALVLLAAATSEAAISCGQVASAIAPCISYARGQGSGPSAGCCSGV
       ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gi|128 MARTQQLAVVATAVVALVLLAAATSEAAISCGQVASAIAPCISYARGQGSGPSAGCCSGV
               10        20        30        40        50        60
               70        80
test.0 RSLNNAARTTADRRALLERG
       ||||||||||||||||||||
gi|128 RSLNNAARTTADRRAACNCLKNAAAGVSGLNAGNAASIPSKCGVSIPYTISTSTDCSRVN
               70        80        90       100       110       120
```

- How about this …?

Conclusion

- Using conventional FASTA to assess potential cross-reactivity:
  - Reduces the number of false positive results
  - Retains sensitivity near current threshold values (35-40%, Bet v 1)
  - Mitigates effect of low-complexity sequence regions
  - Produces more significant (*E*) alignments
- Still only one part of "weight of evidence" approach…
- Important to review alignments on a case by case basis

Acknowledgements

Greg Ladics

Andre Silvanovitch (Monsanto)

Gary Bannon (Monsanto)

Pioneer Crop Genetics Regulatory Science & Registration

The FARRP dataset (www.allergenonline.com)

Allergome (Adriano Mari)