Evaluation and Integration of Existing Methods for Computational Prediction of Allergens

Jing Wang

ICoVax 2012 Workshop  Tongji University  October 13, 2012
What’s allergen?

Allergens

- Food allergens
- Contact allergens
- Aero allergens
- Venom/salivary
- Others
Hazards of allergens

- Chronic ill health are mainly caused by allergy, affecting about **25%** of the population in the world.
- Asthma and atopic dermatitis, respectively, affect **10% and 15%** of the children in some countries.
- Fish allergy and general food allergy were reported in **2.3% and 4%** of the US population, respectively.

... ... ...
Allergen detection methods

- **Experimental methods**
  - High time consumption
  - High cost
  - Difficulty for choosing candidate

- **Bioinformatics prediction methods**
  - Sequence-based method (*FAO/WHO, 2001*)
  - Motif-based method (*Stadler, M. B. et al., FASEB J, 2003*)
  - SVM-based method (*Saha S et al., NAR, 2006*)

**FAO:** Food and Agriculture Organization of the United Nations;  
**WHO:** World Health Organization
Bioinformatics methods

Bioinformatics prediction methods

• Sequence-based method (*FAO/WHO, 2001*)

• Motif-based method (*Stadler, M. B. et al., FASEB J, 2003*)

• SVM-based method (*Saha S et al., NAR, 2006*)

Superior?
Research Objectives

Methods Evaluation

• Compare the performance of a variety of computational methods for allergen prediction
• Find the cons and pros of each method, and perform parameter optimization

Methods integration

• Database search for all known allergen
• One-stop prediction for the protein allergenicity
  • Single protein prediction
  • Batch prediction
Method Evaluation

Workflow of the evaluation

1. Datasets construction
2. Methods implement
3. Parameter optimization
4. Performance comparison
1. Datasets construction

- **Swiss-Prot**
  - 521099 protein sequences

- **Remove (Identities>=30% & AA<50) records**
  - 244538 protein sequences

- **Extract 989 randomly**
  - Negative dataset

- **Reverse allergens’ sequences**

- **Allergens from Swiss-Prot, IUIS, SDAP, ADFS**

- **Eliminate the redundancy**
  - Allergen 989 protein sequences

- **Positive dataset**
1. Datasets construction

Ten-fold cross validation

The dataset was randomly partitioned into ten subsets, where each subset had nearly equal number of allergens and non-allergens. Of the ten subsets, a single set was retained as the validation data for testing the method, and the remaining nine subsets were used as training data. This process was then repeated 10 times with each of the ten subsets used exactly once as the validation data. The overall performance of a method was the average performance over ten subsets.
2. Methods Implement

Sequence-based method (FAO/WHO criteria)

Rule 1: Cutoff = 6

Querying: MQTRSIDNVVNWSRCVHPSCVAWVIFHFCFAKNCISLY.............

AYYVAAGKLDNVVNWSRCVHPSCVAWVIFHFCFAKNCISLY.............

Known allergen sequences

Rule 2: Cutoff = 35%

Querying: EQCFRCLQKQQGGEQRECQRMQSLHKEAPEHTSPEDV...........

EQCFRC-LQKQQGGEQRECQRMQSLHKE-----APEHTSPEDVGTEEEEEHERE
+QC+ +C +Q+Q +Q+E C + C+ +K+ E+ E+ GTG +E HE
KQCKHQCKVQRQYDEQQKEQCVKECEKYYKEKKGREEREHEEEEEEEWGTGVD
ESPTE---

Identities = 31/94 (32%),

Known allergen blast database

Fiers MW et al., BMC Bioinformatics, 2004
2. Methods implement Motif-based method

- Allergens

**E-value \( \leq 0.01 \)?**

- Yes
  - MAST
  - **Motifs**

- No
  - End

Start

**Non-iteration**

**Iteration**

**Start**

- **Allergens**

- **MEME**

- **MAST**

- **Motifs**

- **End**

Querying:

```
MQTRSIDNVNWSRCVHPSCVAWVIFIHFCFAKNCSILY..........
```

*Stadler, M. B. et al., FASEB J, 2003*

*Saha S et al., NAR, 2006*
2. Methods implement

SVM-based method

Allergens → Coding → Vectors → SVM → Model

Non-Allergens

Fraction of amino acid \( i = \frac{\text{total number of amino acids (}i\text{)}}{\text{total number of amino acids in protein}}\)

Amino Acid Composition

MQTRSIDNVVNWSRCVHPSCVAVVIFHFCFAKNCILY...........

Querying:

Saha S et al., *Nucleic Acids Research*, 2006
3. Evaluation results

FAO/WHO

Rule 1

Rule 2
3. Evaluation results

Motif-based method

| MAST E-value | 0.001 | 0.01  | 0.1  | 0.5  | 0.7  | 1    | 10   |
|--------------|-------|-------|------|------|------|------|------|
| **Iteration** |       |       |      |      |      |      |      |
| Sensitivity  | 62.63% | 63.64% | 66.67% | 77.78% | 80.81% | 82.83% | 100% |
| Specificity  | 98.99% | 98.99% | 96.97% | 81.82% | 76.77% | 66.67% | 0%   |
| **Non-iteration** |       |       |      |      |      |      |      |
| Sensitivity  | 13.66% | 13.66% | 13.66% | 13.95% | 14.16% | 14.77% | 23.56% |
| Specificity  | 100%   | 100%   | 99.70% | 99.19% | 98.89% | 98.48% | 86.15% |
4. Methods comparison

- SVM-based method and comparison (92.82%, 90.59%)
## 4. Methods comparison

### Methods comparison
- **Time complexity**

| Approaches | F/W rule 1 | F/W rule 2 | Motif-based | SVM-based |
|------------|------------|------------|-------------|-----------|
| Time (ms*) | 15940      | 58640      | 87          | 10        |

*ms means millisecond*
Methods integration

Why integration?
Welcome to proAP!

proAP, Protein Allergenicity Prediction, is a web-based database of allergenic proteins, providing bioinformatics tools to determine cross-reactivities between potential allergens and known allergens.
Main modules

- Allergen search

Statistics
- Allergen: 1096
- Allergen category: 13
- Species: 249
- Prediction method: 4
Main modules

Allergenicity prediction

Requiring:

Methods:

Select method(s): (One or more methods)

- FAO/WHO: Amino acids sliding window: 80, sequence identity cutoff ≥ 35%
- FAO/WHO: Exact match for ≥ 6 contiguous amino acids
- Motif-based method
- SVM-AAC method
Next step …

- Key features for allergenicity
- Family preference
- Specific structures
Acknowledgments

Dr. Jing Li

Yabin Yu

Yunan Zhao

Prof. Dabing Zhang
Software Demonstration

proAP -- Protein Allergenicity Prediction

Play
Thank you
http://gmobl.sjtu.edu.cn/proAP/main.html
Welcome to proAP!

proAP, Protein Allergenicity Prediction, is a web-based database of allergenic proteins, providing bioinformatics tools to determine cross-reactivities between potential allergens and known allergens.
Protein Allergenicity Prediction

Navigation
- Introduction
- Allergen search
- Allergenicity prediction
- Batch prediction

Allergen search
- By category:
  - Aero Animal
  - Aero Fungi
  - Aero Insect
  - Aero Mite
  - Aero Plant
  - Contact
  - Food Animal
  - Food Fungi
  - Food Plant
  - Gladin
  - Protozoan
  - Venom/Salivary
  - Worm

Search

Contact us
| No. | Name | UniProtAcc | TaxonomicName | CommonName | Category | Species | Epitope | Sugar |
|-----|------|------------|---------------|------------|----------|---------|---------|-------|
| 1   | Bos d 3 | Q28050 | Bos taurus | Bos bovis, Bos primigenius taurus, Bos taurus, bovine, cow, domestic cow, domestic cattle, cattle, | Aero Animal | Cattle |         |       |
| 2   | Can f 1 | O18873 | Canis lupus familiaris | Canis domesticus, Canis familiaris, Canis lupis familiaris, dogs, dog, | Aero Animal | other |         | sugar |
| 3   | Can f 2 | O18874 | Canis lupus familiaris | Canis domesticus, Canis familiaris, Canis lupis familiaris, dogs, dog, | Aero Animal | other |         | sugar |
| 4   | Can f 3 | P45822 | Canis lupus familiaris | Canis domesticus, Canis familiaris, Canis lupis familiaris, dogs, dog, | Aero Animal | other |         | Serum |

*Protein Allergenicity Prediction*
Protein Allergenicity Prediction

Navigation
- Introduction
- Allergen search
- Allergenicity prediction
- Batch prediction

Allergenicity prediction

Predict methods including FAO/WHO criteria, Motif-based method and SVM-AAC method (take amino acid composition as protein features). Choose the method according to your demand. Details about predict methods see Introduction page.

Paste or type your sequence: (One sequence only. Fasta format or pure aa sequence)

Select method(s): (One or more methods)

- [ ] FAO/WHO: Amino acids sliding window: 80, sequence identity cutoff >= 35 %
- [ ] FAO/WHO: Exact match for >= 6 contiguous amino acids
- [ ] Motif-based method
- [ ] SVM-AAC method

Select category: (Only for FAO/WHO methods)

- [ ] By Aero Animal
- [ ] By all.

[Cancel]  [Predict]  [Reset]
Protein Allergenicity Prediction

Navigation
- Introduction
- Allergen search
- Allergenicity prediction
- Batch prediction

Allergenicity prediction

Predict methods including FAO/WHO criteria, Motif-based method and SVM-AAC method (take amino acid composition as protein features). Choose the method according to your demand. Details about predict methods see Introduction page.

Paste or type your sequence: (One sequence only. Fasta format or pure aa sequence)

MGIKHCCYILYTLVLQAGRGLGEEVVDLPSNDRTSSQLNLCHLVIDK
CWRCKPDWAERQALGD
CAQGFGLKATHGGKGWGDYYMVTSDDVTVNPKEGLRFGATQDRPLWII
FQRDMIYLQEMVTSDKTI
DGRGAKVFLVYGGITIMNVKNNIHHNIDIDVRVIQGRIKSNIGPAPRHO

Select method(s): (One or more methods)
- [ ] FAO/WHO: Amino acids sliding window: 80, sequence identity cutoff >= 86%
- [ ] FAO/WHO: Exact match for >= 5 contiguous amino acids
- [ ] Motif-based method
- [ ] SVM-AAC method

Select category: (Only for FAO/WHO methods)
- [ ] By Aer Animal
- [ ] By all.

Predict Reset
Protein Allergenicity Prediction

Home  | Site map  | Contact us

Navigation
- Introduction
- Allergen search
- Allergenicity prediction
- Batch prediction

Prediction results

Query protein:
MQIllLTGALICGLQAEGNHEEPQGGLEELSGRWHSVAlASNKSDLIKPWGHFRVF
HSMASAKDGNLHDULPQDGQCEKVSLTAFKTANSNKFDFLEYWGHNLYLAEVDPKSYL
LYMINQYNDTSLVAHLMVRLSRRQDLPFAPESVCEDIGLHDKQIVVLSDDDRCQGSRD

Prediction result

| FAO/WHO: sequence alignment | FAO/WHO: amino acids match | Motif-based | SVM-AAC |
|-----------------------------|-----------------------------|-------------|---------|
| allergen                    | --                          | --          | --      |

FAO/WHO: Sequence identity cutoff >= 45% (allergen)

Matched records:

| No. | Name  | UniProtAcc | TaxonomicName | CommonName | Category | Species | Epitope | Canis |
|-----|-------|------------|---------------|------------|----------|---------|---------|-------|
|     |       |            |               |            |          |         |         |       |
### Matched records:

| No. | Name   | UniProtAcc | TaxonomicName                          | CommonName                                          | Category   | Species | Epitope | Sugar | Description                                      |
|-----|--------|------------|----------------------------------------|-----------------------------------------------------|------------|---------|---------|-------|--------------------------------------------------|
| 1   | Can f 2| O18874     | Canis lupus familiaris                  | Canis domesticus, Canis familiaris, Canis lupus familiaris, dogs, dog. | Aero Animal| other   |         | sugar | Minor allergen Can f 2 (Allergen Dog 2) (Can f 2) (Flags: Precursor) |

### Alignments detail:

**Query:** (1--80)

>mq018874]ALL2_CANFA Minor allergen Can f 2 OS=Canis familiaris PE=1 SV=1

Length = 180

Score = 166 bits (421), Expect = 4e-44, Method: Compositional matrix adjust.

Identities = 80/80 (100%), Similarity = 100%, Positives = 80/80 (100%)

| Query | Subject |
|-------|---------|
| MQLLLTVGLAGICGFLQAEGNHHEEPFGQGLLEEGRHWSVALASNSKDLIFKFWGHPRFVI 60 | MQLLLTVGLAGICGFLQAEGNHHEEPFGQGLLEEGRHWSVALASNSKDLIFKFWGHPRFVI 60 |
| QLQLLTVGLAGICGFLQAEGNHHEEPFGQGLLEEGRHWSVALASNSKDLIFKFWGHPRFVIH 60 | QLQLLTVGLAGICGFLQAEGNHHEEPFGQGLLEEGRHWSVALASNSKDLIFKFWGHPRFVIH 60 |

**Query:** (2--81)

>mq018874]ALL2_CANFA Minor allergen Can f 2 OS=Canis familiaris PE=1 SV=1

Length = 180

Score = 166 bits (420), Expect = 5e-44, Method: Compositional matrix adjust.

Identities = 80/80 (100%), Similarity = 100%, Positives = 80/80 (100%)

| Query | Subject |
|-------|---------|
| HSMASKDGNLHDGLIPDGD 80 | HSMASKDGNLHDGLIPDGD 80 |
| HSMASKDGNLHDGLIPDGD 80 | HSMASKDGNLHDGLIPDGD 80 |
Protein Allergenicity Prediction

Navigation
- Introduction
- Allergen search
- Allergenicity prediction
- Batch prediction

Allergenicity prediction

Predict methods including FAO/WHO criteria, Motif-based method and SVM-AAC method (take amino acid composition as protein features). Choose the method according to your demand. Details about predict methods see Introduction page.

Paste or type your sequence: (One sequence only. Fasta format or pure aa sequence)
LDPAIIYVÄEQNMSASDVYNAIKNKSGMLALTGASDMRVDVFAKPOENA
VAIKMYVNRV
ADYIAKYNQLLEGNIDLVFTGGIGENASDCVELFINAVKSLGFATDLKLFLVK
YGDSGVV
STPQSKYKIVRVTNEELMIVEDSIRLTK

Select method(s): (One or more methods)
- [ ] FAO/WHO: Amino acids sliding window: 80, sequence identity cutoff >= 35 %
- [x] FAO/WHO: Exact match for >= 6 contiguous amino acids
- [x] Motif-based method
- [x] SVM-AAC method

Select category: (Only for FAO/WHO methods)
- [ ] By: [ ] Aero Animal
- [ ] By all.

[Predict] [Reset]
**Protein Allergenicity Prediction**

**Query protein:**

```
MNDNKILVNNAGSSIKFQLFDYHKKVLAKALCERIFVDGFFKLEFNEQKVEEOKVAFPDH
HAAVTHFLNLTKKKIIEQLSDDILVHRVQGANYFKDSVVIDAEALAKIKEFTKLAAPL
HNKPEADVIIEFKKEPASKNVAVFTDTHTTPCQENLYAYPRSWEQKHLVRKYGHFHT
SYKFINNYLHIKLNQNLNMLVCLHLNGASVCMAIKNGKFNTSMGFTPELEGIMGTSGD
LDPAIIGYVAEQENMSASDVNALSNKSGMLALTGASDMRDFVAKPQENAVAIKMYVNRV
ADYIAKYLNQLGENIDGLVFTGIGIGENASDCVELFIVARKLGFATDLKLFPKVYGDCVY
STPQSKYIKYRVRTNEELMIVESIRLTKK
```

**Prediction result**

| FAO/WHO: sequence alignment | FAO/WHO: amino acids match | Motif-based | SVM-AAC |
|-----------------------------|---------------------------|-------------|---------|
| --                          | allergen                  | allergen    | non-allergen |

**FAO/WHO: Exact match 6 amino acids (allergen)**

**Matched records:**

| No. | Name | UniProtAcc | TaxonomicName | CommonName | Category | Species | Epitope |
|-----|------|------------|---------------|------------|----------|---------|---------|
| No. | Name     | UniProtAcc | TaxonomicName                | CommonName                                                                 | Category  | Species  | Epitope | Sugar |
|-----|----------|------------|------------------------------|----------------------------------------------------------------------------|-----------|----------|---------|-------|
| 1   | Ph a 5   | P56166     | Phalaris aquatica L., Phalaris tuberosa, Phalaris aquatica, canary grass, | P56166 Major pollen allergen Ph a 5.3 OS=Phalaris aquatica PE=1 SV=1 | Aero Plant| Grass    |         |       |
| 2   | Par c 2 ?| A2V734     | Paralithodes camtschaticus   | Paralithodes camtschaticus, Kamchatka crab, red king crab,                | Food Animal| other    |         |       |
Motif-based result (allergen)

Your querying is predicted as allergen. Matched motif is described as:

| SEQUENCE NAME | DESCRIPTION | E-VALUE | LENGTH |
|---------------|-------------|---------|--------|
| name          |             | 0.00018 | 390    |

SVM-AAC result

Your query is predicted as non-allergen (probability=0.975783).
Protein Allergenicity Prediction

Navigation
- Introduction
- Allergen search
- Allergenicity prediction
- Batch prediction

Batch prediction

Upload a FASTA-format file containing multiple protein sequences to be predicted for allergenicity. Results of the prediction will be returned to you at the email address that you specify. Please check the notes below for the restrictions on uploaded sequence files.

Sequences file:

Prediction method:
- FAO/WHO: Amino acids sliding window: 80, sequence identity cutoff >= 35%
- FAO/WHO: Exact match for >= 6 contiguous amino acids
- Motif-based method
- SVM-AAC method

Email address:

[Submit]  [Reset]
Thank you