The COMPARE Database: a public COMprehensive Protein Allergen REsource for protein allergenicity assessment

Lars K. Poulsen
Professor of Basic Allergology
Allergy Clinic, Copenhagen University Hospital at Gentofte
Copenhagen, Denmark

Health and Environmental Sciences Institute (HESI)
Protein Allergens, Toxins and Bioinformatics (PATB) Committee
**Objective**

- Establish weight-of-evidence approach
- Develop new science
- Communicate scientific findings and best practices

**Activity**

**Assessment**

- COMPARE Allergen Dbase
- Protein Toxins Evaluation

**Research**

- Digestibility Study
- Allergen Rebuild Study
- Matrix Study
- GARD Assay Study
- Immunogenicity Study

**Outreach and Training**

- Workshops, e.g. *today* and: “Non-IgE mediated reactions to foods” Workshop. 2016. Rome, Italy.
- Presentations at international meetings
- Publications

➢ Develop new science
➢ Establish weight-of-evidence approach
➢ Communicate scientific findings and best practices
HESI COMPARE Database: www.comparedatabase.org

- Comprehensive and up to date collection of peer reviewed allergen protein sequences.
- Custom process development to identify new allergens initiated in 2016.
  - Includes the allergens listed in AOL v.16 and 14 new allergens identified though bioinformatics approaches.
- COMPARE 2018 (Feb. 2018):
  - Added 68 allergens to COMPARE 2017 (2038 allergens total).
- Updated annually
About the Process

- Automated bioinformatics sorting algorithm.
- Scientific information and publications related to individual identified sequences.
- Independent Peer Review Panel of recognized allergy experts, public sector.
- A quality control and documentation.
- A public release & annual update.

✓ Publicly accessible, transparent, rigorous and reliable resource.
There are already multiple allergen databases available, with different features and purposes…

E.g., Allergome, AllergenOnline (AOL), IUIS, AllFam, SDAP…


… but in recent years the widespread use of genomic sequencing technology brought new challenges….
Number of sequences to be filtered has grown exponentially.

- ~125 Million sequence entries (UniProtKB/TrEMBL)
- Most of them are “predicted”

<table>
<thead>
<tr>
<th>Protein existence (PE)</th>
<th>Entries</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Evidence at protein level</td>
<td>144459</td>
<td>0.12%</td>
</tr>
<tr>
<td>2: Evidence at transcript level</td>
<td>1162753</td>
<td>0.93%</td>
</tr>
<tr>
<td>3: Inferred from homology</td>
<td>30704463</td>
<td>24.60%</td>
</tr>
<tr>
<td>4: Predicted</td>
<td>92785433</td>
<td>74.35%</td>
</tr>
<tr>
<td>5: Uncertain</td>
<td>0</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Source: https://www.ebi.ac.uk/uniprot/TrEMBLstats

The COMPARE process addresses this growth by implementing a cutting-edge and high-throughput bioinformatic pipeline to identify a meaningful subset of these millions sequences, for scientific review.
COMPARE: Bringing together the required strategic, technical and logistical considerations

- Tri-partite by design
- Strategic partnerships: JIFSAN, Academia & Government
- HESI infrastructure support and network (independent non-profit, program coordination, IT, data management, science-focused finance, communications).
- Long-term stability for all of the above
RATIONALE FOR ALLERGEN DATABASE at HESI

(1) Novel components due to HESI public-private partnership.

✓ Multiple stakeholders engagement.

✓ Custom algorithm development

✓ **Independent coordinating body:**
  - Scientific program management
  - Budget and contracts management with external professional services.

**HESI PATB COMPARE**
P-P Science Advisory Team

**Independent Peer Review Panel (PRP)**
Allergy Science Expertise

✓ Decisions based on peer-review

✓ Rigorous process for data management and tracking (custom information management tool).

✓ Literature procurement
RATIONALE FOR ALLERGEN DATABASE at HESI
(2) Modular process allows multiple sourcing of allergen candidates.

Goal: enhance confidence in the efficiency and quality of allergens identified.
What has been done:

➢ COMPARE overview brochure: at initiation of the project (Q1 2016).


➢ Open dialogue w/ stakeholders ([http://comparedatabase.org/contact-us/](http://comparedatabase.org/contact-us/))

In preparation:

➢ Peer-review publication
Overall program and components designed to enable long term, stable support.

<table>
<thead>
<tr>
<th>Advisory Team</th>
<th>Bioinformatics and Data Communication</th>
<th>Program Coordination</th>
<th>Expert Peer Review Panel</th>
<th>Participating Organizations</th>
</tr>
</thead>
</table>
| • 20 Public/Private stakeholders.  
• General program oversight, QC, user feedback.  
• Expert knowledge  
• No influence on decisions regarding sequence inclusion or exclusion. | • **Bioinformatics partner:** sequences screening; deduplication, metadata retrieval and QC.  
• **University of Maryland / JIFSAN:**  
  • Literature search and procurement.  
  • Data sharing.  
  • Review Tool development and maintenance. | • Overall program management.  
• Coordination between partners, PRP assembly and management.  
• Database ‘home’ and maintenance. | • Invited public sector allergy experts.  
• Define inclusion/exclusion criteria.  
• Review of sequences and references/publications.  
• Provide final decisions on new sequences to include in database | **Public Sector:**  
• University of Amsterdam  
• University of Maryland / JIFSAN  
• U.S. EPA  
• U.S. FDA  

**Private Sector:**  
• BASF  
• Bayer CropScience  
• DowDuPont Pioneer  
• KWS  
• Syngenta  
• Vilmorin & Cie

JIFSAN: Joint Institute for Food Safety And Nutrition (UMD/FDA)
Import candidates sequences & metadata to Review Tool

Pass sequences and meta data to JIFSAN

Literature procurement

Import candidates sequences & metadata to Review Tool

Pass on to experts (Peer Review)

KEEP – New and Similar to known allergen\(^1\) (not just putative allergen\(^2\))

Associate existing citation in NCBI with allergen to provide to PRP

Final DBase update build: add old to new, formatted.

Final “approved” sequence list passed back to Bioinformatics partner

1 Known allergen = protein sequence with published clinical allergy data

2 Putative allergen = protein sequence with homology to known allergen, but lacks direct published clinical allergy data
AOL v16: 1956 known allergens

COMPARE 2017

World of Protein Sequences, NCBI: May 2015 to May 2016

Total entries downloaded from NCBI: 55,641
Total entries filtered from custom automatic screening tool: 568
Total entries submitted to Peer Review Panel (PRP) for review: 251
PRP: 14
Final number of allergen sequences for inclusion in database

End: 1970 allergens in COMPARE 2017

a) using keyword “allerg*”
b) using keywords such as “lipid transfer protein”, “profilin”, “tropomyosin”, etc
**COMPARE 2018**

**START: COMPARE 2017**

**END: 2038 allergens in COMPARE 2018**

- **World of Protein Sequences, NCBI:** May 2018 - May 2017
  - a) 50,638 total entries downloaded from NCBI
  - b) 67,327 total entries downloaded from NCBI

- **Total entries filtered from custom automatic screening tool:**
  - a) 1,377
  - b) 151

- **Manual Exam QC:**
  - 204 total entries submitted to Peer Review Panel (PRP) for review

- **PRP:**
  - 68 final number of allergen sequences for inclusion in database

- **2038 allergens in COMPARE 2018**
“COMPARE 2018” numbers and current candidates numbers

<table>
<thead>
<tr>
<th>COMPARE 2018</th>
<th>Reviewed Allergens</th>
<th>Reviewed Articles</th>
<th>Approved Allergens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioinformatics screenings (NCBI)</td>
<td>93</td>
<td>186</td>
<td>27</td>
</tr>
<tr>
<td>Literature Search</td>
<td>-</td>
<td>37</td>
<td>29</td>
</tr>
<tr>
<td>AOL 17</td>
<td>22</td>
<td>35</td>
<td>12</td>
</tr>
<tr>
<td>TOTAL</td>
<td>115</td>
<td>258</td>
<td>68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COMPARE 2019 - preview</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
</tr>
</tbody>
</table>
Overview of www.comparedatabase.org
<table>
<thead>
<tr>
<th>SPECIES</th>
<th>COMMON NAME</th>
<th>GI #</th>
<th>ACCESSION #</th>
<th>LENGTH</th>
<th>YEAR ADOPTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lupinus albus</td>
<td>Lupine (Field)</td>
<td>0075121065</td>
<td>Q8E8C1.1</td>
<td>0533</td>
<td>2017</td>
</tr>
<tr>
<td>Beta vulgaris</td>
<td>Beet</td>
<td>0205820383</td>
<td>P85984.1</td>
<td>0023</td>
<td>2017</td>
</tr>
<tr>
<td>Actinidia deliciosa</td>
<td>Kwifruit</td>
<td>0906849967</td>
<td>4X9U_A</td>
<td>0189</td>
<td>2017</td>
</tr>
<tr>
<td>Amaranthus retroflexus</td>
<td>Pigweed</td>
<td>0914400010</td>
<td>AKV72168.1</td>
<td>0168</td>
<td>2017</td>
</tr>
<tr>
<td>Anisakis simplex</td>
<td>Nematode</td>
<td>0957554293</td>
<td>BAT02430.1</td>
<td>0217</td>
<td>2017</td>
</tr>
<tr>
<td>Lupinus angustifolius</td>
<td>Narrowleaf Lupine</td>
<td>0980951518</td>
<td>F5B8W9.1</td>
<td>0611</td>
<td>2017</td>
</tr>
<tr>
<td>Lupinus angustifolius</td>
<td>Narrowleaf Lupine</td>
<td>0980951548</td>
<td>F5B8W0.1</td>
<td>0603</td>
<td>2017</td>
</tr>
<tr>
<td>Lupinus angustifolius</td>
<td>Narrowleaf Lupine</td>
<td>0980951550</td>
<td>F5B8W1.1</td>
<td>0580</td>
<td>2017</td>
</tr>
<tr>
<td>Lupinus angustifolius</td>
<td>Narrowleaf Lupine</td>
<td>0980951555</td>
<td>F5B8W2.1</td>
<td>0590</td>
<td>2017</td>
</tr>
<tr>
<td>Lupinus angustifolius</td>
<td>Narrowleaf Lupine</td>
<td>0980951561</td>
<td>F5B8W3.1</td>
<td>0637</td>
<td>2017</td>
</tr>
<tr>
<td>Lupinus angustifolius</td>
<td>Narrowleaf Lupine</td>
<td>0980951568</td>
<td>F5B8W5.1</td>
<td>0605</td>
<td>2017</td>
</tr>
<tr>
<td>Arachis hypogaea</td>
<td>Peanut</td>
<td>1018736824</td>
<td>B3EWP3.1</td>
<td>0071</td>
<td>2017</td>
</tr>
<tr>
<td>Arachis hypogaea</td>
<td>Peanut</td>
<td>1018736830</td>
<td>B3EWP4.1</td>
<td>0079</td>
<td>2017</td>
</tr>
<tr>
<td>Arachis hypogaea</td>
<td>Peanut</td>
<td>1018736837</td>
<td>C0HUZ1.1</td>
<td>0072</td>
<td>2017</td>
</tr>
</tbody>
</table>
**COMPARE 2016-2018 and beyond**

### 2016
- Steering team formed
- Bioinfo process v 1.0
- PRP set up
- Literature procurement (JIFSAN)

### 2017
- Bioinfo process v 2.0
- Review Tool Development (JIFSAN) *New!* ;
- Piloting: lit search, IUIS, AOL. *New!*  

### 2018
- **Identified new collaborators for:**  
  - Bioinformatics screening (to meet the needs of the program);  
  - Literature search (in parallel of Bioinfo screening).  

- **Process efficiency improvements:**  
  - For capture of sequences & metadata, deduplication & QC;  
  - Consolidation of multiple sources  

- **New components development and activities:**  
  - COMPARE “historic rebuild”  
  - FASTA Search Tool;  
  - Reviewers decisions online (transparency);  
  - Online Dbase functional improvements.
Current efforts

Developing new features in COMPARE Website:

- FASTA search tool
- Increased transparency: making reviewers decision & comments publicly accessible
- Database improvements: more user friendly (making it easier to retrieve references)

Applying the current process to historic data (2016 data and previous years)

- Bioinformatic screenings from 2016 and previous years;
- Reviewing quality of “older” sequences and updating.

To be available with the “COMPARE 2019” release
a) FASTA Bioinformatics Search

http://fasta.test.jifsan.org/
b) Database improvements

1) Clickable Metadata?
   - http://www.uniprot.org/uniprot/?query=NP_001005208&sort=score
   - Click on Species searched COMPARE?

2) Sequence data added

3) Automatically show all related articles.

4) Make Pubmed ID clickable,

5) Allow alternative links,

6) Pull other COMPARE allergens associated with this article
Access to a transparent and consensus-based allergen database is a key aspect for supporting public safety.

Who?

- **Product developers** providing safety information on novel proteins.
- **Regulatory agencies** responsible for food and feed safety assessments.
- **Clinicians**, medical personnel and caregivers in the allergy field.
- **The Public**, who may be interested in identifying sources of allergens.

How?

- Bioinformatics tools such as the FASTA algorithm are utilized to evaluate the degree of similarity between novel proteins introduced into a crop and known allergens – COMPARE allows these bioinformatic comparisons.
- Researching types of allergens, sequence information and literature associated.
- Staying updated with new allergens identified each year (all users).
Acknowledgements

- **HESI PATB Committee**
- **HESI PATB COMPARE Steering Team:**
  - Katie Bailey, Syngenta
  - Laurent Beuf, Vilmorin et Cie
  - Christal Bowman, formerly, Bayer
  - Supratim Choudhuri, U.S. FDA
  - Eva Gietl, KWS
  - Kevin Glenn, formerly, Monsanto
  - Rod Herman, DowDupont
  - Norma Houston, DowDupont
  - Emir Islamovic, BASF
  - John Kough, U.S. EPA
  - Greg Ladics, Dupont
  - Scott McClain, Syngenta
  - Kyle McKillop, Univ. of Maryland / JIFSAN
  - Henry Mirsky, Dupont
  - Clare Narrod, Univ. of Maryland / JIFSAN
  - Esmeralda Posada C, BASF
  - Andre Silvanovich, Bayer
  - Ping Song, DowDupont
  - Ronald van Ree, Academic Medical Center, University of Amsterdam

- **Peer Review Panel Experts**
  - Cecilia Berin, PhD, Icahn School of Medicine at Mount Sinai, NY, USA
  - Karin Hoffman-Sommergruber, PhD, Medical University of Vienna, Austria
  - Lars Poulsen, PhD, Copenhagen University Hospital, Denmark
  - Suzanne Teuber, MD, University of California, Davis, CA, USA
  - Ronald van Ree, PhD, Academic Medical Center, University of Amsterdam, The Netherlands

- **HESI Staff:**
  - Syril Pettit, Exec. Dir.
  - Lauren Peel, Associate Sc. Prog. Mngr.
  - Oscar Bermudez, Sc. Prog. Associate
  - Brianna Farr, Sc. Prog. Associate
To Learn more:


➢ COMPARE: www.comparedatabase.org

➢ Questions: contact the HESI COMPARE Team at: compare@hesiglobal.org

www.hesiglobal.org