Allergenicity Prediction using Sequence Profiles

M.B. Stadler
University of Bern
Overview

• **Allergy**
  – What is an allergen/allergic reaction?

• **Generalized profiles**
  – Modeling sequence motifs
  – Construction of a profile

• **Allergenicity prediction**
  – Current prediction algorithm (FAO/WHO)
  – Profile-based prediction
Overview

• **Allergy**
  – What is an allergen/allergic reaction?

• **Generalized profiles**
  – Modeling sequence motifs
  – Construction of a profile

• **Allergenicity prediction**
  – Current prediction algorithm (FAO/WHO)
  – Profile-based prediction
Allergens are...

- harmless substances
- (almost) only proteins
- inducing IgE immune response
- rare: \( \sim 800'000 \) proteins (Sp/TrEMBL)
  - \( \sim 800 \) sequences
  - \( < 100 \) allergens

1 in 100’000 proteins:
Common Allergens

house dust mites, pollen & spores, pets, insects, milk, eggs, nuts
Immediate Hypersensitivity Reaction

Sensitize

B-cell

allergen

IgE

mast cell

Trigger

cross-link

mediator release
Response to Mediators

**Mediators:**
- Histamine
- Leukotrienes
- Cytokines
- Chemokines
- Enzymes

**Bronchi:**
Constriction

**Blood vessels:**
Vascular leakage (tissue edema)
<table>
<thead>
<tr>
<th>Reaction</th>
<th>Site</th>
<th>Signs/Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhinitis</td>
<td>Nose</td>
<td>Sneezing, rhinorrhea, nasal itching, congestion</td>
</tr>
<tr>
<td>Asthma</td>
<td>Lungs</td>
<td>Coughing, wheezing, shortness of breath</td>
</tr>
<tr>
<td>Dermatitis</td>
<td>Skin</td>
<td>Itching, rash</td>
</tr>
<tr>
<td>Conjunctivitis</td>
<td>Eye</td>
<td>Itching, redness, tearing</td>
</tr>
<tr>
<td>Anaphylaxis</td>
<td>Systemic</td>
<td>Hypotension, shock, death</td>
</tr>
<tr>
<td>Food</td>
<td>Gut</td>
<td>Bloating, vomiting, diarrhea, cramping</td>
</tr>
</tbody>
</table>
Overview

- **Allergy**
  - What is an allergen/allergic reaction?

- **Generalized profiles**
  - Modeling sequence motifs
  - Construction of a profile

- **Allergenicity prediction**
  - Current prediction algorithm (FAO/WHO)
  - Profile-based prediction
Shared Domain Topologies

EGF

IL-1R

FcεRIα

IgG

MHC I

β2m

TcR

CD4

ICAM-2

SH2
Sequence-Structure Relationship

Domain (Fold):

Sequences:

Sequence Motif:
A Generalized Profile

• position-specific match-, gap- and insert-scores
• score distribution over sequence space
• parameters:

<table>
<thead>
<tr>
<th>begin/end:</th>
<th>insertion/match:</th>
<th>state transitions:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Be</td>
<td>I(X)</td>
<td>T_B→M</td>
</tr>
<tr>
<td>Bi</td>
<td>I(*)</td>
<td>T_B→I</td>
</tr>
<tr>
<td>Ee</td>
<td>M(X)</td>
<td>T_B→D</td>
</tr>
<tr>
<td>Ei</td>
<td>M(*)</td>
<td>T_B→E</td>
</tr>
<tr>
<td>deletion:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>T_I→M</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T_I→I</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T_I→D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T_I→E</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T_D→M</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T_D→I</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T_D→D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T_D→E</td>
</tr>
</tbody>
</table>
Manual Construction of Sequence Profiles

- Trusted sequences
- Add new sequences
- Multiple alignment
- Sequence weights
- Create profile
- Calibrate profile
- New matches?
- External HMM training
- New profile
Clinical cross-reactions include:
birch, celery, carrot, apple, hornbeam, hazelnut, alder, apricot, cherry, pear
Creating a Bet v 1-Profile

extensive literature search:
13 allergens:
• 6 structures
• 94 sequences
Good and Bad Sequence Alignments

Sequence based alignment

Structure based alignment

ASTQSPSVFPLTRCCK
NI PSNATSVTLGCLAT
GYFPEPVMVTWDTGSL
NGTTMTPATTLTLSSG
HYATISLLTVSCAWAK
QMFTCRVAHTPSSTDW
VDNKTSVCSRDFTPP
TVKILQSSCDGGHFPP
PTIQLLCLVSGYTPGT
INITWLEDGQVMDVDL
Creating a Bet v 1-Profile: Structural Alignment

trusted sequences → multiple alignment

6 structures → 94 sequences

T_COFFEE
structure based alignment

guide-profile

profile based alignment

PFTOOLS

ASTQSPSVFPLTRCCK
NIPSNATSVTGLCLAT
GYFPEPVMVTDTGL
NGTMTLPATTTLTSLG
EMPEWNHTPSLCK
VDKTFVCSRDFTPP
TVKILQSSCDGGHFP
PTIQLCLVSGYTPGT
INITWLEDGQVMVDL
Sequence Weights

Profile should be:

<table>
<thead>
<tr>
<th></th>
<th>focused</th>
<th>unbiased</th>
</tr>
</thead>
<tbody>
<tr>
<td>outliers</td>
<td>– –</td>
<td>+ +</td>
</tr>
<tr>
<td>similar sequences</td>
<td>+ +</td>
<td>– –</td>
</tr>
</tbody>
</table>
Sequence Weights: Tree Derived

$U_0$

$I_1 + I_2$

$I_1$

$1$

$I_2$

$2$

$I_3 + I_4 + I_5$

$3$

$I_3$

$4$

$I_4$

$5$

$I_5$

$I_6 + I_7 + I_8 + I_9 + I_{10}$

$6$

$I_6$

$7$

$I_7$

$8$

$I_8$

$9$

$I_9$

$10$

$R \sim \text{branch length}$
Sequence Weights: ‘Voronoi’
Empiric Profile Calibration

- **create profile**
- **calibrate profile:**
  - search randomized sequence database
  - fit EVD to scores
  - use EVD-parameters to normalize scores

Swiss-Prot, window shuffled

\[ p(x) = \lambda e^{-\lambda(x-\mu)} \approx e^{-\lambda(x-\mu)} \]

\[
S_{\text{norm}} = R_1 + R_2x
\]

\[
E(x, A) = A \cdot 10^{-S_{\text{norm}}}
\]

- \(x\): raw score
- \(A\): no. of residues
- \(N\): no. of sequences

\[ R_1 = \frac{\ln A}{\ln 10} - \lambda \mu \]

\[ R_2 = \frac{\lambda}{\ln 10} \]
Normalized Score and E-value

\[ E(x, A) = A \cdot 10^{-R_1 - R_2 x} \]

**Swiss-Prot 40:**
\[ A = 37'315'215 \]
\[ x = 814 \]
\[ S_{\text{norm}} = 6.5 \]
\[ E\text{-value} = 10 \]
Profile Construction is an Iterative Process

1. Trusted sequences
2. Multiple alignment
3. Sequence weights
4. Create profile
5. Calibrate profile
6. New matches?
7. External HMM training
8. New profile

Add new sequences if new matches are found.

The process starts with trusted sequences and iteratively adds new sequences, aligns them, weights them, creates a profile, calibrates the profile, checks for new matches, and trains the HMM based on these matches.
## FHA Domain Profile: PS50006

Hofmann & Bucher, 1995, TIBS

<table>
<thead>
<tr>
<th>Database Entry</th>
<th>Profile 1 weight</th>
<th>Profile 1 score</th>
<th>Profile 2 weight</th>
<th>Profile 2 score</th>
<th>Profile 3 weight</th>
<th>Profile 3 score</th>
<th>Profile 4 weight</th>
<th>Profile 4 score</th>
<th>Profile 5 weight</th>
<th>Profile 5 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS1_SCHPO</td>
<td>2.3</td>
<td>2.9</td>
<td>2.7</td>
<td>3.9</td>
<td>14.7</td>
<td>31.9</td>
<td>12.7</td>
<td>24.7</td>
<td>23.8</td>
<td></td>
</tr>
<tr>
<td>FHL1_YEAST</td>
<td>14.5</td>
<td>50.0</td>
<td>40.6</td>
<td>12.7</td>
<td>14.7</td>
<td>31.9</td>
<td>12.7</td>
<td>24.7</td>
<td>23.8</td>
<td></td>
</tr>
<tr>
<td>FKH1_YEAST</td>
<td>7.9</td>
<td>19.0</td>
<td>23.9</td>
<td>35.6</td>
<td>13.5</td>
<td>34.1</td>
<td>12.1</td>
<td>24.7</td>
<td>24.4</td>
<td></td>
</tr>
<tr>
<td>FKH2_YEAST</td>
<td>6.3</td>
<td>18.6</td>
<td>23.2</td>
<td>35.5</td>
<td>13.3</td>
<td>33.8</td>
<td>11.3</td>
<td>24.0</td>
<td>24.8</td>
<td></td>
</tr>
<tr>
<td>FRAH_ANASP</td>
<td>3.4</td>
<td>5.7</td>
<td>5.1</td>
<td>6.6</td>
<td>20.5</td>
<td>15.3</td>
<td>27.7</td>
<td>18.4</td>
<td>20.8</td>
<td></td>
</tr>
<tr>
<td>KAPP_ARATH</td>
<td>2.8</td>
<td>5.1</td>
<td>5.3</td>
<td>5.7</td>
<td>9.7</td>
<td>8.9</td>
<td>6.5</td>
<td>5.7</td>
<td>5.7</td>
<td></td>
</tr>
<tr>
<td>KI67_HUMAN</td>
<td>3.1</td>
<td>2.7</td>
<td>3.6</td>
<td>4.0</td>
<td>7.8</td>
<td>7.7</td>
<td>6.5</td>
<td>5.7</td>
<td>5.7</td>
<td></td>
</tr>
<tr>
<td>MNF_MOUSE</td>
<td>100.0</td>
<td>49.1</td>
<td>50.0</td>
<td>37.4</td>
<td>29.4</td>
<td>28.2</td>
<td>18.4</td>
<td>27.7</td>
<td>15.6</td>
<td>20.8</td>
</tr>
<tr>
<td>SPKI_YEAST</td>
<td>5.0</td>
<td>6.6</td>
<td>6.3</td>
<td>6.9</td>
<td>4.8</td>
<td>4.7</td>
<td>5.8</td>
<td>5.1</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>YHR5_YEAST</td>
<td>2.5</td>
<td>4.4</td>
<td>5.1</td>
<td>7.1</td>
<td>3.6</td>
<td>3.1</td>
<td>3.4</td>
<td>3.8</td>
<td>4.8</td>
<td></td>
</tr>
<tr>
<td>YKI5_CAEEL</td>
<td>3.6</td>
<td>6.4</td>
<td>5.8</td>
<td>9.0</td>
<td>4.8</td>
<td>4.7</td>
<td>5.5</td>
<td>8.6</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td>HUMKIAA10_1</td>
<td>3.1</td>
<td>3.1</td>
<td>3.4</td>
<td>3.8</td>
<td>4.8</td>
<td>4.7</td>
<td>5.5</td>
<td>8.6</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td>MLB1770_16</td>
<td>4.8</td>
<td>4.7</td>
<td>5.8</td>
<td>5.1</td>
<td>4.8</td>
<td>4.7</td>
<td>5.5</td>
<td>8.6</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td>SC9346_10</td>
<td>4.1</td>
<td>6.8</td>
<td>4.4</td>
<td>6.3</td>
<td>2.7</td>
<td>3.6</td>
<td>5.5</td>
<td>8.6</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td>SCCXIV38K_16</td>
<td>2.7</td>
<td>3.6</td>
<td>3.1</td>
<td>2.9</td>
<td>4.2</td>
<td>6.5</td>
<td>8.7</td>
<td>11.8</td>
<td>14.3</td>
<td>18.6</td>
</tr>
<tr>
<td>SCD9717_7</td>
<td>3.1</td>
<td>3.1</td>
<td>3.3</td>
<td>2.9</td>
<td>3.1</td>
<td>3.1</td>
<td>3.3</td>
<td>2.9</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>SCPR1GEN_4</td>
<td>3.1</td>
<td>3.1</td>
<td>3.3</td>
<td>2.9</td>
<td>3.1</td>
<td>3.1</td>
<td>3.3</td>
<td>2.9</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>SPAC17G8_10</td>
<td>6.0</td>
<td>8.5</td>
<td>10.2</td>
<td>19.2</td>
<td>6.0</td>
<td>8.5</td>
<td>10.2</td>
<td>19.2</td>
<td>18.5</td>
<td>17.1</td>
</tr>
<tr>
<td>SYCSLRG_6</td>
<td>4.0</td>
<td>5.5</td>
<td>4.8</td>
<td>6.4</td>
<td>4.0</td>
<td>5.5</td>
<td>4.8</td>
<td>6.4</td>
<td>6.9</td>
<td></td>
</tr>
<tr>
<td>SYCSLRG_63</td>
<td>8.9</td>
<td>6.2</td>
<td>6.5</td>
<td>8.7</td>
<td>8.9</td>
<td>6.2</td>
<td>6.5</td>
<td>8.7</td>
<td>9.7</td>
<td></td>
</tr>
<tr>
<td>YSCL8083_15</td>
<td>3.7</td>
<td>6.5</td>
<td>5.0</td>
<td>6.4</td>
<td>3.7</td>
<td>6.5</td>
<td>5.0</td>
<td>6.4</td>
<td>6.8</td>
<td></td>
</tr>
<tr>
<td>YSCL9470_15</td>
<td>8.2</td>
<td>8.9</td>
<td>9.7</td>
<td>18.9</td>
<td>8.2</td>
<td>8.9</td>
<td>9.7</td>
<td>18.9</td>
<td>15.3</td>
<td>20.5</td>
</tr>
<tr>
<td>B61188</td>
<td>2.8</td>
<td>3.5</td>
<td>2.9</td>
<td>3.6</td>
<td>2.8</td>
<td>3.5</td>
<td>2.9</td>
<td>3.6</td>
<td>3.8</td>
<td></td>
</tr>
</tbody>
</table>

**Highest false positive**

<p>|          | 8.1 | 8.2 | 8.7 | 7.7 | 7.8 |</p>
<table>
<thead>
<tr>
<th>Database Entry</th>
<th>Profile 6 weight</th>
<th>Profile 6 score</th>
<th>Profile 7 weight</th>
<th>Profile 7 score</th>
<th>Profile 8 weight</th>
<th>Profile 8 score</th>
<th>Profile 9 weight</th>
<th>Profile 9 score</th>
<th>Profile 10 weight</th>
<th>Profile 10 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS1_SCHPO</td>
<td>5.5</td>
<td></td>
<td>7.0</td>
<td></td>
<td>7.3</td>
<td></td>
<td>7.9</td>
<td></td>
<td>7.8</td>
<td></td>
</tr>
<tr>
<td>FHL1_YEAST</td>
<td>8.5</td>
<td>20.4</td>
<td>6.5</td>
<td>17.5</td>
<td>6.6</td>
<td>17.9</td>
<td>6.2</td>
<td>15.9</td>
<td>6.0</td>
<td>16.4</td>
</tr>
<tr>
<td>FKH1_YEAST</td>
<td>9.6</td>
<td>23.7</td>
<td>7.1</td>
<td>19.2</td>
<td>6.3</td>
<td>18.7</td>
<td>5.9</td>
<td>16.6</td>
<td>6.0</td>
<td>16.7</td>
</tr>
<tr>
<td>FKH2_YEAST</td>
<td>8.8</td>
<td>22.7</td>
<td>6.6</td>
<td>18.4</td>
<td>6.2</td>
<td>18.3</td>
<td>5.7</td>
<td>15.9</td>
<td>5.6</td>
<td>16.2</td>
</tr>
<tr>
<td>FRAH_ANASP</td>
<td>9.6</td>
<td>6.3</td>
<td>16.0</td>
<td></td>
<td>6.0</td>
<td>17.1</td>
<td>5.7</td>
<td>15.4</td>
<td>5.6</td>
<td>15.8</td>
</tr>
<tr>
<td>KAPP_ARATH</td>
<td>5.3</td>
<td></td>
<td>7.2</td>
<td></td>
<td>8.2</td>
<td></td>
<td>10.6</td>
<td></td>
<td>6.0</td>
<td>12.9</td>
</tr>
<tr>
<td>HUM7_HUMAN</td>
<td>3.7</td>
<td></td>
<td>5.0</td>
<td></td>
<td>5.4</td>
<td></td>
<td>5.8</td>
<td></td>
<td>5.7</td>
<td></td>
</tr>
<tr>
<td>MNF_MOUSE</td>
<td>10.1</td>
<td>18.1</td>
<td>7.4</td>
<td>15.2</td>
<td>7.6</td>
<td>15.6</td>
<td>7.0</td>
<td>13.8</td>
<td>6.8</td>
<td>13.8</td>
</tr>
<tr>
<td>SPKI_YEAST</td>
<td>6.2</td>
<td></td>
<td>6.7</td>
<td></td>
<td>7.0</td>
<td></td>
<td>8.1</td>
<td></td>
<td>7.8</td>
<td></td>
</tr>
<tr>
<td>YHR5_YEAST</td>
<td>9.2</td>
<td>21.1</td>
<td>6.8</td>
<td>19.2</td>
<td>6.1</td>
<td>19.6</td>
<td>5.6</td>
<td>17.2</td>
<td>5.7</td>
<td>17.8</td>
</tr>
<tr>
<td>YKI5_CAEL</td>
<td>8.5</td>
<td></td>
<td>10.3</td>
<td></td>
<td>8.4</td>
<td>15.0</td>
<td>7.8</td>
<td>13.6</td>
<td>7.6</td>
<td>14.0</td>
</tr>
<tr>
<td>HUMKIAA10_1</td>
<td>5.4</td>
<td></td>
<td>8.1</td>
<td></td>
<td>8.4</td>
<td></td>
<td>8.4</td>
<td></td>
<td>8.4</td>
<td></td>
</tr>
<tr>
<td>MLB1770_16</td>
<td>6.3</td>
<td></td>
<td>8.5</td>
<td></td>
<td>10.0</td>
<td></td>
<td>7.5</td>
<td>13.5</td>
<td>6.4</td>
<td>13.9</td>
</tr>
<tr>
<td>SC9346_10</td>
<td>10.2</td>
<td></td>
<td>7.0</td>
<td>17.2</td>
<td>5.7</td>
<td>17.5</td>
<td>5.4</td>
<td>15.4</td>
<td>5.0</td>
<td>15.9</td>
</tr>
<tr>
<td>SCCXIV38K_16</td>
<td>9.7</td>
<td>20.8</td>
<td>6.7</td>
<td>19.1</td>
<td>6.0</td>
<td>19.4</td>
<td>5.5</td>
<td>17.1</td>
<td>5.4</td>
<td>17.9</td>
</tr>
<tr>
<td>SCD9717_7</td>
<td>9.7</td>
<td>16.4</td>
<td>7.5</td>
<td>13.9</td>
<td>7.7</td>
<td>14.2</td>
<td>7.1</td>
<td>12.6</td>
<td>6.6</td>
<td>12.7</td>
</tr>
<tr>
<td>SPPR1GEN_4</td>
<td>4.7</td>
<td></td>
<td>7.1</td>
<td></td>
<td>7.8</td>
<td></td>
<td>7.6</td>
<td></td>
<td>8.0</td>
<td></td>
</tr>
<tr>
<td>SPAC17G8_10</td>
<td>11.1</td>
<td>17.8</td>
<td>8.7</td>
<td>17.4</td>
<td>6.5</td>
<td>18.5</td>
<td>6.0</td>
<td>16.3</td>
<td>5.5</td>
<td>17.4</td>
</tr>
<tr>
<td>SYCSR1G_6</td>
<td>9.1</td>
<td>6.4</td>
<td>15.6</td>
<td></td>
<td>6.1</td>
<td>16.7</td>
<td>5.2</td>
<td>15.5</td>
<td>5.3</td>
<td>15.7</td>
</tr>
<tr>
<td>SYCSR1G_63</td>
<td>13.2</td>
<td>15.7</td>
<td>8.7</td>
<td>13.5</td>
<td>7.7</td>
<td>13.7</td>
<td>7.0</td>
<td>13.8</td>
<td>4.9</td>
<td>13.8</td>
</tr>
<tr>
<td>YSCI8083_15</td>
<td>9.6</td>
<td>7.0</td>
<td>16.8</td>
<td></td>
<td>6.0</td>
<td>17.1</td>
<td>5.9</td>
<td>14.8</td>
<td>5.4</td>
<td>15.5</td>
</tr>
<tr>
<td>YSCI9470_15</td>
<td>10.1</td>
<td>18.7</td>
<td>7.3</td>
<td>15.7</td>
<td>7.2</td>
<td>16.3</td>
<td>6.5</td>
<td>14.3</td>
<td>6.2</td>
<td>14.2</td>
</tr>
<tr>
<td>B61188</td>
<td>3.7</td>
<td></td>
<td>5.3</td>
<td></td>
<td>6.1</td>
<td></td>
<td>7.2</td>
<td></td>
<td>7.5</td>
<td></td>
</tr>
</tbody>
</table>

Highest false positive

|                  | 7.0 | 7.2 | 7.4 | 6.9 | 7.0 |
Overview

• **Allergy**
  – What is an allergen/allergic reaction?

• **Generalized profiles**
  – Modeling sequence motifs
  – Construction of a profile

• **Allergenicity prediction**
  – Current prediction algorithm (FAO/WHO)
  – Profile-based prediction
Allergens in Transgenic Food

Nordlee et al., 1996, N Engl J Med

Brazil nut

Methionine-rich 2S Albumin

Soy

Met-rich

tg-Soy

allergic reaction in nut-sensitized patients
Prediction of Allergens

No common:

• Structure
• Biochemical properties

⇒ no direct prediction

indirect prediction:

similarity with known allergen

potential cross-reactivity

potentially allergenic
Immunological Cross-reactivity

**antigen A**
birch pollen allergen
(Bet v 1, PDB:1BTV)

**antigen B**
cherry allergen
(Pru av 1, PDB:1E09)
Prediction of Cross-reactivity

'linear' sequence
single loops
secondary structure elements

surface
shape
physico-chemical properties
Cross-reactivity and Sequence Similarity

Sequence-based Clustering

Bet v 1 - cluster

- ALL1_APIGR
- ALL2_APIGR
- BV1A_BETVE
- BV1B_BETVE
- BV1C_BETVE
- BV1D_BETVE
- BV1E_BETVE
- BV1F_BETVE
- BV1G_BETVE
- BV1H_BETVE
- BV1I_BETVE
- BV1J_BETVE
- BV1K_BETVE
- BV1L_BETVE
- BV1M_BETVE
- DAU1_DAUC
- MAL1_MALDO
- MPA1_CARBE
- MPA2_CARBE
- MPAA_CORAV
- MPAG_ALNGL
- PRU1_PRUAR
- PRU1_PRUAV
- PYR1_PYRCO

Bet v 1
- Api g 1 celery
- Api g 2
- Bet v 1 birch
- Dau c 1 carrot
- Mal d 1 apple
- Car b 1 hornbeam
- Car b 1
- Cor a 1 hazelnut
- Aln g 1 alder
- Pru ar 1 apricot
- Pru av 1 cherry
- Pyr c 1 pear
**Proposed Allergenicity Evaluation (WHO/FAO)**

- **Identity-test:**
  - \( n = 6 \) contiguous amino acids

- **Similarity-test:**
  - 35% (80 residues)
# Allergen Prediction

**According to Guidelines**

<table>
<thead>
<tr>
<th>Database:</th>
<th>Adb</th>
<th>Swiss-Prot</th>
<th>Rice</th>
<th>trGEN (human)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
<td>allergen database</td>
<td>general protein database</td>
<td>rice genome (TIGR OsGI)</td>
<td>human genome (translated)</td>
</tr>
<tr>
<td><strong>#Proteins:</strong></td>
<td>779</td>
<td>101’602</td>
<td>10’891</td>
<td>330’743</td>
</tr>
<tr>
<td><strong>Allergens (predicted):</strong></td>
<td>98.6%</td>
<td>67.3%</td>
<td>75.9%</td>
<td>42.9%</td>
</tr>
</tbody>
</table>

**Clinical observation:** < 0.5 %
Performance of Current Allergen Prediction

unsuitable for allergenicity prediction.
**Allergen Database (Adb)**

### Data source (February 11th, 2002)

<table>
<thead>
<tr>
<th>Data source</th>
<th>#accessions</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.allergen.org">www.allergen.org</a> (isoallergens)</td>
<td>165</td>
</tr>
<tr>
<td><a href="http://www.expasy.org">www.expasy.org</a> (Swiss-Prot allergens)</td>
<td>276</td>
</tr>
<tr>
<td><a href="http://www.iit.edu/~sgenel">www.iit.edu/~sgenel</a> (food allergens)</td>
<td>140</td>
</tr>
<tr>
<td><a href="http://www.iit.edu/~sgenel">www.iit.edu/~sgenel</a> (non-food allergens)</td>
<td>341</td>
</tr>
<tr>
<td>add_list.txt</td>
<td>105</td>
</tr>
<tr>
<td><strong>Total accessions:</strong></td>
<td><strong>809</strong></td>
</tr>
</tbody>
</table>

### Database

<table>
<thead>
<tr>
<th>Database</th>
<th>#sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swiss-Prot</td>
<td>439</td>
</tr>
<tr>
<td>Swiss-Prot (variants)</td>
<td>99</td>
</tr>
<tr>
<td>GenBank</td>
<td>269</td>
</tr>
<tr>
<td>PIR</td>
<td>97</td>
</tr>
<tr>
<td>skipped:</td>
<td></td>
</tr>
<tr>
<td>redundant</td>
<td>-111</td>
</tr>
<tr>
<td>unreachable</td>
<td>-7</td>
</tr>
<tr>
<td>annotation problems</td>
<td>-7</td>
</tr>
<tr>
<td><strong>Total retrieved sequences:</strong></td>
<td><strong>779</strong></td>
</tr>
</tbody>
</table>
Automatic Identification of Allergen Profiles

- Adb
- Dataset
- Motif discovery (MEME)
- Motifs
- Matches
- Matches remaining sequences
- Remove matches
- Search dataset
- Collect
# 52 Allergen Profiles

<table>
<thead>
<tr>
<th>Profile Identifier</th>
<th>MEME E-value</th>
<th>Matching allergens</th>
<th>Predominant PROSITE matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM000001</td>
<td>$1.8 \cdot 10^{-4123}$</td>
<td>101</td>
<td>Pathogenesis-related Bet v 1 family</td>
</tr>
<tr>
<td>AM000002</td>
<td>$2.0 \cdot 10^{-1477}$</td>
<td>68</td>
<td>Profilins Pollen proteins (Ole e I)</td>
</tr>
<tr>
<td>AM000003</td>
<td>$1.3 \cdot 10^{-919}$</td>
<td>36</td>
<td>Globins</td>
</tr>
<tr>
<td>AM000004</td>
<td>$3.0 \cdot 10^{-845}$</td>
<td>35</td>
<td>none</td>
</tr>
<tr>
<td>AM000005</td>
<td>$4.8 \cdot 10^{-794}$</td>
<td>22</td>
<td>SCP/Tpx-1/Ag5/PR-1/Sc7</td>
</tr>
</tbody>
</table>

...
Profile-based Allergenicity Prediction

query protein

allergens:

83%
17%

52 profiles

profile match?

yes
no

135 sequences

pair match?

yes
no

non-allergen

allergen
Performance of Prediction: Random Sequences

Profile-based

FAO/WHO

BLAST E-value cut-off

identity length n

performance

100% 90% 80% 70% 60% 50% 40% 30% 20% 10% 0%

10^{-1} 10^{-10} 10^{-20} 10^{-30}

recall

precision
### Performance of Prediction: Swiss-Prot Sequences

<table>
<thead>
<tr>
<th>Prediction method</th>
<th>Predicted Allergens</th>
<th>True Allergens</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAO/WHO</td>
<td>68'356</td>
<td>351</td>
</tr>
<tr>
<td></td>
<td>67.3 %</td>
<td>0.5 %</td>
</tr>
<tr>
<td>Profile-based</td>
<td>4’096</td>
<td>351</td>
</tr>
<tr>
<td></td>
<td>4.0 %</td>
<td>8.6 %</td>
</tr>
</tbody>
</table>
Conclusion

• **Allergen Database (Adb):**
  - most complete allergen sequence resource
  - semi-automated

• **WHO/FAO method for allergenicity evaluation:**
  - unspecific (precision = 0.5%)
  - algorithmic limitation

• **Profile-based prediction:**
  - improved performance (precision = 8.6%)
Outlook

• **Improved prediction:**
  - pure profile based approach
  - two phase prediction (fold, surface)

• **Search for unknown allergens:**
  - xeno- vs. autologous structures
  - production of pan-allergic structures
    (for diagnosis and therapy)