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
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.

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 segments 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
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

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
Grasses as a single genetic system
Evolutionary history of cereals – AND GLUTEN COMPLEX Divergence


Copyright © 2007 by the Genetics Society of America
Probable Evolution of modern wheat
D. Kasarda, JAFC 2013 61:1155-1159

Gluten genetics are complex, gene duplication + divergence

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 resistant peptide

Tissue transglutaminase

PQPQLPYPQ

Inflammatory Reactions

(Adadie 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

(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 proinflammatory 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
QLQPFPQPQLPY - 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)

<table>
<thead>
<tr>
<th>Test (clone 6)</th>
<th>CPM</th>
<th>SI</th>
<th>IFN-γ</th>
<th>IL-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T+APC only</td>
<td>1053 (911)</td>
<td>—</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>PQPELPYPQPQLPY</td>
<td>14764 (3870)</td>
<td>14</td>
<td>232</td>
<td>0</td>
</tr>
<tr>
<td>PQAPELPYPQPQLPY</td>
<td>343 (110)</td>
<td>&lt;1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PQPEAPYPQPQLPY</td>
<td>244 (59)</td>
<td>&lt;1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PQPELAYYPQPQLPY</td>
<td>182 (42)</td>
<td>&lt;1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PQPELPAPQPQLPY</td>
<td>236 (25)</td>
<td>&lt;1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PQPELPYAQPQLPY</td>
<td>211 (33)</td>
<td>&lt;1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PQPELPYAPAQLPY</td>
<td>798 (123)</td>
<td>&lt;1</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>PQPELPYPAQLPY</td>
<td>341 (70)</td>
<td>&lt;1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PQPELPYQPAPLY</td>
<td>14454 (197)</td>
<td>14</td>
<td>205</td>
<td>0</td>
</tr>
<tr>
<td>PQPELPYQPQAPY</td>
<td>13681 (1209)</td>
<td>13</td>
<td>225</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test (clone 8)</th>
<th>CPM</th>
<th>SI</th>
<th>IFN-γ</th>
<th>IL-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T+APC</td>
<td>4238 (418)</td>
<td>—</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>PQPELPYPQPQLPY</td>
<td>51683 (913)</td>
<td>12</td>
<td>575</td>
<td>130</td>
</tr>
<tr>
<td>PQPELPYPQPQLAY</td>
<td>54782 (771)</td>
<td>13</td>
<td>600</td>
<td>nd</td>
</tr>
<tr>
<td>PQPELPYPQPQLPA</td>
<td>57467 (1999)</td>
<td>14</td>
<td>600</td>
<td>nd</td>
</tr>
</tbody>
</table>
48 unique exact peptide matches

GI7209265 α-gliadin

[Triticum aestivum] 290 aa
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
1,016 peptides → Exact sequence matching (MySQL)
68 proteins → 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 farro allergen protein database

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

Celiac Tools
Browse Entries
By Peptides
By References
By Proteins
Sequence Search
Exact peptide match
Full FASTA

Celiac disease database also includes a FASTA reducing proteins that are the sources of the inclusion of peptides and proteins in
### Peptide Exact match
**alpha-gliadin**
Triticum spelta var. arduini
GI:3928509

#### Your Search returned 33 results

<table>
<thead>
<tr>
<th>ID</th>
<th>Type</th>
<th>Description</th>
<th>Toxicity</th>
<th>Form</th>
<th>Refs</th>
<th>Sequence</th>
<th>HLADQ</th>
<th>SeqLen</th>
<th># of Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>alpha-gliadin</td>
<td>alpha-gliadin CT-1</td>
<td>Toxic</td>
<td>Native</td>
<td>41</td>
<td>VPVPQLQPQNSQQPQQEQQVPL</td>
<td>Unknown</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>alpha-gliadin</td>
<td>alpha-gliadin p14</td>
<td>Immunogenic</td>
<td>Native</td>
<td>34</td>
<td>VRVPVQLQPQNSQQPQQPQQ</td>
<td>DQ2</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>alpha-gliadin</td>
<td>alpha-gliadin p15</td>
<td>Immunogenic</td>
<td>Native</td>
<td>34</td>
<td>QNPSQQQPGEOVPLQQQ</td>
<td>DQ2</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>alpha-gliadin</td>
<td>alpha-gliadin p62</td>
<td>Immunogenic</td>
<td>Native</td>
<td>34</td>
<td>QFYQPQIPPSQFFYLQL</td>
<td>DQ2</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>alpha-gliadin</td>
<td>alpha-gliadin p44</td>
<td>Immunogenic</td>
<td>Native</td>
<td>10</td>
<td>PQQPQIPPSQQPY</td>
<td>HLA-B*</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>160</td>
<td>alpha-gliadin</td>
<td>DQ2-Gli-a alpha epitope (p58-p72, 569)</td>
<td>Immunogenic</td>
<td>Native</td>
<td>59</td>
<td>LPFPQPLPYSQPOQ</td>
<td>DQ2</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>164</td>
<td>alpha-gliadin</td>
<td>Wheat peptide W08</td>
<td>Immunogenic</td>
<td>Native</td>
<td>62</td>
<td>QFPFPQQLPYSQ</td>
<td>DQ2</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>166</td>
<td>alpha-gliadin</td>
<td>Gli-a alpha</td>
<td>Immunogenic</td>
<td>Native</td>
<td>59</td>
<td>PFPQQLPYSQ</td>
<td>DQ2</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>168</td>
<td>alpha-gliadin</td>
<td>alpha-gliadin (p202-p220)</td>
<td>Toxic</td>
<td>Native</td>
<td>11</td>
<td>QYTPLQGQSFQPSQNPQA</td>
<td>DQ2</td>
<td>19</td>
<td>1</td>
</tr>
</tbody>
</table>
Search FASTA with alpha gliadin from *Triticum spelta*
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

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
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

<table>
<thead>
<tr>
<th>Pooideae monocots</th>
<th>Non-Pooideae monocots</th>
</tr>
</thead>
<tbody>
<tr>
<td>History of CD</td>
<td>History of Safe Use</td>
</tr>
<tr>
<td>Gluten</td>
<td>Zein</td>
</tr>
<tr>
<td>Prolamin</td>
<td>Kafirin</td>
</tr>
<tr>
<td>Glutelin</td>
<td>Coixin</td>
</tr>
<tr>
<td>Gliadin</td>
<td>Canein</td>
</tr>
<tr>
<td>Glutenin</td>
<td>Pennisetin</td>
</tr>
<tr>
<td>Hordein</td>
<td>Oryzin</td>
</tr>
<tr>
<td>Secalin</td>
<td>Oryzenin</td>
</tr>
<tr>
<td>Avenin</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Protein</th>
<th>Contain exact CD peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Prolamins in Pooideae</td>
<td>2,104</td>
<td>Yes</td>
</tr>
<tr>
<td>Prolamins in Pooideae</td>
<td>562</td>
<td>No</td>
</tr>
<tr>
<td>II Prolamins &amp; prolamin-like proteins in other Monocots</td>
<td>1,059</td>
<td>No</td>
</tr>
<tr>
<td>III Prolamin-like proteins in Dicots</td>
<td>1,050</td>
<td>No</td>
</tr>
<tr>
<td>IV Unrelated proteins</td>
<td>48</td>
<td>No</td>
</tr>
</tbody>
</table>
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

= 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**

= 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**

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)

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

Bacterial protein alignments with 77% ID, E=0.33 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)

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)**
      - **Test Positive**
        - **Reject as probable CD elicitor**
      - **Uncertain**
        - **Test**
          - **Toxicity test on defined CD patients**

- **> 45% ID**
  - **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