Bioinformatics tools to evaluate potential risks of celiac disease from novel proteins

Richard E. Goodman

Research Professor Dept. of Food Science & Technology
Food Allergy Research and Resource Program (FARRP)
University of Nebraska—Lincoln, USA
rgoodman2@unl.edu

Plaimein Amnuaycheewa

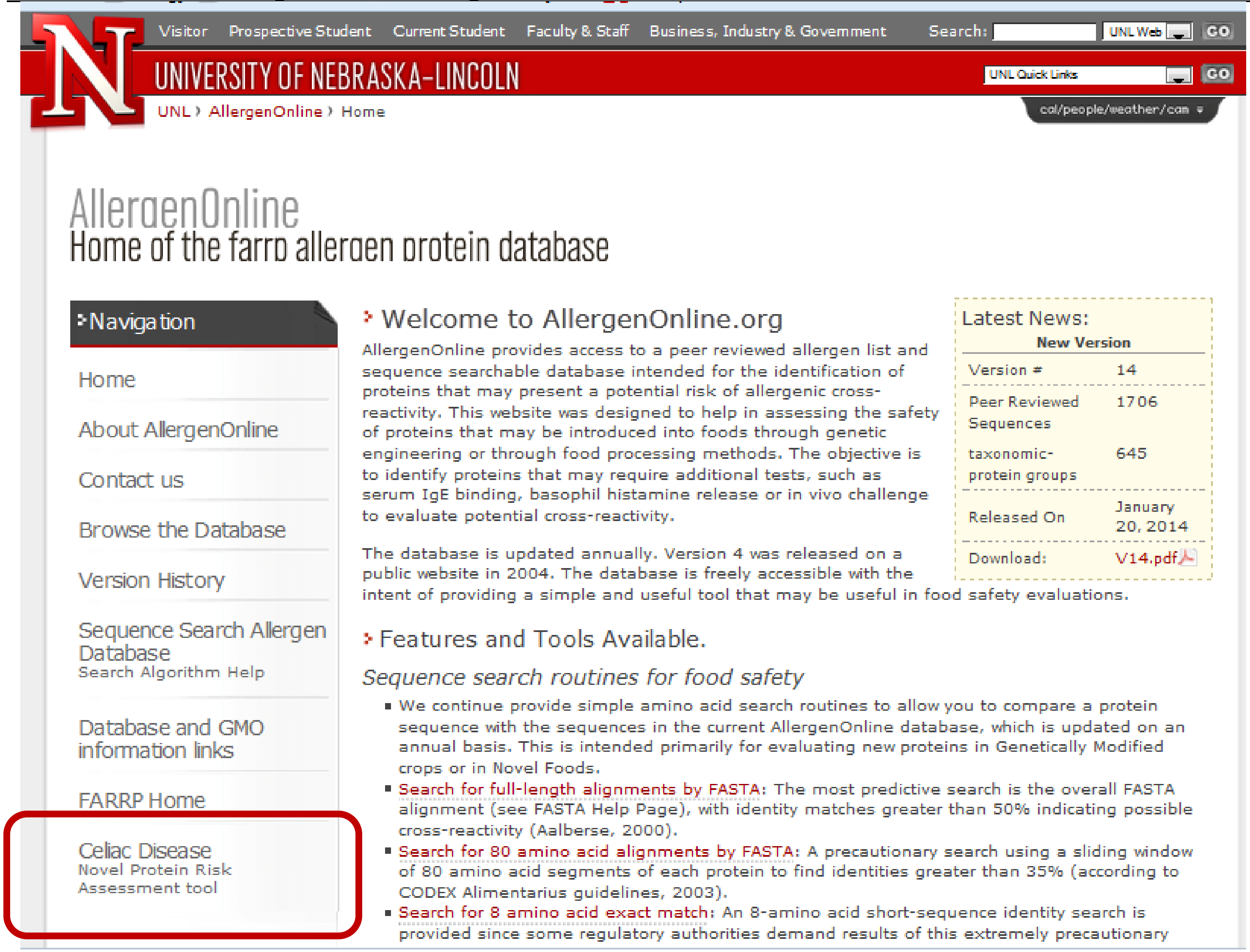
Ph.D. Student from Thailand - UNL

Sponsors of AllergenOnline.org: 8 Biotechnology Companies

17 June, 2015 EFSA Workshop on
Allergenicity Assessment of GM Plants

food allergy research
& resource program
farrp

www.Allergenonline.org – see Celiac Disease (below)



AllergenOnline
Home of the farrp allergen protein database

Welcome to AllergenOnline.org

AllergenOnline provides access to a peer reviewed allergen list and sequence searchable database intended for the identification of proteins that may present a potential risk of allergenic cross-reactivity. This website was designed to help in assessing the safety of proteins that may be introduced into foods through genetic engineering or through food processing methods. The objective is to identify proteins that may require additional tests, such as serum IgE binding, basophil histamine release or in vivo challenge to evaluate potential cross-reactivity.

The database is updated annually. Version 4 was released on a public website in 2004. The database is freely accessible with the intent of providing a simple and useful tool that may be useful in food safety evaluations.

Latest News:

New Version

Version #	14
Peer Reviewed Sequences	1706
taxonomic-protein groups	645
Released On	January 20, 2014
Download:	V14.pdf

Features and Tools Available.

Sequence search routines for food safety

- We continue provide simple amino acid search routines to allow you to compare a protein sequence with the sequences in the current AllergenOnline database, which is updated on an annual basis. This is intended primarily for evaluating new proteins in Genetically Modified crops or in Novel Foods.
- Search for full-length alignments by FASTA: The most predictive search is the overall FASTA alignment (see FASTA Help Page), with identity matches greater than 50% indicating possible cross-reactivity (Aalberse, 2000).
- Search for 80 amino acid alignments by FASTA: A precautionary search using a sliding window of 80 amino acid segments of each protein to find identities greater than 35% (according to CODEX Alimentarius guidelines, 2003).
- Search for 8 amino acid exact match: An 8-amino acid short-sequence identity search is provided since some regulatory authorities demand results of this extremely precautionary

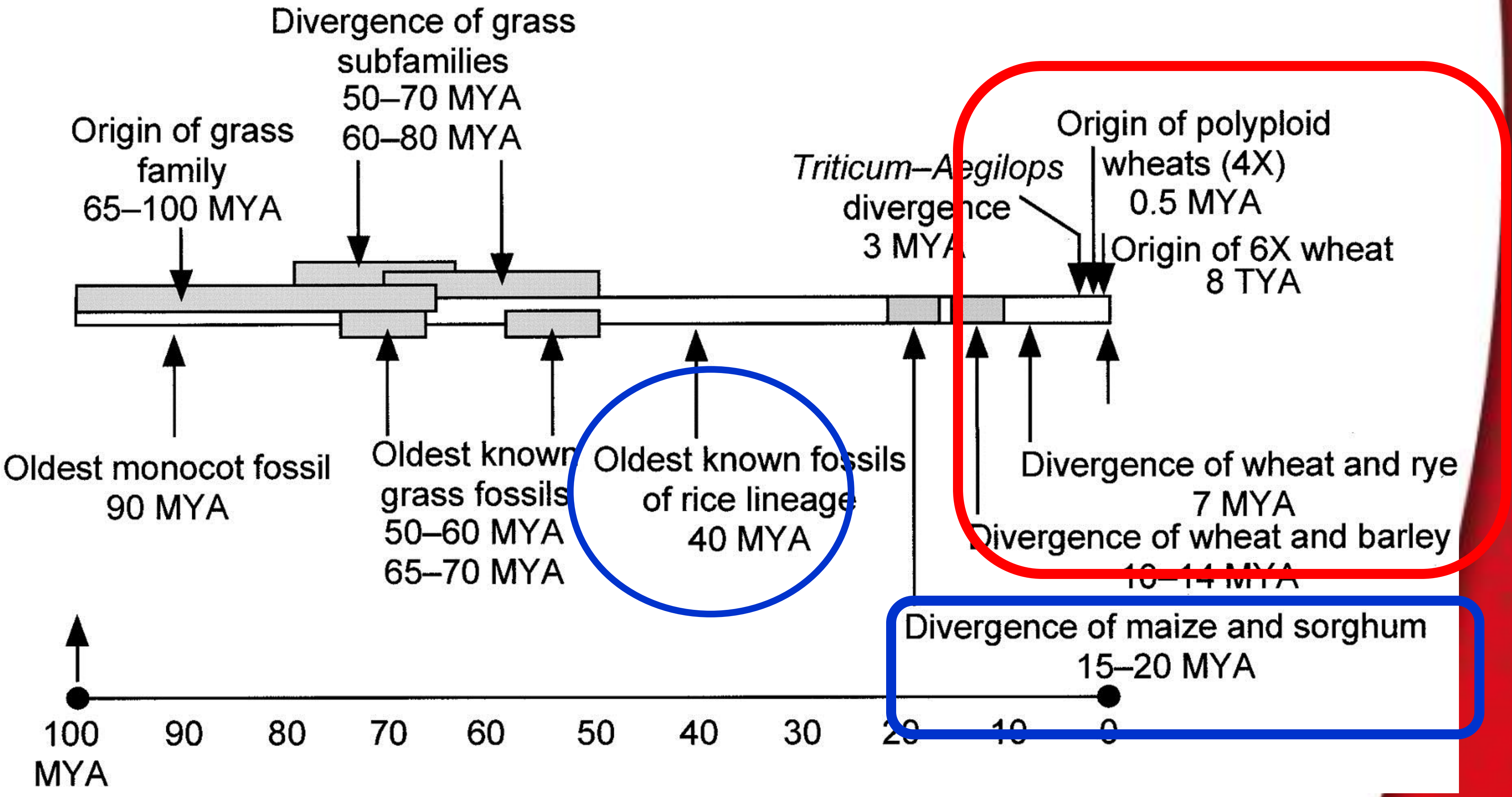
Outline: Development of Celiac Database and bioinformatics tests

- Celiac Disease (CD) affects ~ 1.3% of the population
- Developers of biotechnology products (genetically modified crops – GM) or food processors need a test:
  - **Codex Alimentarius Guideline for GM Safety (2003): any gene (protein) transferred from wheat or near wheat relative should be evaluated for potential to cause Celiac Disease**
- Bioinformatics should provide efficient evaluation to demonstrate:
  - the SAFETY of >98+% of wheat proteins that would not cause CD
  - And RISK ~ 100% identification of those that would cause CD……http://www.allergenonline.org/celiachome.shtml
- I will present selection of Celiac eliciting peptides and representative proteins (for immunogenic and toxic properties)
- Testing
  - Exact peptide matching
  - FASTA Identity and E-score evaluation to set criteria

food allergy research & resource program
farrp

Grasses as a single genetic system

Evolutionary history of cereals – AND GLUTEN COMPLEX Divergence

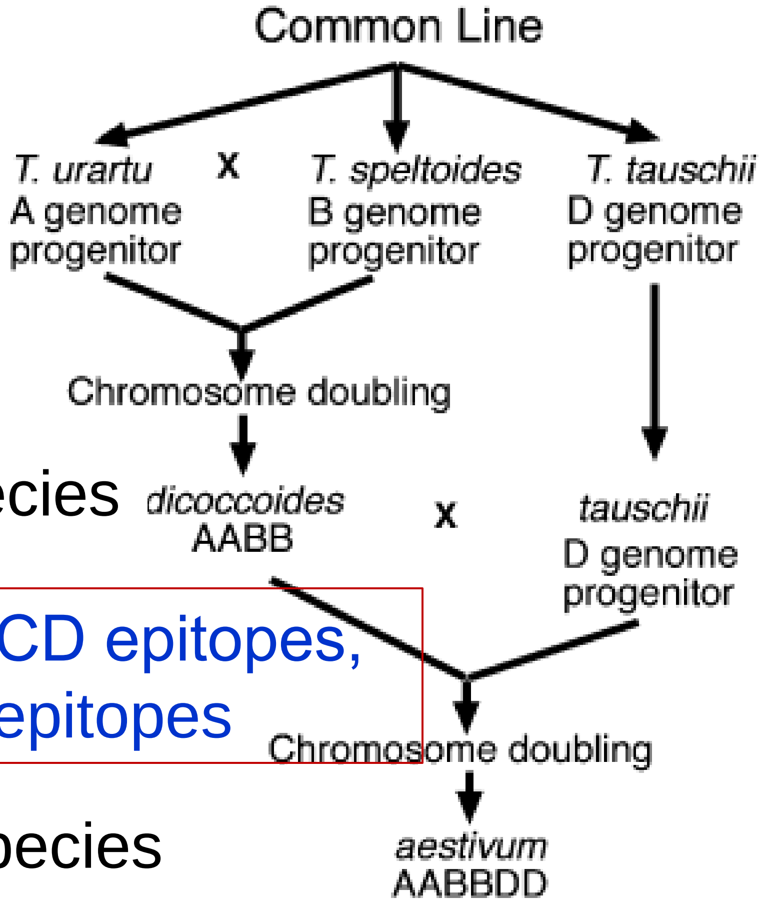

Divergence of grass subfamilies
50–70 MYA
60–80 MYA
Origin of grass family
65–100 MYA
Triticum–Aegilops divergence
3 MYA
Origin of polyploid wheats (4X)
0.5 MYA
Origin of 6X wheat
8 TYA
Oldest monocot fossil
90 MYA
Oldest known grass fossils
50–60 MYA
65–70 MYA
Oldest known fossils of rice lineage
40 MYA
Divergence of wheat and rye
7 MYA
Divergence of wheat and barley
10–14 MYA
Divergence of maize and sorghum
15–20 MYA
100 MYA
90
80
70
60
50
40
30
20
10
0

Gill, B. S. et al. Genetics 2004;168:1087-1096

Probable Evolution of modern wheat

D. Kasarda, JAFC 2013 61:1155-1159

Gluten genetics are complex, gene duplication + divergence



Common Line

T. urartu
A genome
progenitor

x

T. speltoides
B genome
progenitor

T. tauschii
D genome
progenitor

Chromosome doubling

dicoccoides
AABB

x

tauschii
D genome
progenitor

Chromosome doubling

aestivum
AABBDD

Diploid wheat species

Tetraploid wheat species

"D" genome has more CD epitopes, but all 3 have multiple epitopes

Hexaploid wheat species

Figure 3. Combinations of diploid wheats leading to the polyploid forms.

Immunogenic effect is the most common problem, associated with MHCII DQ 2 or 8

+

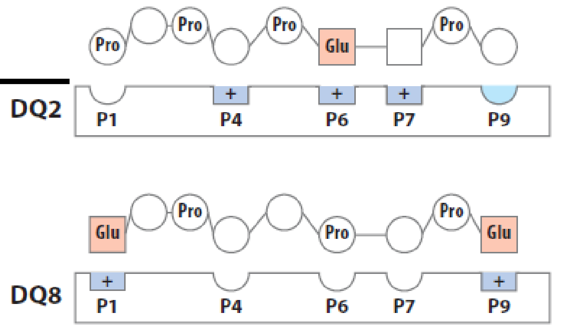
Digestion

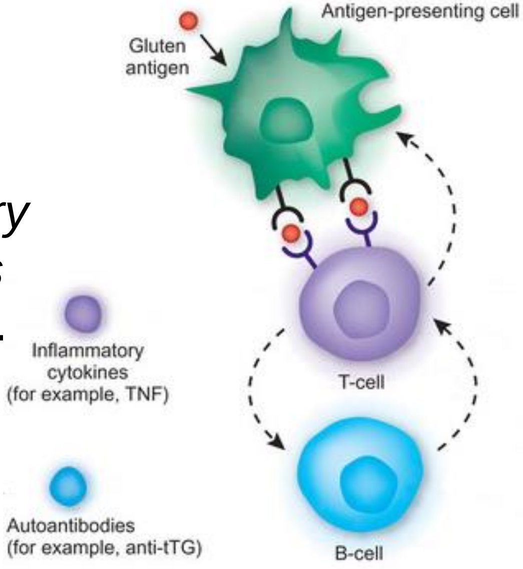
PQPQLPYPQ

Digestion resistant peptide

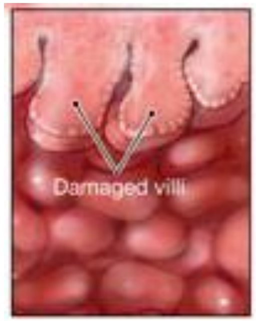
Tissue transglutaminase

PQP E L P Y



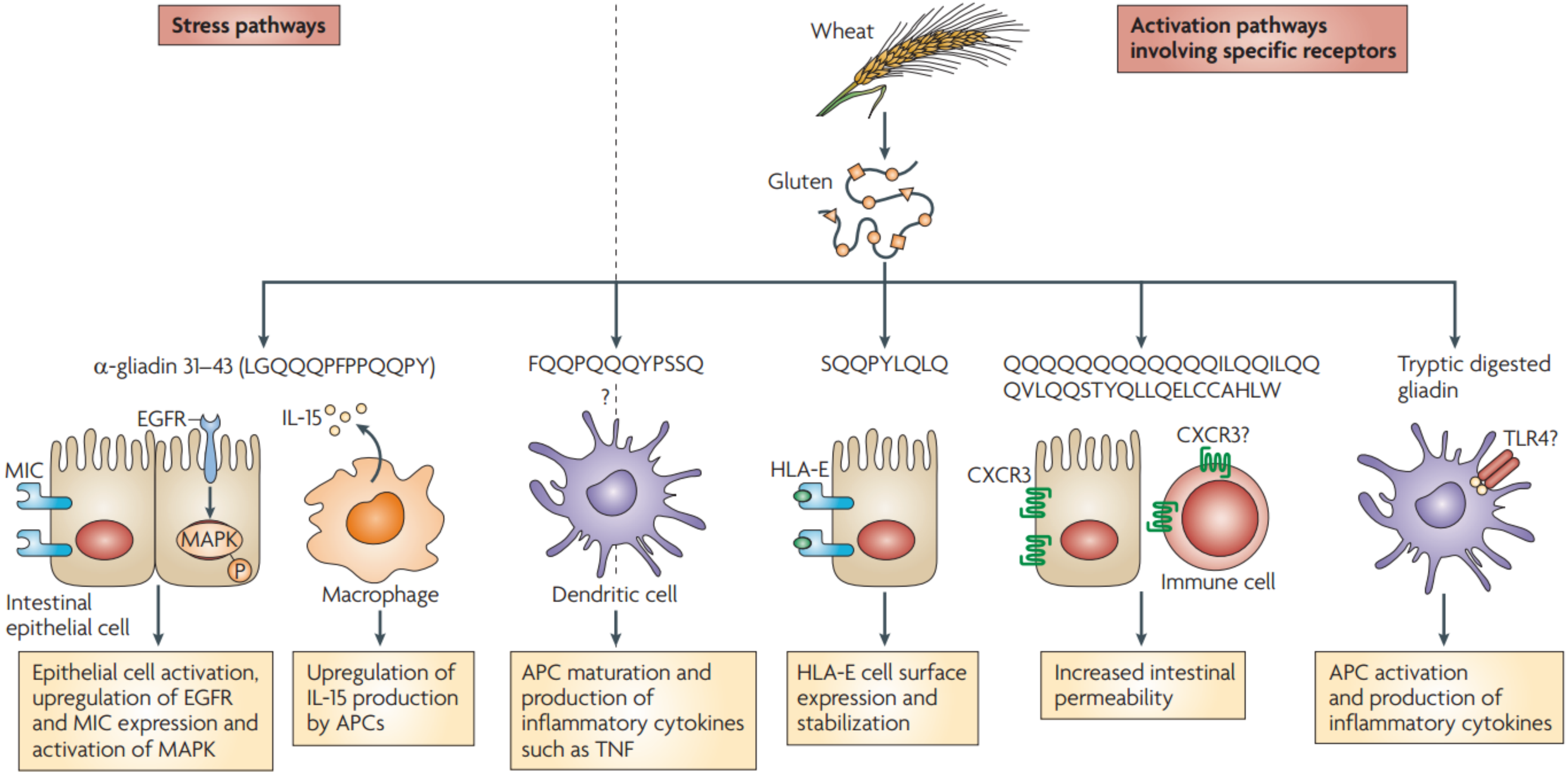

Antigen-presenting cell
Gluten antigen
T-cell
Inflammatory cytokines (for example, TNF)
Autoantibodies (for example, anti-tTG)
B-cell

Inflammatory Reactions



(Abadie et al., 2011 ; http://bnljceliacdisease.wordpress.com /2010/11/22/celiac-disease-at-a-cellular-level/)

Toxic effects - less well defined, - may bind innate immune receptors, rather than MHC / TCR


Stress pathways
Activation pathways involving specific receptors
Wheat
Gluten
α-gliadin 31–43 (LGQQQPFPPQQPY)
FQQPQQQYPSSQ
SQQPYLQLQ
QQQQQQQQQQQQILQQILQQ QVLQQSTYQLLQELCCAHLW
Tryptic digested gliadin
EGFR
IL-15
MIC
MAPK
P
Intestinal epithelial cell
Macrophage
?
Dendritic cell
HLA-E
CXCR3
CXCR3?
Immune cell
TLR4?
Epithelial cell activation, upregulation of EGFR and MIC expression and activation of MAPK
Upregulation of IL-15 production by APCs
APC maturation and production of inflammatory cytokines such as TNF
HLA-E cell surface expression and stabilization
Increased intestinal permeability
APC activation and production of inflammatory cytokines

(Jabri & Sollid, 2011)

Building the Database (2010-2012)

1. Plaimein reviewed 53 CD research papers identified from PubMed
2. Identified 1,016 peptides with CD related activity → Peptide database
   a. **464 native** & **552 deamidated** peptides
   b. **Toxic** peptides = **18**
   c. Length of peptides from 8 to 55 aa
3. BLAST search with each peptide against NCBI - Protein database *(non-redundant sequences)* to identify the “parent proteins”

Method of selecting CD peptides

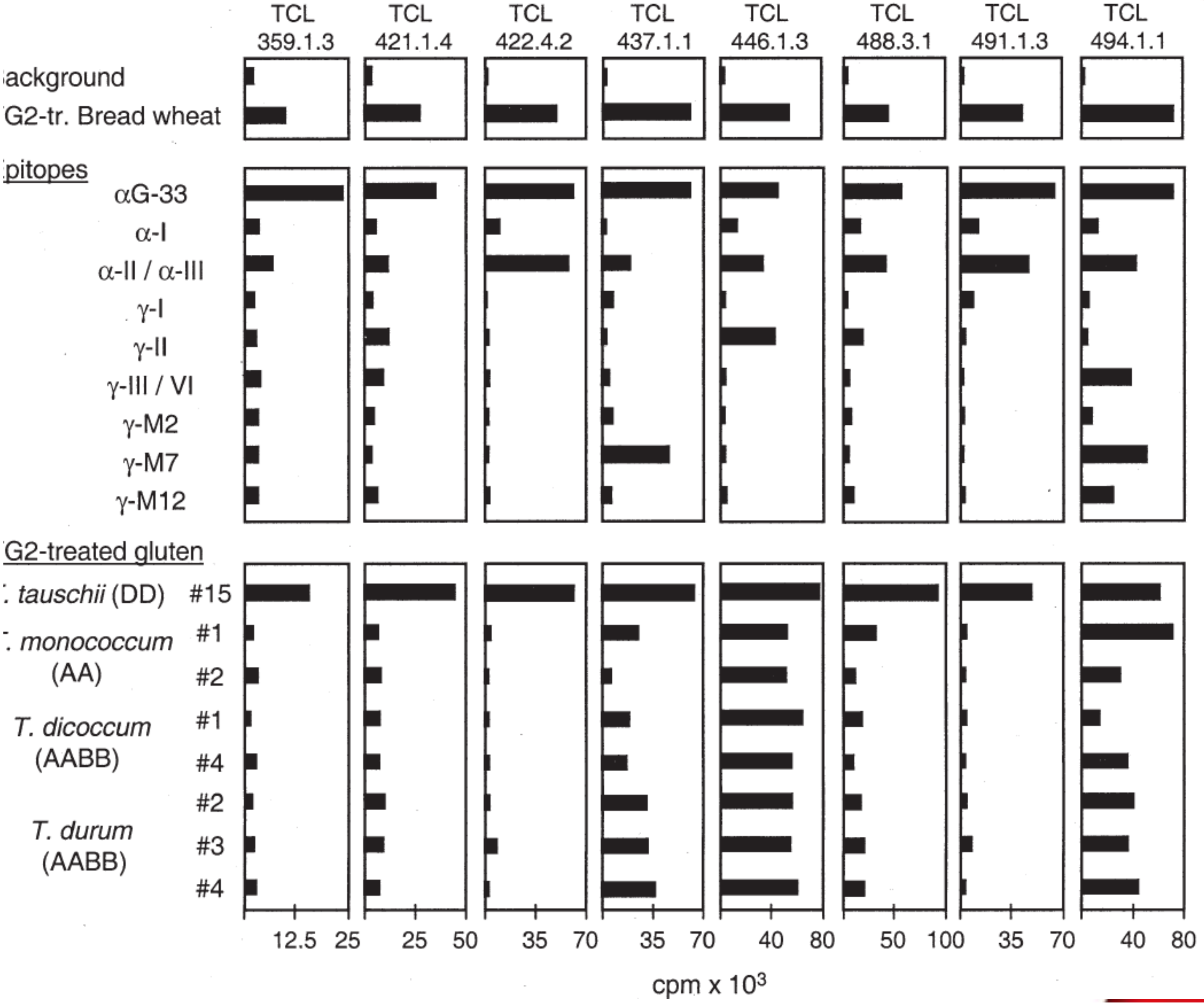
1. Literature search for “celiac” and “coeliac” in PubMed
   52 publications 1984 - 2011

2. Peptides/proteins are selected based on evidence in CD patients:
   - Immunogenic: able to bind the HLA-DQ molecules & stimulate T-cell response (proliferation &/or IFNγ secretion)
   - Toxic: able to
     - Stimulate & release proinflamatory cytokine IL-15
     - Reduce brush border alkaline phosphatase activity
     - Reduce mean enterocyte surface cell height
     - Induce partial villus atrophy
     - Induce increase in intestinal permeability
   - Collected sequences:
     - **473 native peptides and 558 deamidated peptides**
     - **18 peptides were toxic and not immunogenic**

8 - T cell LINES diversity in proliferation

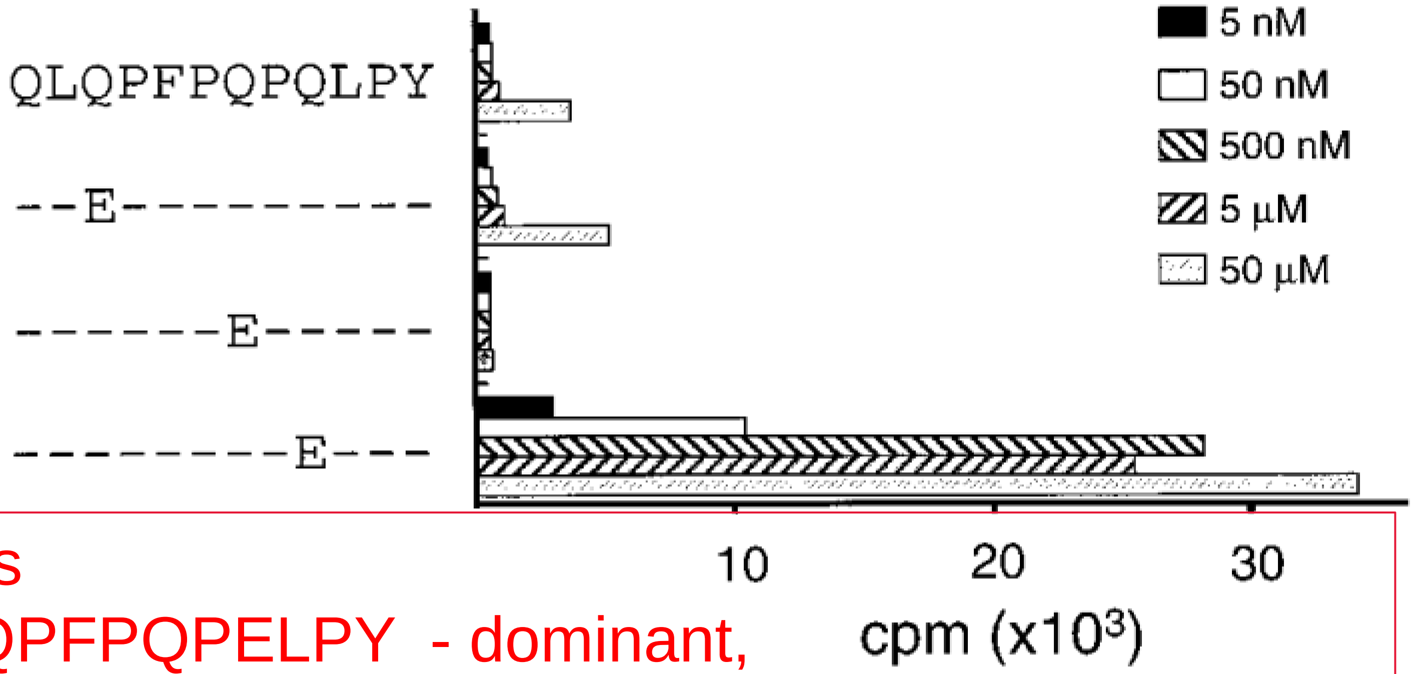
Tested with various gluten peptides

Or wheat gluten extracts treated with TG2



APC; DQ2 homozygous, EBV transformed B-LCL were irradiated ; incubated with TCL cells and 100µg/mL TG2-treated gluten or 5µmol/L gliadin peptide epitopes

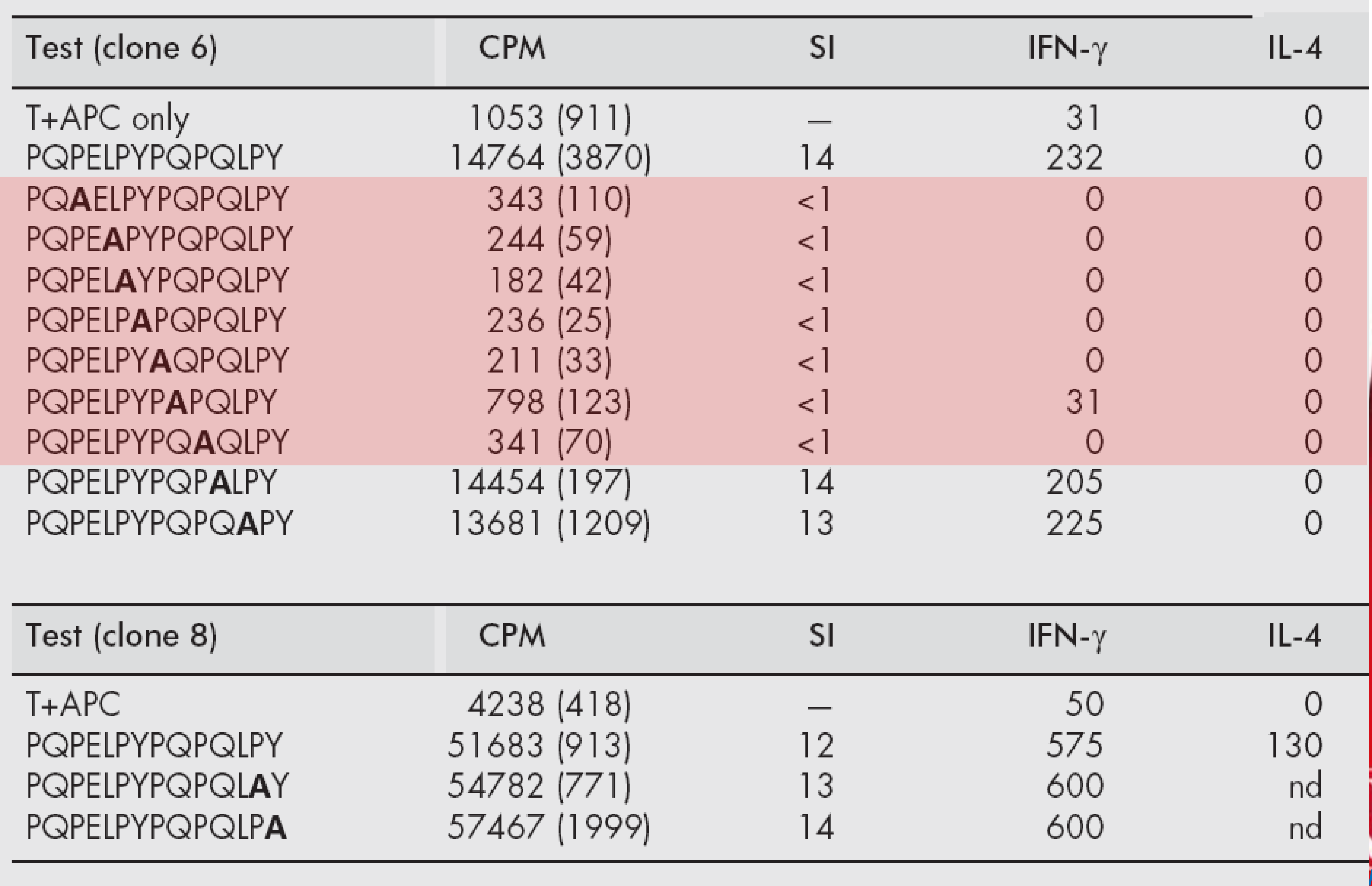
Testing specific deamidation (is rarely done this well, but this is only 1 TCL, T cell line, not a clone)



Thus
QLQPFPQPELPY - dominant,
But 3 other peptides stimulate slightly in this study…

MHC II DQ2.5, DR3+ DQ2+ B-LCL were irradiated and incubated with TCL cells using native peptide or synthetic peptides to simulate specific deamidation (Q > E)

Peptide divergence test: alanine substitutions in alpha-2 gliadin peptide G4: T cell clones Ellis et al. Gut 2003, 52:212-217

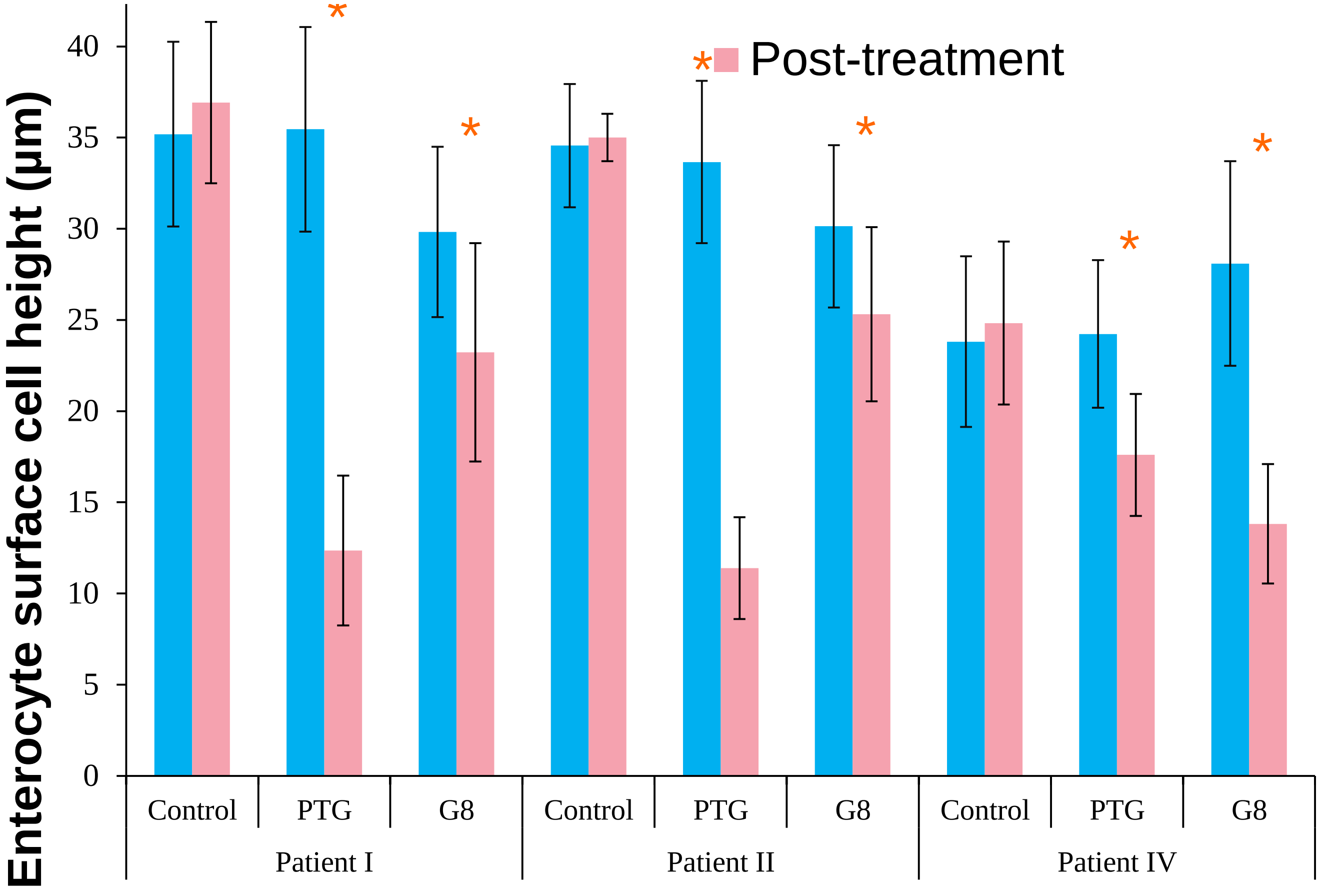


Test (clone 6)	CPM	SI	IFN-γ	IL-4
T+APC only	1053 (911)	—	31	0
PQPELPYPQPQLPY	14764 (3870)	14	232	0
PQ**A**ELPYPQPQLPY	343 (110)	<1	0	0
PQPE**A**PYPQPQLPY	244 (59)	<1	0	0
PQPEL**A**YPQPQLPY	182 (42)	<1	0	0
PQPELP**A**PQPQLPY	236 (25)	<1	0	0
PQPELPY**A**QPQLPY	211 (33)	<1	0	0
PQPELPYP**A**PQLPY	798 (123)	<1	31	0
PQPELPYPQ**A**QLPY	341 (70)	<1	0	0
PQPELPYPQP**A**LPY	14454 (197)	14	205	0
PQPELPYPQPQ**A**PY	13681 (1209)	13	225	0

Test (clone 8)	CPM	SI	IFN-γ	IL-4
T+APC	4238 (418)	—	50	0
PQPELPYPQPQLPY	51683 (913)	12	575	130
PQPELPYPQPQL**A**Y	54782 (771)	13	600	nd
PQPELPYPQPQLP**A**	57467 (1999)	14	600	nd

Toxic effect of α-gliadin G8 peptide (LQLQPFPQPQLPYPQPQLPY) following proteolysis (pepsin and trypsin) of peptide and dosing 4 hours in vitro Enterocyte cell height of 3 CD patient biopsies

Fraser et al. Gut 2003, 52: 1698-1702.



MVRVPVPQLQPQNPSQQQPQEQVPLVQQQQFPGQQQPFPPQQPYPQPQPFPSQQPYLQLQPFPQPQLPYPQPQLPYPQPQLPYPQPQPFRPQQPYPQSQPQ YSQPQQPISQQQQQQQQQQQQKQQ

GI7209265 α-gliadin [*Triticum aestivum*] 290 aa

VRVPVPQLQPQNPSQQQPQEQVPLVQQQQF
VRVPVPQLQPQNPSQQQPQ
VPVPQLQPQNPSQQQPQEQVPL
QNPSQQQPQEQVPLVQQQ
VQQQQFPGQQQPFPPQQPYPQPQPFPSQQPY
FPGQQQPFPPQQPYPQPQPF
QPYPQPQPFPSQQPYLQL
PQPQPFPSQQPY
YLQLQPFPQPQLPYPQPQLP
YLQLQPFPQPQLPYP
LQLQPFPQPQLPYPQPQLPYPQPQLPYPQPQPF
LQLQPFPQPQLPYPQPQLPY
LQLQPFPQPQLPY
QLQPFPQPQLPYPQPQ
QLQPFPQPQLPYPQP
QLQPFPQPQLPYPQ
QLQPFPQPQLPY
LQPFPQPQLPYPQPQ
QPFPQPQLPYPQ
QPFPQPQLPY
PFPQPQLPYPQPQLP
PFPQPQLPYPQ
PFPQPQLPY
PQPQLPYPQPQLPY
PQPQLPYPQPQL
PQPQLPYPQPQ
PQPQLPYPQ
QPQLPYPQPQLPYPQ
PQLPYPQPQLPY
QLPYPQPQLPYPQPQ
QLPYPQPQLPYPQ
LPYPQPQLPYPQ
PYPQPQLPY
PQLPYPQPQLPYPQPQPFRP
YPQPQLPYPQPQPFRP
YPQPQLPYPQPQPFR
FRPQQPYPQ

48 unique exact peptide matches

QQQQQQILQQILQQQLIPCRDVVLQQHSIAYGSSQVLQQSTYQLVQQLCCQQLWQIPEQSRCQAIHNVVHAIILHQQQQQQQQQQQQPLSQVSFQQPQQQYPSGQGSFQPSQQNPQAQGSVQPQQLPQFE

QQPQQQYPSGQGSFQPSQQNPQAQG
QQPQQQYPSGQGSFQPSQQNPQAQ
QPQQQYPSGQGSFQPSQQNP
QQYPSGQGSFQPSQQNPQ
YPSGQGSFQPSQQNP
PSGQGSFQPSQQNPQAQG
PSGQGSFQPSQQ
PSGQGSFQPSQ
QGSFQPSQQ
GSFQPSQQNPQAQGS
QAQGSVQPQQLPQFE

EIRNLALETLPAMCNVYIPPYCTIAPVGIFGTNYR

Database 68 Proteins (2010-2012)

3. The proteins were aligned by ClustalW2 to remove redundant protein sequences

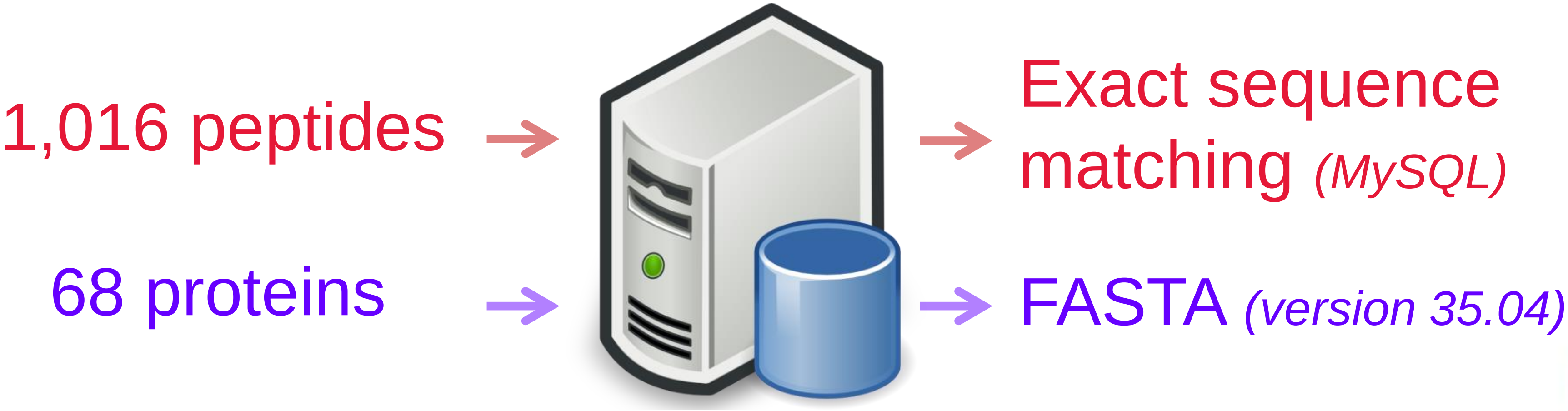
4. 68 representative parent proteins → Protein database

- *Triticum aestivum* (43)
- *Triticum monococcum* (2)
- *Hordeum vulgare* (11)
- *Hordeum vulgare subsp. vulgare* (7)
- *Secale cereale* (6)
- *Avena sativa* (3)
- *Avena nuda* (2)
- HMW glutenin synthetic construct (1)

Short protein fragments : 20, 29, 43, 52, 54, 68, 72 aa

Full protein lengths : 150-839 aa

CD Database and search routines
(2010-2012)



1,016 peptides →

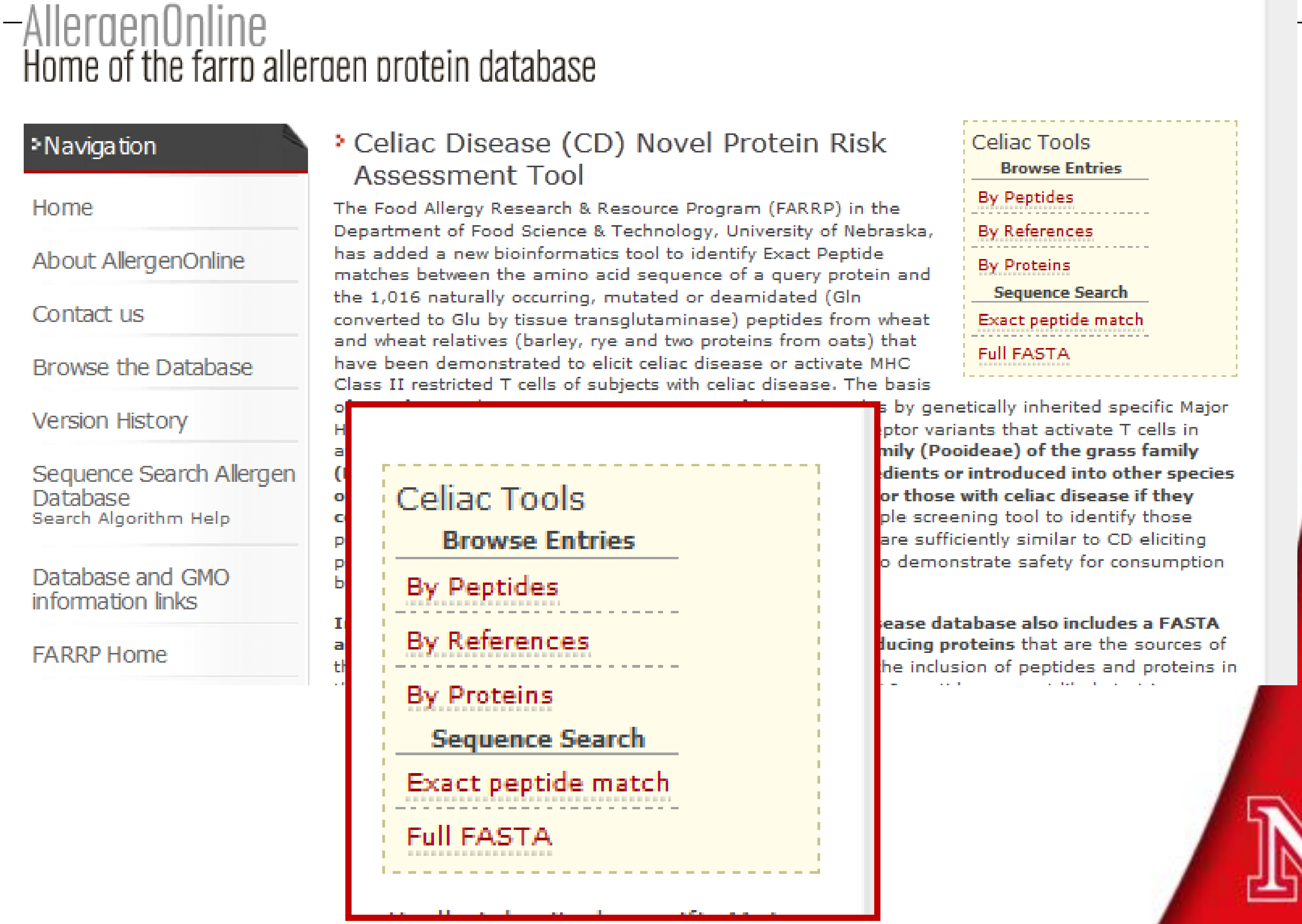
68 proteins →

→ Exact sequence matching *(MySQL)*

→ FASTA *(version 35.04)*

Plaimein Amnuaycheewa identified sequences
John Wise constructed the MYSQL database and search routines
Verified by Afua Tetteh and Rick Goodman

www.AllergenOnline.org/celiachome.shtml
Browse and Search functions



AllergenOnline
Home of the farrp allergen protein database

Navigation

- Home
- About AllergenOnline
- Contact us
- Browse the Database
- Version History
- Sequence Search Allergen Database
  Search Algorithm Help
- Database and GMO information links
- FARRP Home

Celiac Disease (CD) Novel Protein Risk Assessment Tool

The Food Allergy Research & Resource Program (FARRP) in the Department of Food Science & Technology, University of Nebraska, has added a new bioinformatics tool to identify Exact Peptide matches between the amino acid sequence of a query protein and the 1,016 naturally occurring, mutated or deamidated (Gln converted to Glu by tissue transglutaminase) peptides from wheat and wheat relatives (barley, rye and two proteins from oats) that have been demonstrated to elicit celiac disease or activate MHC Class II restricted T cells of subjects with celiac disease. The basis

o… s by genetically inherited specific Major
H… ptor variants that activate T cells in
a… **mily (Pooideae) of the grass family**
(… dients or introduced into other species
o… or those with celiac disease if they
c… ple screening tool to identify those
p… are sufficiently similar to CD eliciting
p… o demonstrate safety for consumption
b…

I… ease database also includes a FASTA
a… ducing proteins that are the sources of
th… he inclusion of peptides and proteins in

Celiac Tools

Browse Entries

- By Peptides
- By References
- By Proteins

Sequence Search

- Exact peptide match
- Full FASTA

Celiac Tools

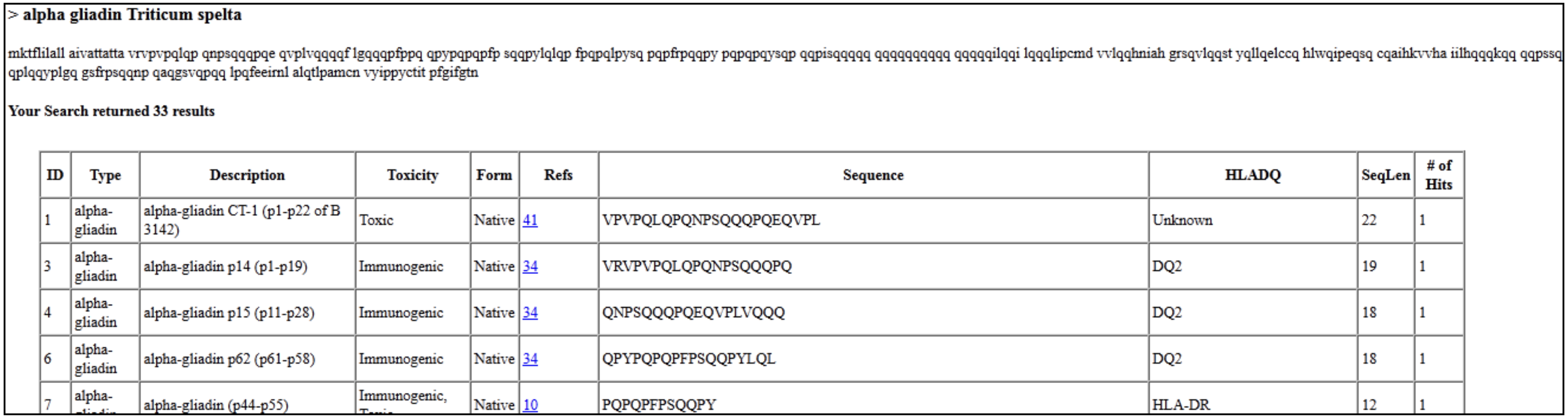
Browse Entries

- By Peptides
- By References
- By Proteins

Sequence Search

- Exact peptide match
- Full FASTA

Peptide Exact match alpha-gliadin Triticum spelta var. arduini GI:3928509

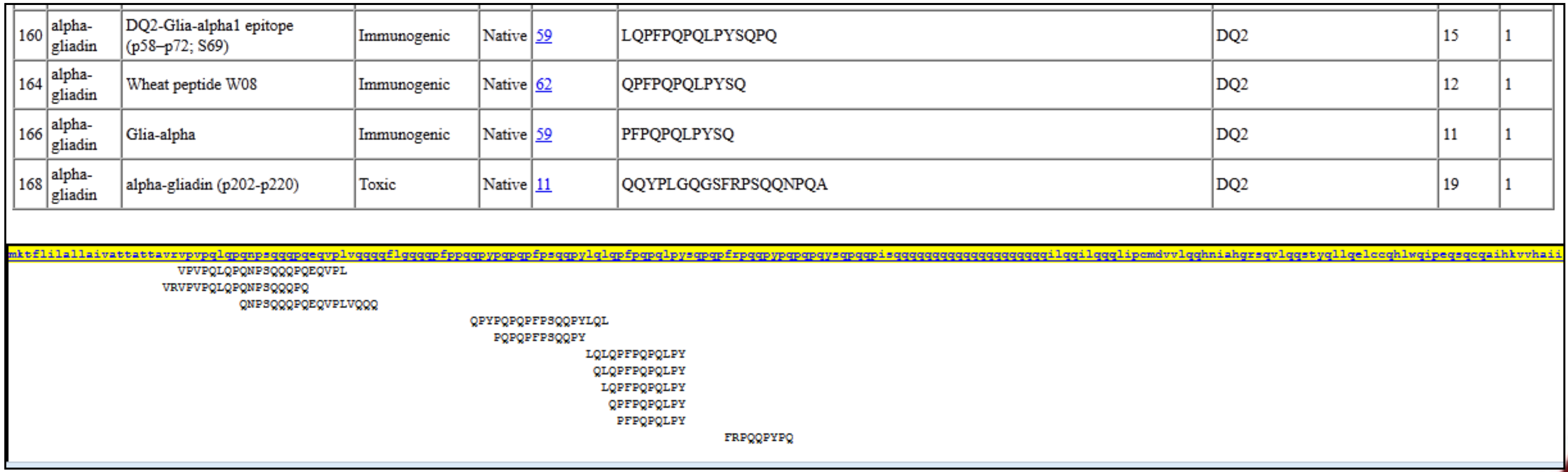


> alpha gliadin Triticum spelta

mktflilall aivattatta vrvpvpqlqp qnpsqqqpqe qvplvqqqqf lgqqqpfppq qpypqpqpfp sqqpylqlqp fpqpqlpysq pqpfrpqqpy pqpqpqysqp qqpisqqqqq qqqqqqqqq qqqqqilqqi lqqqlipcmd vvlqqhniah grsqvlqqst yqllqelccq hlwqipeqsq cqaihkvvha iilhqqqkqq qqpssq qplqqyplgq gsfrpsqqnp qaqgsvqpqq lpqfeeirnl alqtlpamcn vyippyctit pfgifgtn

Your Search returned 33 results

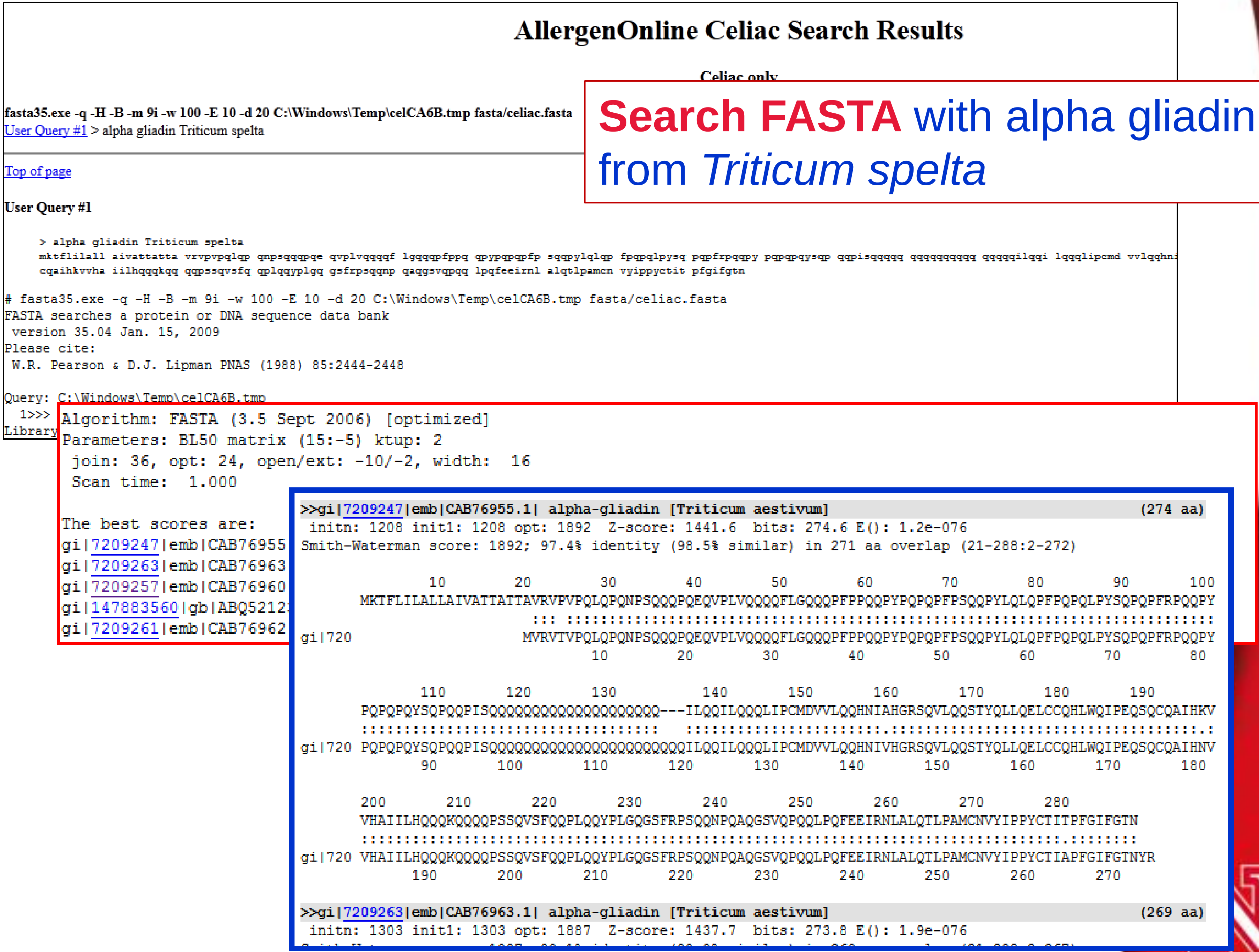
ID	Type	Description	Toxicity	Form	Refs	Sequence	HLADQ	SeqLen	# of Hits
1	alpha-gliadin	alpha-gliadin CT-1 (p1-p22 of B 3142)	Toxic	Native	41	VPVPQLQPQNPSQQQPQEQVPL	Unknown	22	1
3	alpha-gliadin	alpha-gliadin p14 (p1-p19)	Immunogenic	Native	34	VRVPVPQLQPQNPSQQQPQ	DQ2	19	1
4	alpha-gliadin	alpha-gliadin p15 (p11-p28)	Immunogenic	Native	34	QNPSQQQPQEQVPLVQQQ	DQ2	18	1
6	alpha-gliadin	alpha-gliadin p62 (p61-p58)	Immunogenic	Native	34	QPYPQPQPFPSQQPYLQL	DQ2	18	1
7	alpha-	alpha-gliadin (p44-p55)	Immunogenic,	Native	10	PQPQPFPSQQPY	HLA-DR	12	1



160	alpha-gliadin	DQ2-Glia-alpha1 epitope (p58–p72; S69)	Immunogenic	Native	59	LQPFPQPQLPYSQPQ	DQ2	15	1
164	alpha-gliadin	Wheat peptide W08	Immunogenic	Native	62	QPFPQPQLPYSQ	DQ2	12	1
166	alpha-gliadin	Glia-alpha	Immunogenic	Native	59	PFPQPQLPYSQ	DQ2	11	1
168	alpha-gliadin	alpha-gliadin (p202-p220)	Toxic	Native	11	QQYPLGQGSFRPSQQNPQA	DQ2	19	1

```
mktflilallaivattattavrvpvpqlqpqnpsqqqpqeqvplvqqqqflqqqqpfppqqpypqpqpfpsqqpylqlqpfpqpqlpysqpqpfrpqqpypqpqpqysqpqqpisqqqqqqqqqqqqqqqqqqqqilqqilqqqlipcmdvvlqqhniahgrsqvlqqstyqllqelccqhlwqipeqsqcqaihkvvhaii
                      VPVPQLQPQNPSQQQPQEQVPL
                    VRVPVPQLQPQNPSQQQPQ
                              QNPSQQQPQEQVPLVQQQ
                                                            QPYPQPQPFPSQQPYLQL
                                                               PQPQPFPSQQPY
                                                                           LQLQPFPQPQLPY
                                                                            QLQPFPQPQLPY
                                                                             LQPFPQPQLPY
                                                                              QPFPQPQLPY
                                                                               PFPQPQLPY
                                                                                             FRPQQPYPQ
```

Search FASTA with alpha gliadin from *Triticum spelta*



AllergenOnline Celiac Search Results

Celiac only

fasta35.exe -q -H -B -m 9i -w 100 -E 10 -d 20 C:\Windows\Temp\celCA6B.tmp fasta/celiac.fasta
User Query #1 > alpha gliadin Triticum spelta

Top of page

User Query #1

```
> alpha gliadin Triticum spelta
mktflilall aivattatta vrvpvpqlqp qnpsqqqpqe qvplvqqqqf lgqqqpfppq qpypqpqpfp sqqpylqlqp fpqpqlpysq pqpfrpqqpy pqpqpqysqp qqpisqqqqq qqqqqqqqqq qqqqqilqqi lqqqlipcmd vvlqqhni
cqaihkvvha iilhqqqkqq qqpssqvsfq qplqqyplgq gsfrpsqqnp qaqgsvqpqq lpqfeeirnl alqtlpamcn vyippyctit pfgifgtn
```

```
# fasta35.exe -q -H -B -m 9i -w 100 -E 10 -d 20 C:\Windows\Temp\celCA6B.tmp fasta/celiac.fasta
FASTA searches a protein or DNA sequence data bank
 version 35.04 Jan. 15, 2009
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

Query: C:\Windows\Temp\celCA6B.tmp
  1>>>
Library
```

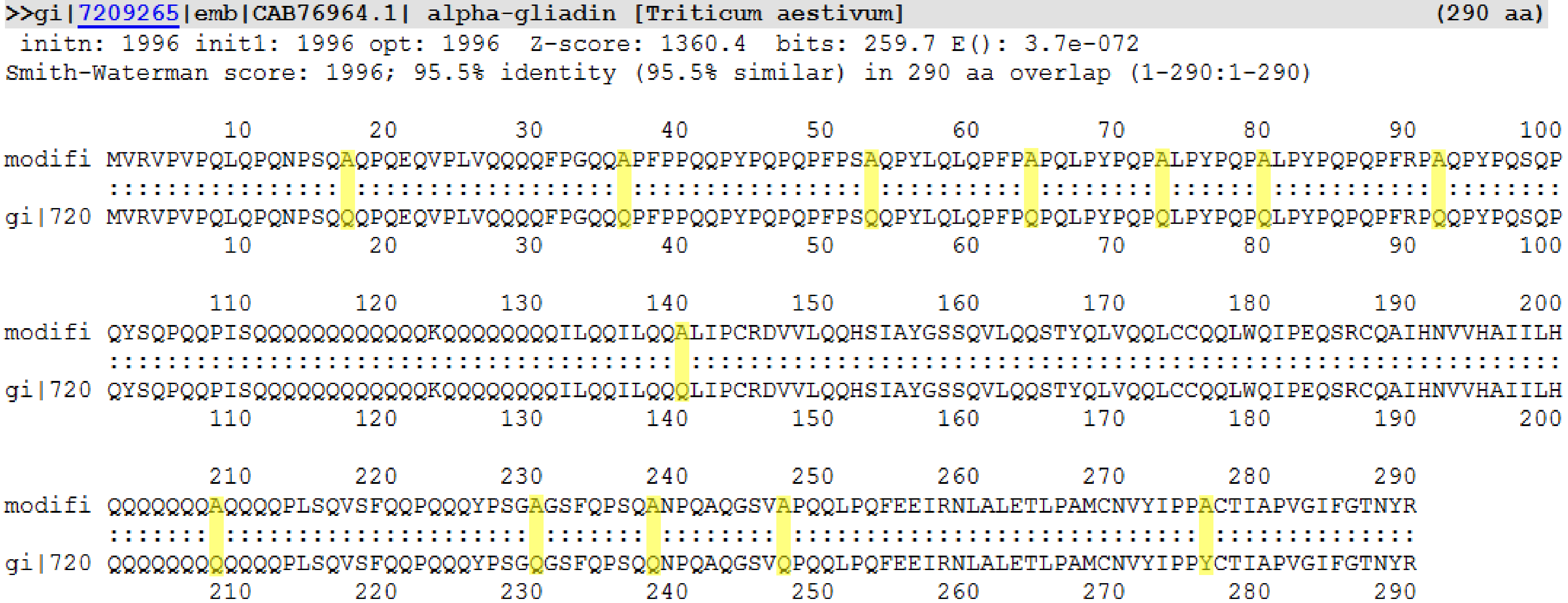
```
Algorithm: FASTA (3.5 Sept 2006) [optimized]
Parameters: BL50 matrix (15:-5) ktup: 2
 join: 36, opt: 24, open/ext: -10/-2, width:  16
 Scan time:  1.000

The best scores are:
gi|7209247|emb|CAB76955
gi|7209263|emb|CAB76963
gi|7209257|emb|CAB76960
gi|147883560|gb|ABQ5212
gi|7209261|emb|CAB76962
```

```
>>gi|7209247|emb|CAB76955.1| alpha-gliadin [Triticum aestivum]                 (274 aa)
 initn: 1208 init1: 1208 opt: 1892  Z-score: 1441.6  bits: 274.6 E(): 1.2e-076
Smith-Waterman score: 1892; 97.4% identity (98.5% similar) in 271 aa overlap (21-288:2-272)

               10        20        30        40        50        60        70        80        90       100
       MKTFLILALLAIVATTATTAVRVPVPQLQPQNPSQQQPQEQVPLVQQQQFLGQQQPFPPQQPYPQPQPFPSQQPYLQLQPFPQPQLPYSQPQPFRPQQPY
                           ::: ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|720                    MVRVTVPQLQPQNPSQQQPQEQVPLVQQQQFLGQQQPFPPQQPYPQPQPFPSQQPYLQLQPFPQPQLPYSQPQPFRPQQPY
                                  10        20        30        40        50        60        70        80

              110       120       130          140       150       160       170       180       190
       PQPQPQYSQPQQPISQQQQQQQQQQQQQQQQQQQQ---ILQQILQQQLIPCMDVVLQQHNIAHGRSQVLQQSTYQLLQELCCQHLWQIPEQSQCQAIHKV
       :::::::::::::::::::::::::::::::::::   :::::::::::::.::::::::::::::::::::::::::::::::::::::::::::::.:
gi|720 PQPQPQYSQPQQPISQQQQQQQQQQQQQQQQQQQQQQQILQQILQQQLIPCMDVVLQQHNIVHGRSQVLQQSTYQLLQELCCQHLWQIPEQSQCQAIHNV
               90       100       110       120       130       140       150       160       170       180

       200       210       220       230       240       250       260       270       280
       VHAIILHQQQKQQQQPSSQVSFQQPLQQYPLGQGSFRPSQQNPQAQGSVQPQQLPQFEEIRNLALQTLPAMCNVYIPPYCTITPFGIFGTN
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::.::::::::
gi|720 VHAIILHQQQKQQQQPSSQVSFQQPLQQYPLGQGSFRPSQQNPQAQGSVQPQQLPQFEEIRNLALQTLPAMCNVYIPPYCTIAPFGIFGTNYR
              190       200       210       220       230       240       250       260       270

>>gi|7209263|emb|CAB76963.1| alpha-gliadin [Triticum aestivum]                 (269 aa)
 initn: 1303 init1: 1303 opt: 1887  Z-score: 1437.7  bits: 273.8 E(): 1.9e-076
```

Why do a FASTA? Why not just exact match?

- Either Natural or artificial mutations could lead to proteins with no exact match, but that pose high risk.
- Example, modified alpha gliadin (next slide)
- Would such a protein be safe?
- FASTA may prove a good backup for exact match….BUT WE NEEDED to establish CRITERIA and LIMITS

MODIFIED α-gliadin [*Triticum aestivum*] alanine substitutions for glutamine



```
>>gi|7209265|emb|CAB76964.1| alpha-gliadin [Triticum aestivum]                      (290 aa)
 initn: 1996 init1: 1996 opt: 1996  Z-score: 1360.4  bits: 259.7 E(): 3.7e-072
Smith-Waterman score: 1996; 95.5% identity (95.5% similar) in 290 aa overlap (1-290:1-290)

               10        20        30        40        50        60        70        80        90       100
modifi MVRVPVPQLQPQNPSQAQPQEQVPLVQQQQFPGQQAPFPPQQPYPQPQPFPSAQPYLQLQPFPAPQLPYPQPALPYPQPALPYPQPQPFRPAQPYPQSQP
       :::::::::::::::: :::::::::::::::::: :::::::::::::::: :::::::::: :::::::: :::::: ::::::::::: ::::::::
gi|720 MVRVPVPQLQPQNPSQQQPQEQVPLVQQQQFPGQQQPFPPQQPYPQPQPFPSQQPYLQLQPFPQPQLPYPQPQLPYPQPQLPYPQPQPFRPQQPYPQSQP
               10        20        30        40        50        60        70        80        90       100

              110       120       130       140       150       160       170       180       190       200
modifi QYSQPQQPISQQQQQQQQQQQQKQQQQQQQQILQQILQQALIPCRDVVLQQHSIAYGSSQVLQQSTYQLVQQLCCQQLWQIPEQSRCQAIHNVVHAIILH
       ::::::::::::::::::::::::::::::::::::::: ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|720 QYSQPQQPISQQQQQQQQQQQQKQQQQQQQQILQQILQQQLIPCRDVVLQQHSIAYGSSQVLQQSTYQLVQQLCCQQLWQIPEQSRCQAIHNVVHAIILH
              110       120       130       140       150       160       170       180       190       200

              210       220       230       240       250       260       270       280       290
modifi QQQQQQQAQQQQPLSQVSFQQPQQQYPSGAGSFQPSQANPQAQGSVAPQQLPQFEEIRNLALETLPAMCNVYIPPACTIAPVGIFGTNYR
       ::::::: ::::::::::::::::::::: ::::::: :::::::: :::::::::::::::::::::::::::: ::::::::::::::
gi|720 QQQQQQQQQQQQPLSQVSFQQPQQQYPSGQGSFQPSQQNPQAQGSVQPQQLPQFEEIRNLALETLPAMCNVYIPPYCTIAPVGIFGTNYR
              210       220       230       240       250       260       270       280       290
```

Theoretical 13 alanine substitutions in CD inducing peptides

E score = 3.7E-072, 95.5 % id over full length overlap

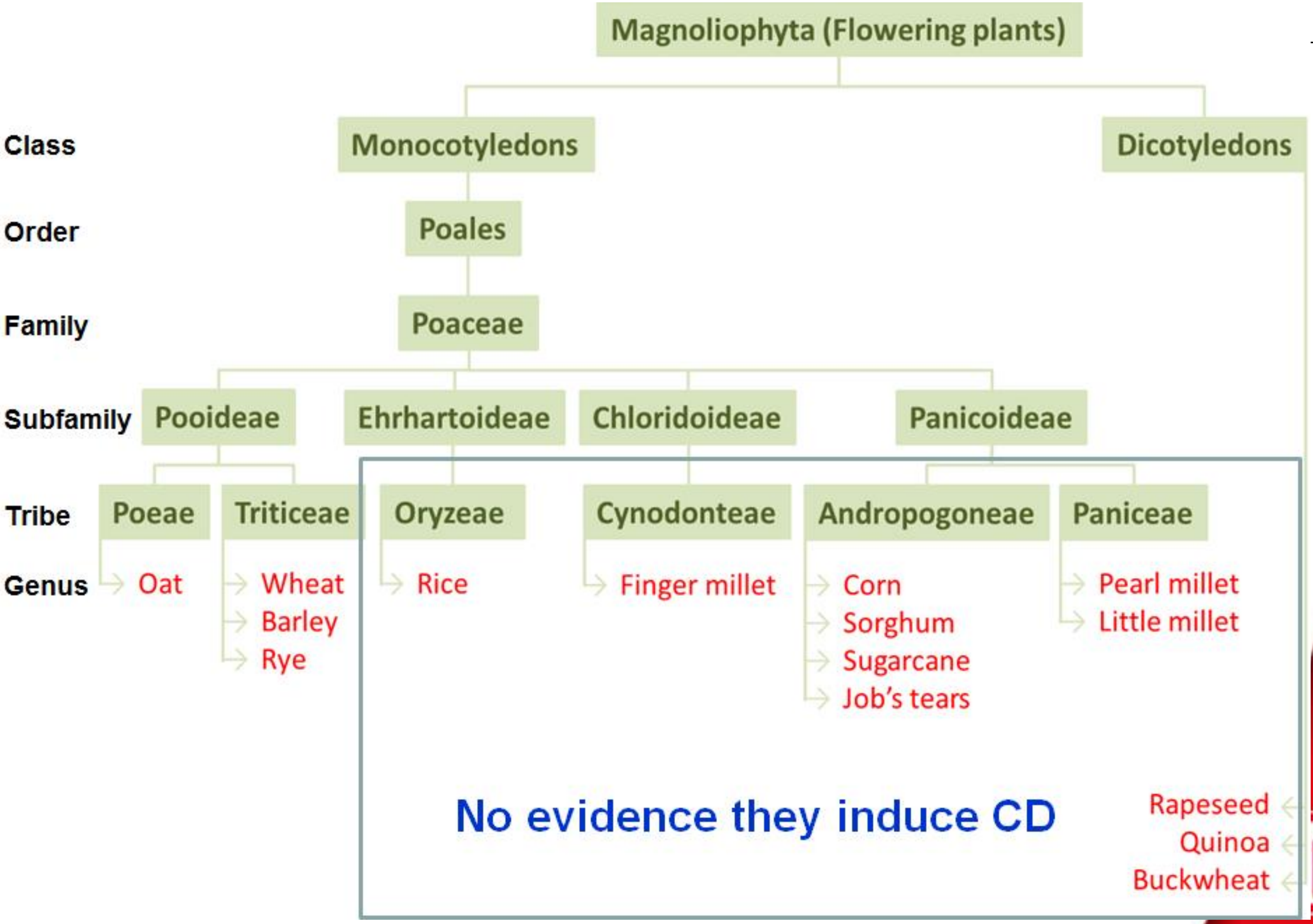
No exact peptide matches, if exact match, no risk..?

WOULD THIS PROTEIN POSE A RISK?

Tests of Allergenonline.org Celiac Database

With proteins from Pooideae and other taxonomic groups

Taxonomy of Plants: CD database test



No evidence they induce CD

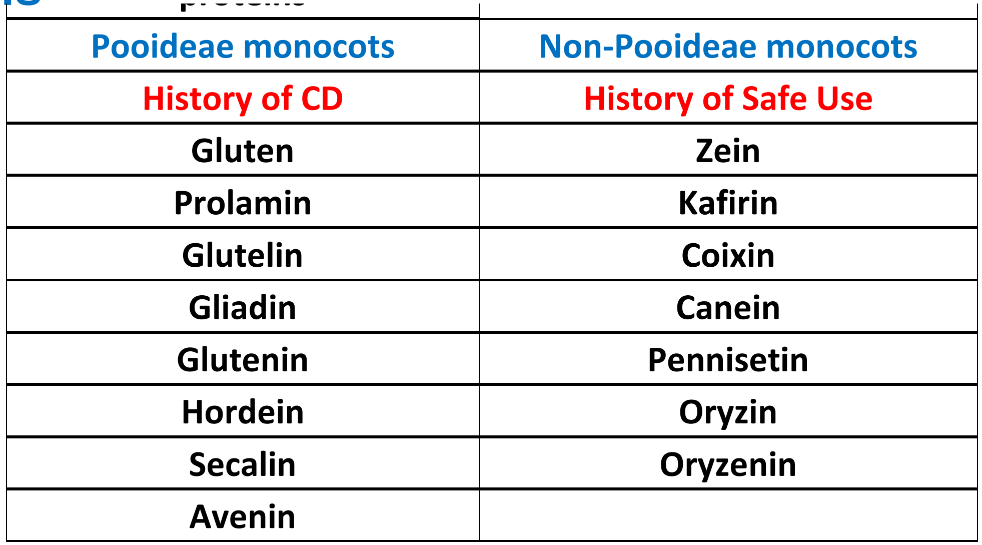
Tests for Exact peptide matches using randomly selected NCBI “gluten-like” proteins from taxonomic groups

- Proteins in Pooideae
  - – 2,666 from NCBI tested
  - – 2,104 exact matches to CD peptide
- Proteins from other monocots (non-Pooidea)
  - – 1,059 from NCBI tested
  - – 0 exact matches
- Proteins from dicotyledons
  - – 1,050 from NCBI tested
  - – 0 exact matches

Testing Efficiency of Protein Screening

1. Tested proteins from sources with evidence of CD
2. Tested proteins from sources without evidence of CD

2.1 Representative proteins from the NCBI database keyword inclusion by protein, exclusion by taxonomic group homologous proteins



Pooideae monocots	Non-Pooideae monocots
History of CD	History of Safe Use
Gluten	Zein
Prolamin	Kafirin
Glutelin	Coixin
Gliadin	Canein
Glutenin	Pennisetin
Hordein	Oryzin
Secalin	Oryzenin
Avenin	

2.2 Tested by BLASTP vs. NCBI with peptides in our CD

NO matches of 1,016 peptides in non-Pooideae monocots

Exact peptide Sequence Match Testing

Group		Number of Protein	Contain exact CD peptides
I	Prolamins in Pooideae	2,104	Yes
	Prolamins in Pooideae	562	No
II	Prolamins & prolamin-like proteins in other Monocots	1,059	No
III	Prolamin-like proteins in Dicots	1,050	No
IV	Unrelated proteins	48	No

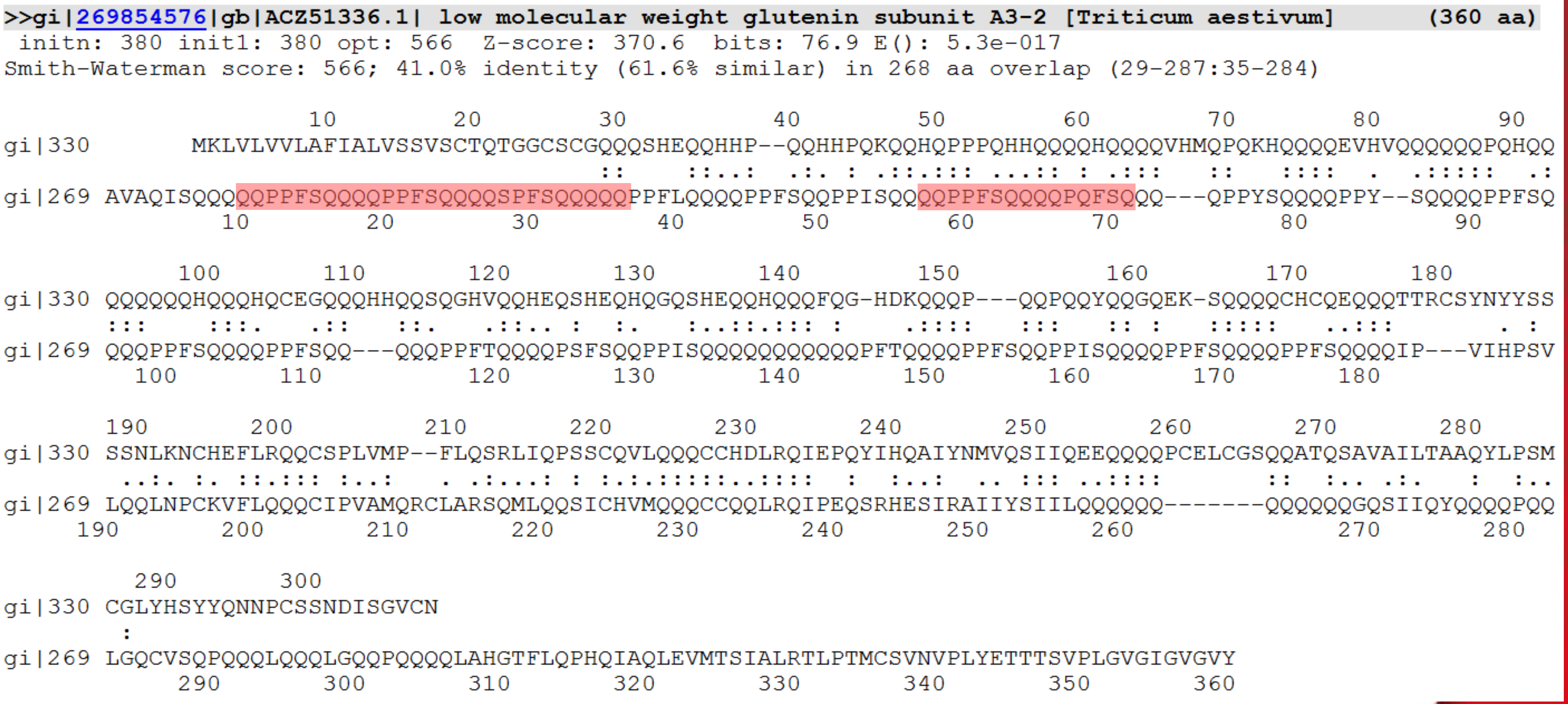
Examples of additional testing by exact match and FASTA unrelated proteins

- Some non-Pooideae query proteins were found to have “FASTA” alignments with the 68 CD proteins, but were NOT significant
- Short protein segment alignments of 20 to 29 aa
  - Yielded high percent identities & moderate to low E scores
  - NO epitope alignments

Group II – FASTA GI:330732090 unnamed protein [*Zea mays*]

41% identity over 268 AA, 5.3e-17

NO EXACT matches



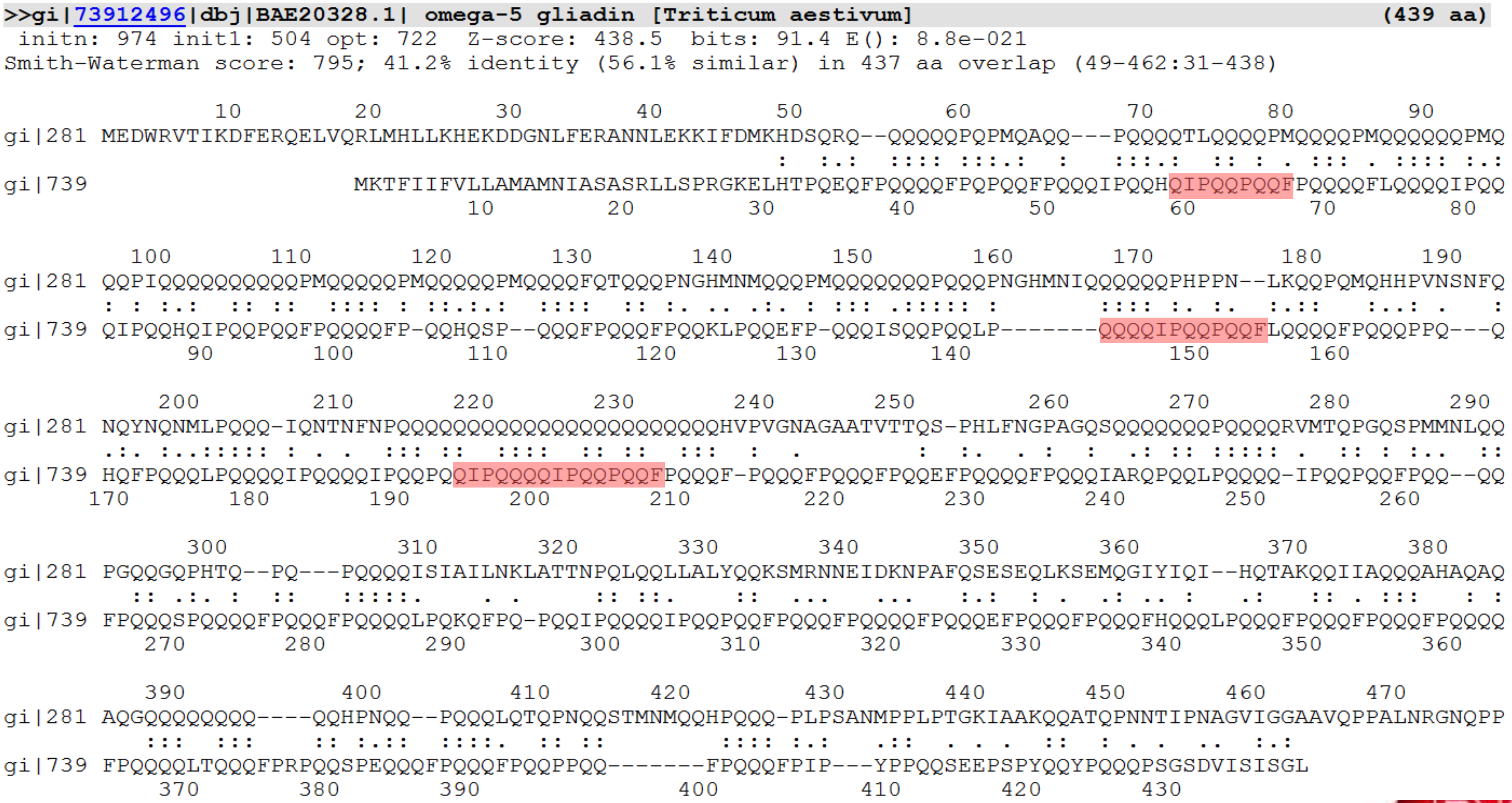
```
>>gi|269854576|gb|ACZ51336.1| low molecular weight glutenin subunit A3-2 [Triticum aestivum]       (360 aa)
 initn: 380 init1: 380 opt: 566  Z-score: 370.6  bits: 76.9 E(): 5.3e-017
Smith-Waterman score: 566; 41.0% identity (61.6% similar) in 268 aa overlap (29-287:35-284)

                     10        20        30          40        50        60        70        80        90
gi|330       MKLVLVVLAFIALVSSVSCTQTGGCSCGQQQSHEQQHHP--QQHHPQKQQHQPPPQHHQQQQHQQQQVHMQPQKHQQQQEVHVQQQQQQPQHQQ
                                         ::    ::..:  .:. : .::.::: ...:: : .:::   ::   ::::  .  .:::::  .:
gi|269 AVAQISQQQQQPPFSQQQQPPFSQQQQSPFSQQQQQPPFLQQQQPPFSQQPPISQQQQPPFSQQQQPQFSQQQ---QPPYSQQQQPPY--SQQQQPPFSQ
               10        20        30        40        50        60        70           80          90

            100       110       120       130       140        150          160        170       180
gi|330 QQQQQQHQQQHQCEGQQQHHQQSQGHVQQHEQSHEQHQGQSHEQQHQQQFQG-HDKQQQP---QQPQQYQQGQEK-SQQQQCHCQEQQQTTRCSYNYYSS
       :::    :::.   .::   ::.   .::.. :  :.   :..::.::: :    .::::   :::   :: :   :::::   ..:::       . :
gi|269 QQQPPFSQQQQPPFSQQ---QQQPPFTQQQQPSFSQQPPISQQQQQQQQQQQPFTQQQQPPFSQQPPISQQQQPPFSQQQQPPFSQQQQIP---VIHPSV
         100       110          120       130       140       150       160       170       180

       190       200         210       220       230       240       250       260       270       280
gi|330 SSNLKNCHEFLRQQCSPLVMP--FLQSRLIQPSSCQVLQQQCCHDLRQIEPQYIHQAIYNMVQSIIQEEQQQQPCELCGSQQATQSAVAILTAAQYLPSM
        ..:. :. ::.::: :..:   . .:...: : :.:.:::::..::::  :  :..:  .. ::: ..::::       ::  :.. .:.   :  :..
gi|269 LQQLNPCKVFLQQQCIPVAMQRCLARSQMLQQSICHVMQQQCCQQLRQIPEQSRHESIRAIIYSIILQQQQQQ-------QQQQQQGQSIIQYQQQQPQQ
     190       200       210       220       230       240       250       260              270       280

         290       300
gi|330 CGLYHSYYQNNPCSSNDISGVCN
        :
gi|269 LGQCVSQPQQQLQQQLGQQPQQQQLAHGTFLQPHQIAQLEVMTSIALRTLPTMCSVNVPLYETTTSVPLGVGIGVGVY
            290       300       310       320       330       340       350       360
```

🟥 = Region of CD inducing peptides in LMW glutenin of *T. aestivum*

Group IV – FASTA with GI:281206089 hypothetical protein PPL_07106 [*Polysphondylium pallidum* ; slime mold]

41% identical in 437 AA alignment, E score 0.8 e-21

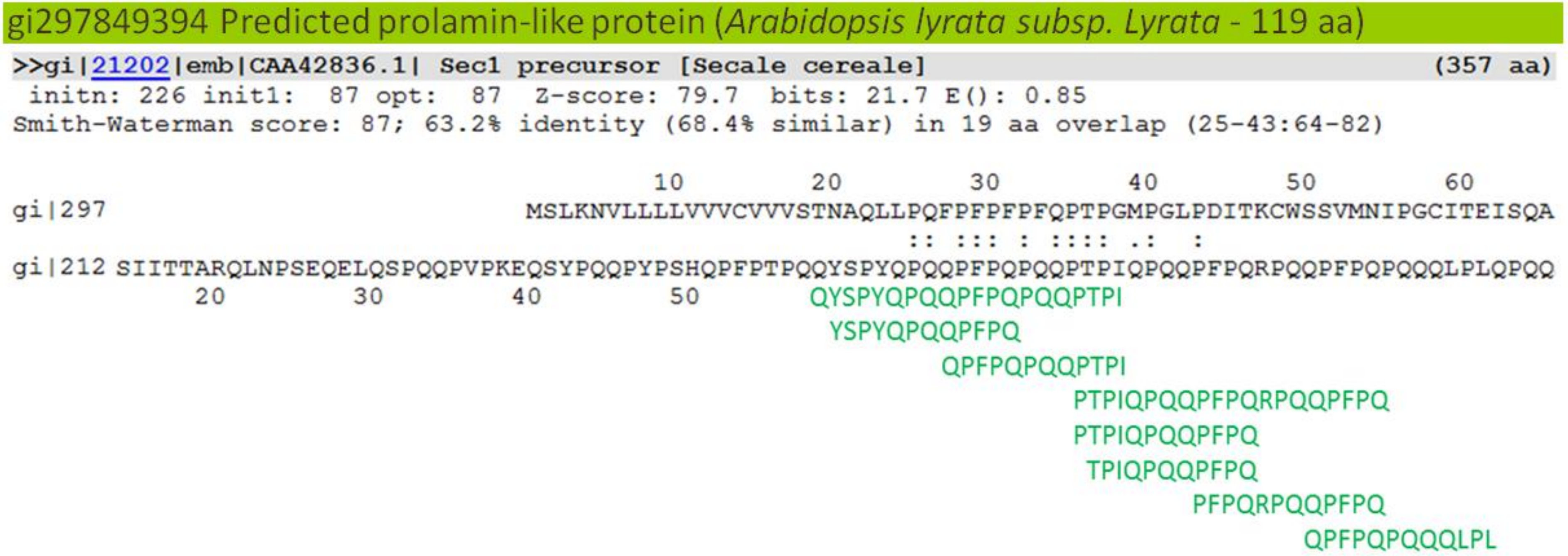
NO EXACT PEPTIDE MATCH



```
>>gi|73912496|dbj|BAE20328.1| omega-5 gliadin [Triticum aestivum]                                    (439 aa)
 initn: 974 init1: 504 opt: 722  Z-score: 438.5  bits: 91.4 E(): 8.8e-021
Smith-Waterman score: 795; 41.2% identity (56.1% similar) in 437 aa overlap (49-462:31-438)

               10        20        30        40        50          60           70        80        90
gi|281 MEDWRVTIKDFERQELVQRLMHLLKHEKDDGNLFERANNLEKKIFDMKHDSQRQ--QQQQQPQPMQAQQ---PQQQQTLQQQQPMQQQQPMQQQQQQPMQ
                                                       :  :.:  :::: :::.:  :   :::.:  :: : . ::: . :::: :.:
gi|739                   MKTFIIFVLLAMAMNIASASRLLSPRGKELHTPQEQFPQQQQFPQPQQFPQQQIPQQHQIPQQPQQFPQQQQFLQQQQIPQQ
                                 10        20        30        40        50        60        70        80

         100       110       120       130       140       150       160       170         180       190
gi|281 QQPIQQQQQQQQQQPMQQQQQPMQQQQQPMQQQQFQTQQQPNGHMNMQQQPMQQQQQQQPQQQPNGHMNIQQQQQQPHPPN--LKQQPQMQHHPVNSNFQ
       : : :.:  :: ::  :::: : ::.:.:  ::::  :: :. .. .:. : ::: .::::: :       :::: :. :.  :.::   :..: .   :
gi|739 QIPQQHQIPQQPQQFPQQQQFP-QQHQSP--QQQFPQQQFPQQKLPQQEFP-QQQISQQPQQLP-------QQQQIPQQPQQFLQQQQFPQQQPPQ---Q
             90       100        110         120       130        140              150       160

           200        210       220       230       240       250        260       270       280       290
gi|281 NQYNQNMLPQQQ-IQNTNFNPQQQQQQQQQQQQQQQQQQQQQQQHVPVGNAGAATVTTQS-PHLFNGPAGQSQQQQQQQPQQQQRVMTQPGQSPMMNLQQ
       .:. :..::::: : . .  ::: ::  ::::  :: ::  :::  :  .        :  :.  . :  :  .: :: ::::: .  :: : :..  ::
gi|739 HQFPQQQLPQQQQIPQQQQIPQQPQQIPQQQQIPQQPQQFPQQQF-PQQQFPQQQFPQQEFPQQQQFPQQQIARQPQQLPQQQQ-IPQQPQQFPQQ--QQ
      170       180       190       200       210        220       230       240       250        260

             300            310       320       330       340       350       360         370       380
gi|281 PGQQGQPHTQ--PQ---PQQQQISIAILNKLATTNPQLQQLLALYQQKSMRNNEIDKNPAFQSESEQLKSEMQGIYIQI--HQTAKQQIIAQQQAHAQAQ
         :: .:. :  ::   :::::.    . .     :: ::.    ::  ...   ...   :.:  : .  .: .. :   .:   :: . :::   : :
gi|739 FPQQQSPQQQQFPQQQFPQQQQLPQKQFPQ-PQQIPQQQQIPQQPQQFPQQQFPQQQQFPQQQEFPQQQFPQQQFHQQQLPQQQFPQQQFPQQQFPQQQQ
          270       280       290        300       310       320       330       340       350       360

          390           400         410       420        430       440       450       460       470
gi|281 AQGQQQQQQQQ----QQHPNQQ--PQQQLQTQPNQQSTMNMQQHPQQQ-PLPSANMPPLPTGKIAAKQQATQPNNTIPNAGVIGGAAVQPPALNRGNQPP
          :::  :::    :: :.::  ::::.  :: ::        :::: :.:   .::  . . .  ::  : . .  ..  :.:
gi|739 FPQQQQLTQQQFPRPQQSPEQQQFPQQQFPQQPPQQ-------FPQQQFPIP---YPPQQSEEPSPYQQYPQQQPSGSDVISISGL
           370       380       390              400          410       420       430
```

🟥 = Region of CD inducing peptides in ω-5 gliadin of *T. aestivum*

Group III – FASTA dicot protein homologue No exact match, showing alignment Secalin **GI: 212**



gi297849394 Predicted prolamin-like protein (*Arabidopsis lyrata subsp. Lyrata* - 119 aa)

```
>>gi|21202|emb|CAA42836.1| Sec1 precursor [Secale cereale]                    (357 aa)
 initn: 226 init1:  87 opt:  87  Z-score: 79.7  bits: 21.7 E(): 0.85
Smith-Waterman score: 87; 63.2% identity (68.4% similar) in 19 aa overlap (25-43:64-82)

                                         10        20        30        40        50        60
gi|297                           MSLKNVLLLLVVVCVVVSTNAQLLPQFPFPFPFQPTPGMPGLPDITKCWSSVMNIPGCITEISQA
                                                         :: ::: : :::: .:  :
gi|212 SIITTARQLNPSEQELQSPQQPVPKEQSYPQQPYPSHQPFPTPQQYSPYQPQQPFPQPQQPTPIQPQQPFPQRPQQPFPQPQQQLPLQPQQ
            20        30        40        50
```

QYSPYQPQQPFPQPQQPTPI

YSPYQPQQPFPQ

QPFPQPQQPTPI

PTPIQPQQPFPQRPQQPFPQ

PTPIQPQQPFPQ

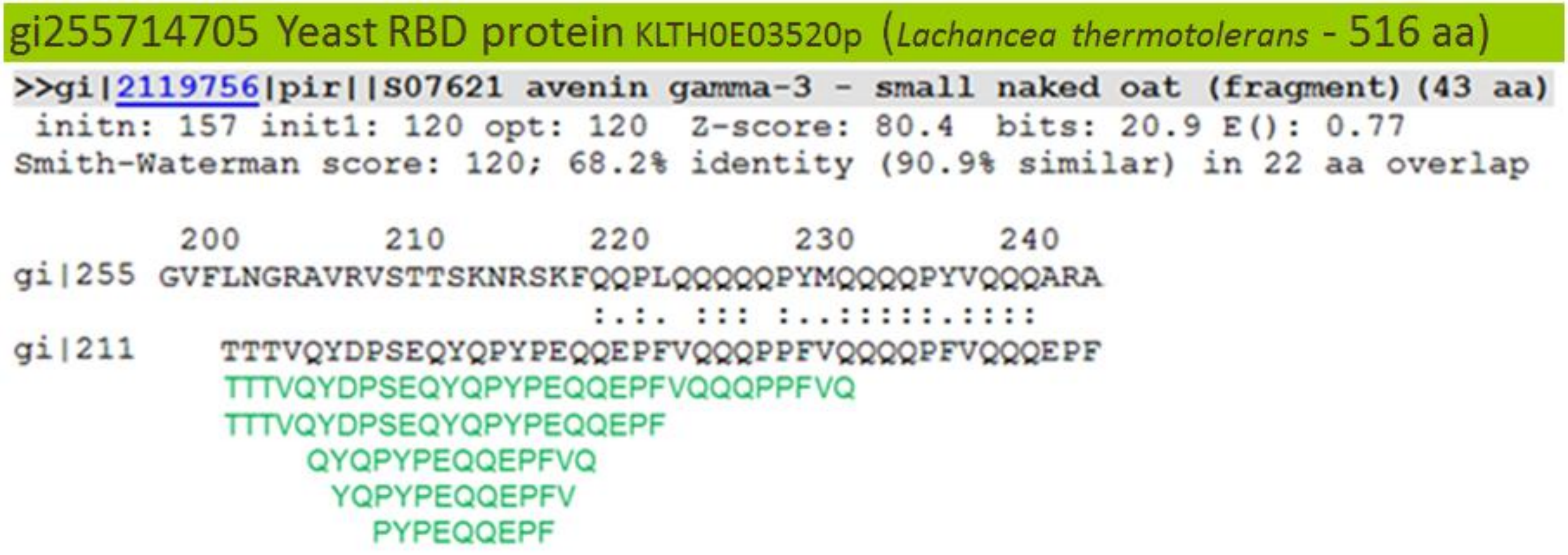
TPIQPQQPFPQ

PFPQRPQQPFPQ

QPFPQPQQQLPL

Arabidopsis (mustard) alignment with rye Sec 1 precursor does NOT have exact matches to the 8 CD inducing peptides

Group IV – FASTA with unrelated protein yeast 516 AA showing region of epitopes in Avenin (GI: 2119756)

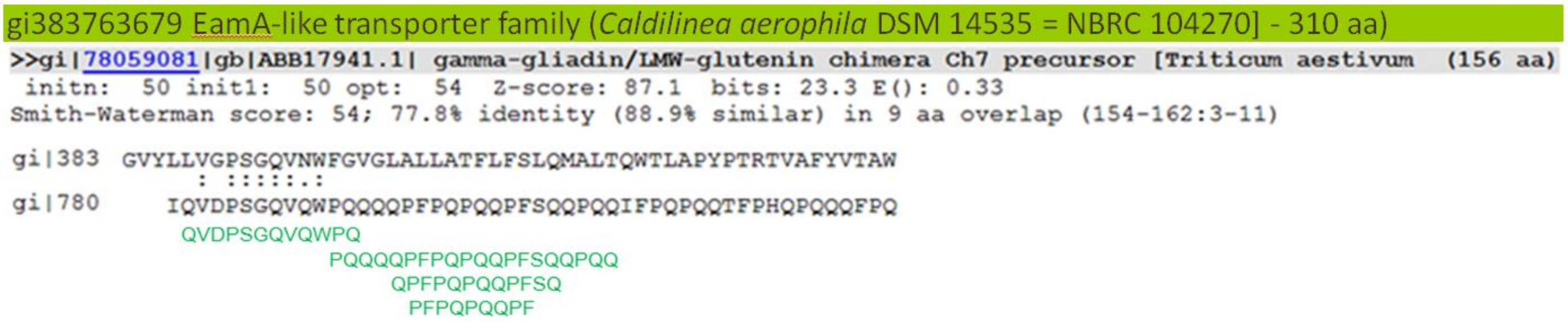


gi255714705 Yeast RBD protein KLTH0E03520p (*Lachancea thermotolerans* - 516 aa)

```
>>gi|2119756|pir||S07621 avenin gamma-3 - small naked oat (fragment) (43 aa)
 initn: 157 init1: 120 opt: 120  Z-score: 80.4  bits: 20.9 E(): 0.77
Smith-Waterman score: 120; 68.2% identity (90.9% similar) in 22 aa overlap

        200       210       220       230       240
gi|255 GVFLNGRAVRVSTTSKNRSKFQQPLQQQQQPYMQQQQPYVQQQARA
                            :.:. ::: :..:::::.::::
gi|211    TTTVQYDPSEQYQPYPEQQEPFVQQQPPFVQQQQPFVQQQEPF
          TTTVQYDPSEQYQPYPEQQEPFVQQQPPFVQ
          TTTVQYDPSEQYQPYPEQQEPF
              QYQPYPEQQEPFVQ
               YQPYPEQQEPFV
                 PYPEQQEPF
```

Yeast RNA binding protein FASTA alignment with 68% ID and E=0.77 over 22 AA alignment with Avenin γ3 adjacent to, but does not have exact matches with the 5 overlapping CD inducing peptides

Group IV – Unrelated Protein (bacteria) 310 AA alignment with LMW gamma gliadin GI:78059081

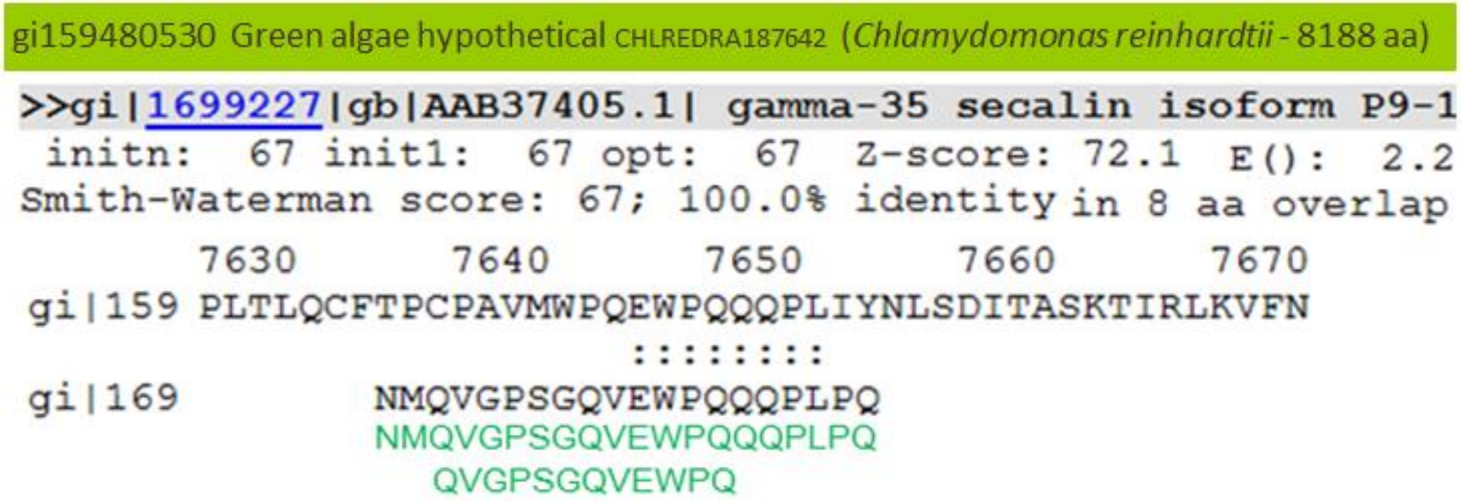


gi383763679 EamA-like transporter family (*Caldilinea aerophila* DSM 14535 = NBRC 104270] - 310 aa)

```
>>gi|78059081|gb|ABB17941.1| gamma-gliadin/LMW-glutenin chimera Ch7 precursor [Triticum aestivum  (156 aa)
 initn:  50 init1:  50 opt:  54  Z-score: 87.1  bits: 23.3 E(): 0.33
Smith-Waterman score: 54; 77.8% identity (88.9% similar) in 9 aa overlap (154-162:3-11)

gi|383 GVYLLVGPSGQVNWFGVGLALLATFLFSLQMALTQWTLAPYPTRTVAFYVTAW
            : :::::.:
gi|780    IQVDPSGQVQWPQQQQPFPQPQQPFSQQPQQIFPQPQQTFPHQPQQQFPQ
           QVDPSGQVQWPQ
                     PQQQQPFPQPQQPFSQQPQQ
                         QPFPQPQQPFSQ
                          PFPQPQQPF
```

Bacterial protein alignments with 77% ID, E=0.33m over 9 AA overlap to γ-gliadin/LMW-glutenin clearly no exact match to the 4 CD inducing peptides

Group IV – Unrelated Protein (Chlamydomonas sp. 8188 AA alignment with Secalin GI:169

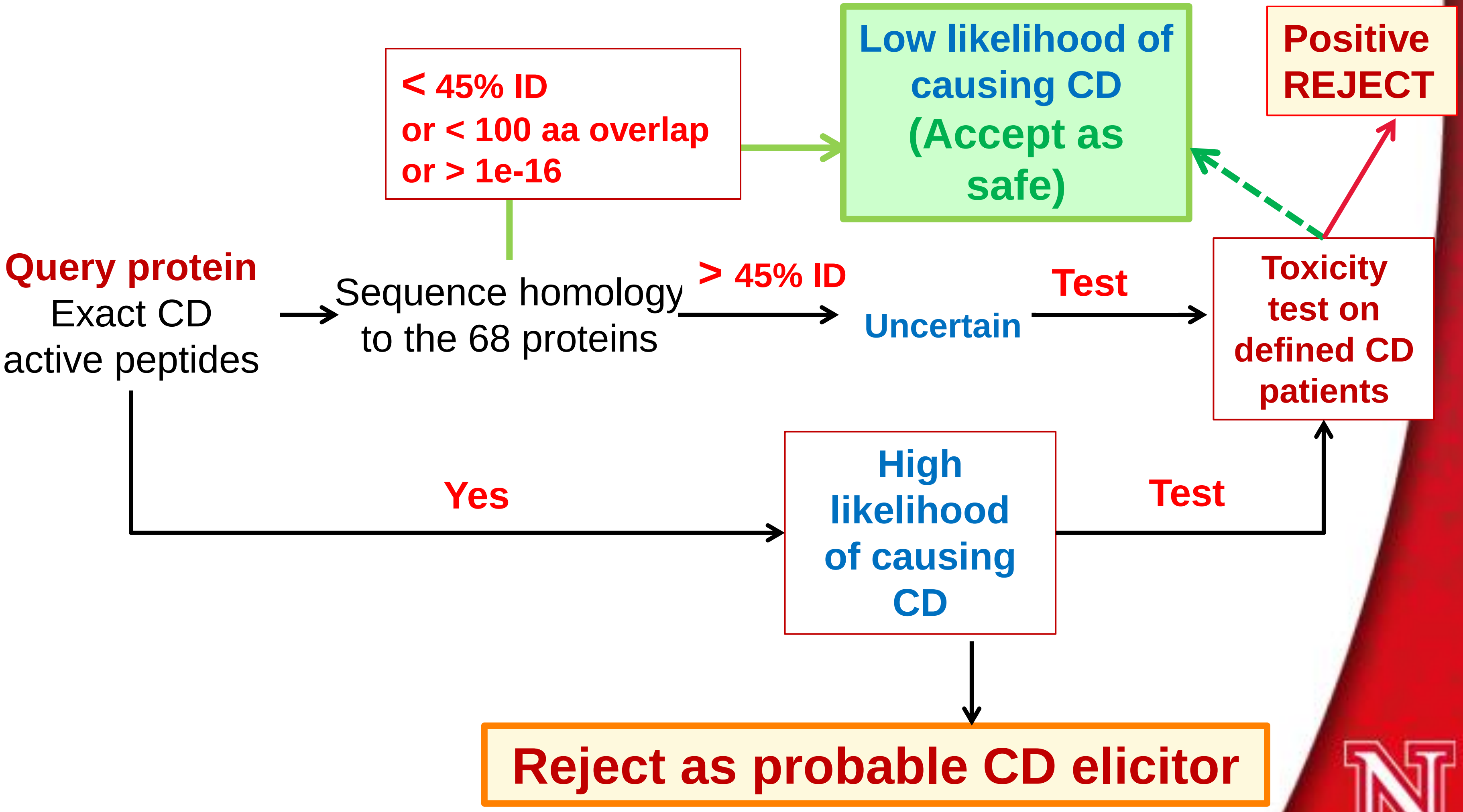


gi159480530 Green algae hypothetical CHLREDRA187642 (*Chlamydomonas reinhardtii* - 8188 aa)

```
>>gi|1699227|gb|AAB37405.1| gamma-35 secalin isoform P9-1
 initn:  67 init1:  67 opt:  67  Z-score: 72.1  E():  2.2
Smith-Waterman score: 67; 100.0% identity in 8 aa overlap

       7630      7640      7650      7660      7670
gi|159 PLTLQCFTPCPAVMWPQEWPQQQPLIYNLSDITASKTIRLKVFN
                        ::::::::
gi|169        NMQVGPSGQVEWPQQQPLPQ
              NMQVGPSGQVEWPQQQPLPQ
                QVGPSGQVEWPQ
```

Algae protein alignment with 100% ID, E = 2.2 to 8 AA segment of γ-35 secalin but clearly no exact-match to the 2 CD inducing peptides

Predicting the likelihood of a query protein to cause CD using AllergenOnline.org Celiac DB



Query protein

Exact CD active peptides

Sequence homology to the 68 proteins

< 45% ID
or < 100 aa overlap
or > 1e-16

Low likelihood of causing CD (Accept as safe)

> 45% ID

Uncertain

Test

Toxicity test on defined CD patients

Positive REJECT

Yes

High likelihood of causing CD

Test

Reject as probable CD elicitor

GM proteins from Pooideae that pass the bioinformatics evaluation and are transferred to a non-Pooideae crop should be as safe as these

- Rice
- Maize (corn)
- Sorghum (Jowar)
- Millet (Bajra)
- Amaranth
- Arrowroot
- Buckwheat
- Flax
- Oats (if pure), although some varieties??

- Potato
- Quinoa
- Tapioca
- Flours from nuts and beans

Acknowledgements

- Goodman lab
  - Plaimein Amnuaycheewa
  - John Wise
  - Afua Tetteh
- Allergenonline experts
  - Steve Taylor (FARRP)
  - Joe Baumert (FARRP)
  - Barbara Bohle
  - Fatima Ferreira
- Authors of many CD papers
- Comments from
  - Bana Jabri – Chicago
  - Frits Koenig – Leiden
- Database sponsors
  - BASF
  - Bayer
  - Dow
  - Monsanto
  - Pioneer / DuPont
  - Syngenta
  - KWS Seeds
  - LimaGrain