МЕДИЦИНСКИ УНИВЕРСИТЕТ
СОФИЯ

ФАРМАЦЕВТИЧЕН ФАКУЛТЕТ
СОФИЯ

Ivan Dimitrov

Faculty of Pharmacy, Medical University of Sofia, Bulgaria

In silico approaches for allergenicity (IgE and non- IgE) assessment

Phases in allergic reaction

Sensitisation

- The primary immune response to the first exposure to an allergen
- No symptoms occur while the capacity of the immune system to react increases dramatically

Elicitation

- The secondary immune response to an allergen
- Begins after re-exposure to the same allergen
- Patients exhibit clinical manifestations

Bioinformatic approaches for allergenicity assessment

- Current *in silico* approaches cannot differentiate between sensitization and elicitation phase of allergy.
- The *in silico* approaches are based on the comparison of the structure of a new protein with the structure of proteins already known to possess allergenic potential.
- Computational protocols and algorithms could be considered as methods primarily for assessment of IgE cross-reactivity rather than allergenicity in general.

Bioinformatic approaches for allergenicity assessment

1. Sequence alignment methods
   - Following FAO/WHO Codex Allimentarius guidelines
   - Search for sequence similarity of the whole protein sequences using specific algorithms (FASTA, BLAST)

2. Identification of conserved allergenicity-related linear motifs
   - Applications of different approaches for protein sequence presentation and different machine learning methods

Both approaches are based on the assumption that the allergenicity is a **linearly coded property**.

Bioinformatic approaches for allergenicity assessment

3. Modeling the primary structure of the proteins and search for similarity by machine learning and computational methods

4. Search for similarity in the tertiary structure of a novel protein and the known tertiary structure of allergens.

5. Modeling the physicochemical properties of allergenic proteins and classification with a machine learning algorithm.

6. Application of artificial intelligence (AI)

The Codex Allimentarius recommendations

A query protein is potentially allergenic if:

- **It has an identity of 6 to 8 contiguous amino acids with a known allergen**
- **It has > 35% sequence identity over a window of 80 amino acids when compared with known allergens.**

- High specificity for proteins that are structurally identical with known allergens
- Produce many false positives and low specificity i.e. does not recognize non-allergens.
- Does not take into account the presence of gaps and of amino acids with similar physicochemical properties in the protein sequence e.g. Arg and Lys, Glu and Asp
- Assessment of a novel protein as an allergen may be limited by the lack of identity to known allergens

Reasons for the Codex recommendations

- The use of 6-8 contiguous, identical amino acids as a match is predicated according the minimal length of known IgE and T cell epitopes:
  - The minimal IgE-binding epitopes of Ara h 1 and Ara h 2 involve 6 contiguous amino acids.
  - The minimum peptide length for a T cell–binding epitope is 8 contiguous amino acids.
- The sliding window of 80 amino acids corresponds to a typical domain size of a protein and recognizes that single protein domain may contain epitopes that mediate antibody binding.

Reasons for the Codex recommendations II

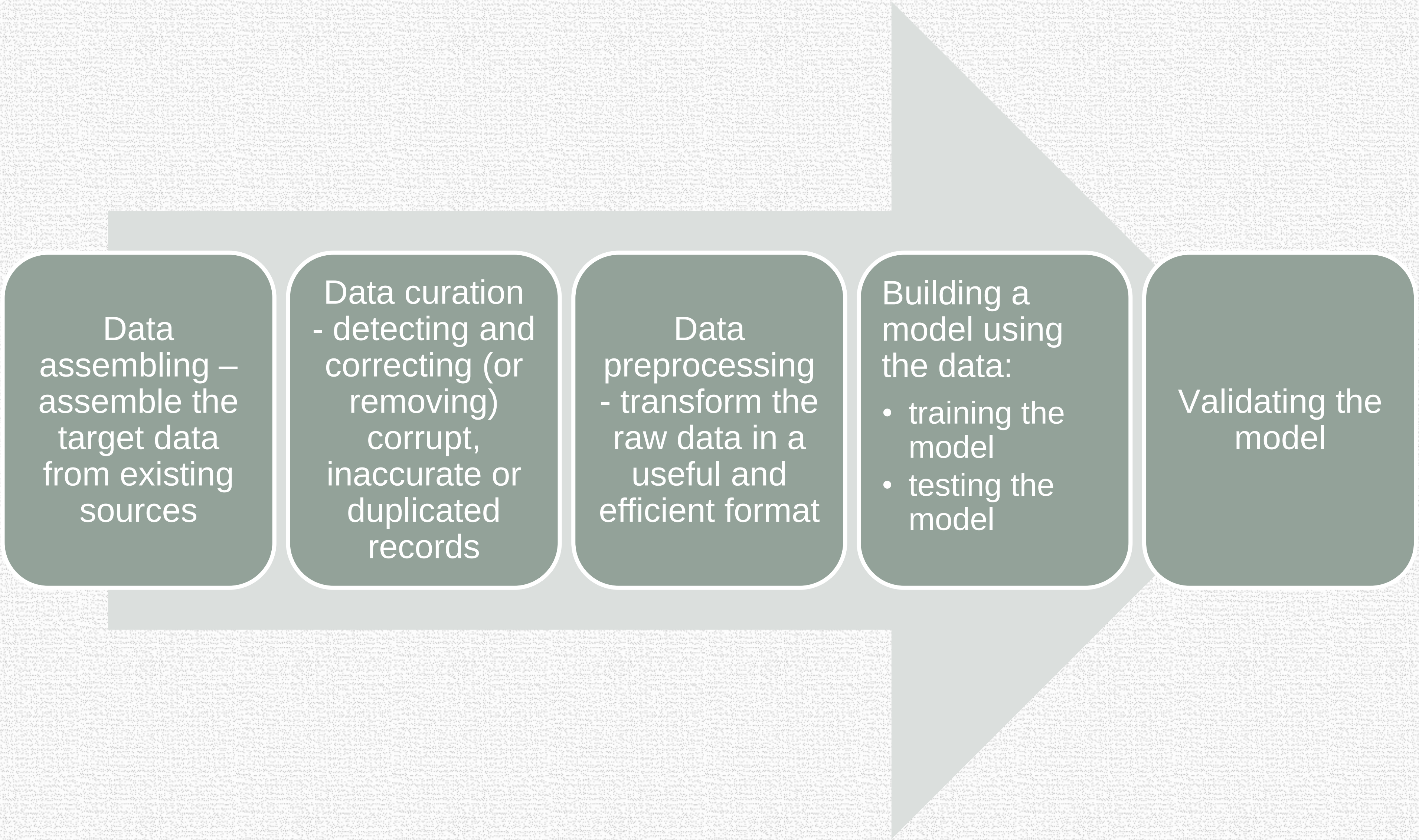
- **35% sequence identity?**
  - Karluss et al., *In Silico* Methods for Evaluating Human Allergenicity to Novel Proteins: International Bioinformatics Workshop Meeting Report, 23–24 February 2005, *Toxicological Sciences* 88(2), 307–310 (2005) doi:10.1093/toxsci/kfi277
  - Ronald Ross Watson and Victor R. Preedy Genetically Modified Organisms in Food Production, Safety, Regulation and Public Health. 2016, Academic Press

“Proteins homologues to the main birch pollen allergen Bet V 1 failed to bind or bind very poorly to IgE of patients allergic to birch pollen when sequence identity drops below 40%.”

Scheurer et al., Cross-reactivity and epitope analysis of Pru a 1, the major cherry allergen. *Molecular Immunology* 36 (1999) 155-167.

“In summary, we have shown that cross-reactive epitopes on Bet v 1 and related food allergens can be destroyed by single point mutations of amino acid residues in identical positions”

Main steps of *in silico* modeling



Data assembling – assemble the target data from existing sources

Data curation - detecting and correcting (or removing) corrupt, inaccurate or duplicated records

Data preprocessing - transform the raw data in a useful and efficient format

Building a model using the data:

- training the model
- testing the model

Validating the model

Databases with allergenic proteins

- WHO/IUIS Allergen Nomenclature Database - 1037 records
- AllergenOnline – 2233 records of proteins
- COMPARE - 2348 records of proteins
- Allergome – 4981 records
- AllerBase – 2187 records of proteins
- SDAP – 1526 records of proteins
- InformAll – 84 food records
- AllFam – 1042 records of families
- Immune Epitope Database (IEDB)
- Uniprot database

Data used in *in silico* methods

- Primary structure (amino acid sequence) of the proteins.
  - application of the FAO/WHO Codex Allimentarius rules
  - sequence alignment algorithms (FASTA, BLAST)
  - motif search
  - machine learning methods for sequence comparison
- Secondary and tertiary structure of the proteins
  - approaches that compare the 3D structures of protein
- Properties of the whole protein
- algorithms that uses physicochemical or biochemical properties of proteins
- Structure of known IgE epitopes of allergens
  - algorithms for mapping IgE epitopes of allergens

Performance evaluation

$$\boldsymbol{Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}}$$

$$\boldsymbol{Specificity = \frac{Thrue\ Negatives}{True\ Negatives + False\ Positives}}$$

- Sensitivity - How accurate the method predicts allergens.
- Specificity - How accurate the method predicts non-allergens.
- Accuracy - the proportion of correct predictions (both true positives and true negatives) among the total number of all cases examined
- Determination of both sensitivity and specificity requires a negative set i.e. set of non-allergens.

Selection of non-allergens

The available bioinformatics approaches have different methodologies, use diverse positive and negative sets of proteins and have widely varying validation procedures due to the lack of conventional criteria for non-allergenic proteins:

- The proteins non-labeled as allergen in a protein database e.g. Uniprot
- Human proteins
- Proteins from the same species as the allergens but lacking homology
- Proteins from species of common human food but lacking homology to allergens

Questions, challenges, opportunities

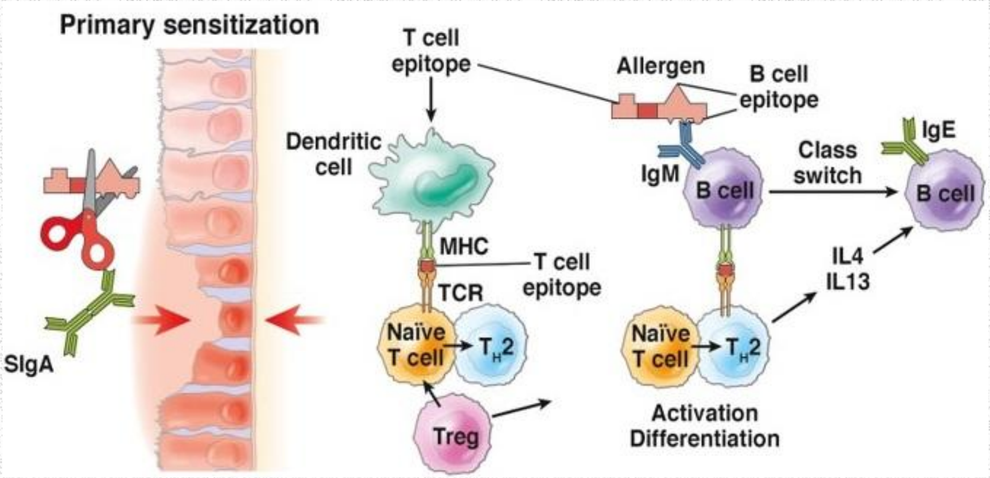
- The selection of data for positive (allergens) and negative (non-allergenic) sets is paramount to developing of reliable models. Data on the tertiary structure and conformational IgE epitopes of allergens are still insufficient.
- Since heterogeneity and imbalance in the data sets exist, the use of Matthews' correlation coefficient together with sensitivity, specificity, and accuracy would be more beneficial for assessing the predictive performance of in silico models.
- Searching for sequence similarity with known allergens is useful for detecting allergens that are evolutionarily similar. The challenge in the *in silico* methods for allergenicity assessment is to recognize allergenic proteins with no homology to existing proteins.

Questions, challenges, opportunities II

- The application of the methods of artificial intelligence could boost the allergenicity prediction. Unfortunately, the amount of existing data is still a limiting factor for their application.
- Application of methods of other scientific domains e.g. digital signal processing could be also an opportunity to improve the methods for the *in silico* assessment of allergenicity.

Sensitisation - main processes

- Is it possible to assess the sensitisation potential of a protein *in silico*?



Valenta et al. Food allergies: the basics. *Gastroenterology* 2015;148:1120–1131. (Licensed under CC BY 4.0)

- Digestion of protein by GIT proteases
- **Antigen presentation**
- **Naïve T cells differentiation to T_H2 cell**
- **Activation of B cells by T_H2 cells**
- **Proliferation and differentiation of activated B cells**

MHC class II basics

- In humans, the MHC class II protein complex is encoded by the Human Leukocyte Antigen (HLA) gene complex.
- There are 3 major MHC class II loci encoded by HLA: HLA-DP, HLA-DQ, and HLA-DR.
- Human MHC genes are highly polymorphic, i.e. each locus has many alleles.
- Not all MHC class II binders are T-cell epitopes, but all T-cell epitopes are MHC class II binders.

HLA class II basics

- The structure of the binding cleft on the HLA class II proteins limits the length of the binding peptide core to nine residues.
- In order to be recognized as an antigen by the immune system, the sequence of a given protein should contain at least one nonamer (9 amino acid long peptide) that binds to an HLA class II protein.
- Due to their polymorphism, HLA proteins bind to a large repertoire of antigen peptides.

Why HLA binding?

- It is involved in almost every key event in the sensitisation process
- There are evidences from genome-wide association studies that HLA loci are involved in allergen sensitisation.
- HLA binders (potential T-cell epitopes) are linear, in contrast to B-cell epitopes.
- Linked recognition - B cell activation by a helper T cell that responds to the same, or physically associated, antigen.

Bønnelykke K, Matheson MC, Pers TH, et al. Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. *Nat Genet*. 2013;45(8):902-906. doi:10.1038/ng.2694

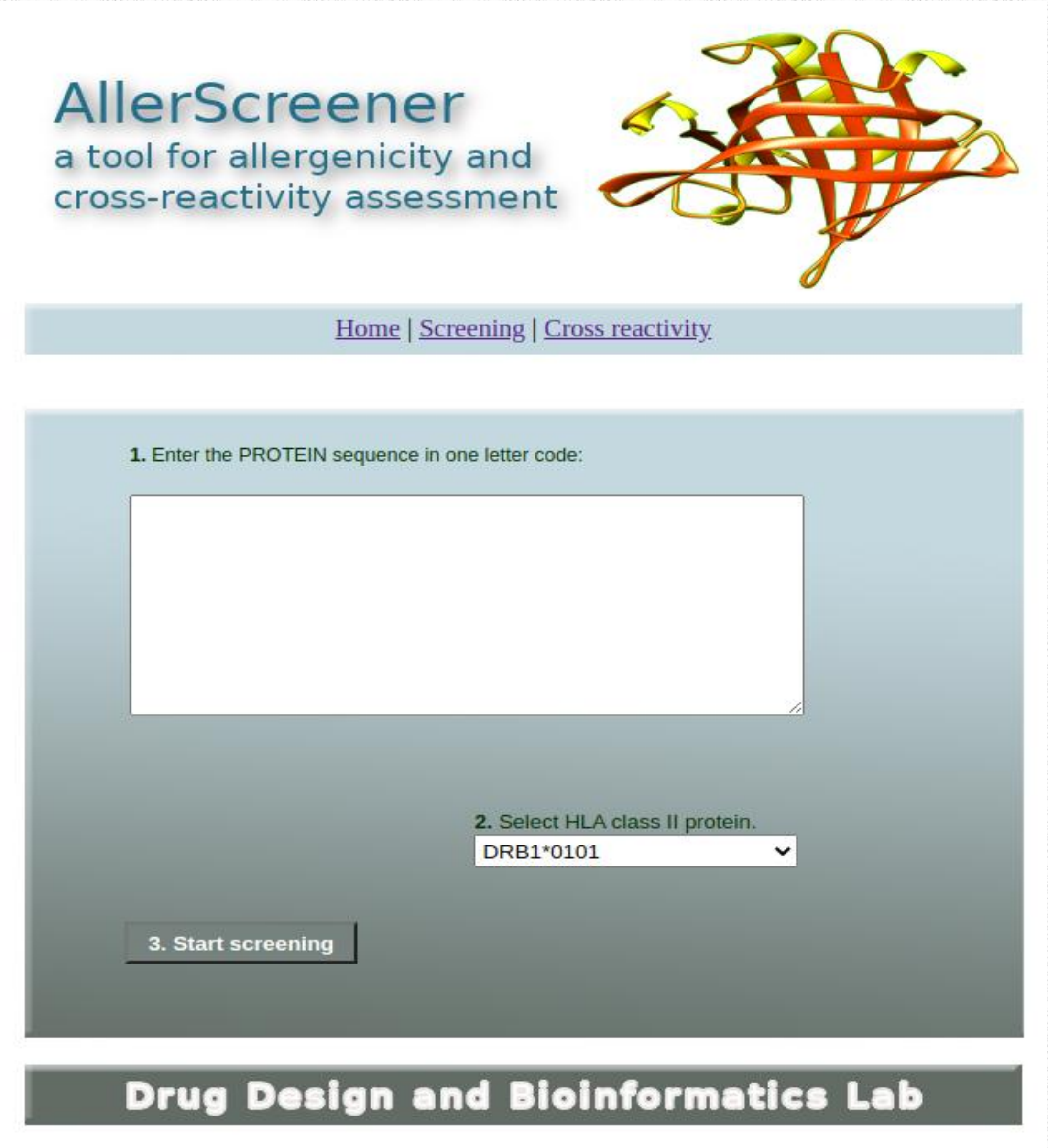
AllerScreener – a tool for allergenicity and cross-reactivity assessment

- Collect allergenic proteins with known sequences and species from four databases:
    - Allergome
    - COMPARE
    - Allerbase
    - AllergenOnline
    - Collect sequence data for the proteins in Allergome from the Uniprot database. (January 2020)
- Remove duplicated proteins and proteins with unidentified amino acids in their sequence.
- Collect data for known T cell and B cell epitopes from the IEDB

AllerScreener II

- Build a database with HLA binders (nonamers) to the most common human alleles for each of the allergenic proteins predicted by two different *in silico* tools:
  - EpiTOP - based on the data for the known HLA binders to 12 DRB1, 5 DQ and 7 DP alleles and the methods of the ligand-based drug design. Predicts binding affinity.
  - EpiDOCK – based on the data for the 3D structure of 12 HLA DRB1, 6 DQ and 5 DP alleles proteins and the methods of the structure-based drug design
- The nonamers from allergenic proteins are accepted to be binders only if the binding affinity (pIC_{50}) of the nonamers predicted by EpiTOP is higher than 5.3 ($IC_{50} < 5000$nM) and is assessed by EpiDOCK as binder to the same HLA allele.

AllerScreener for allergen screening



AllerScreener
a tool for allergenicity and
cross-reactivity assessment

Home | Screening | Cross reactivity

1. Enter the PROTEIN sequence in one letter code:

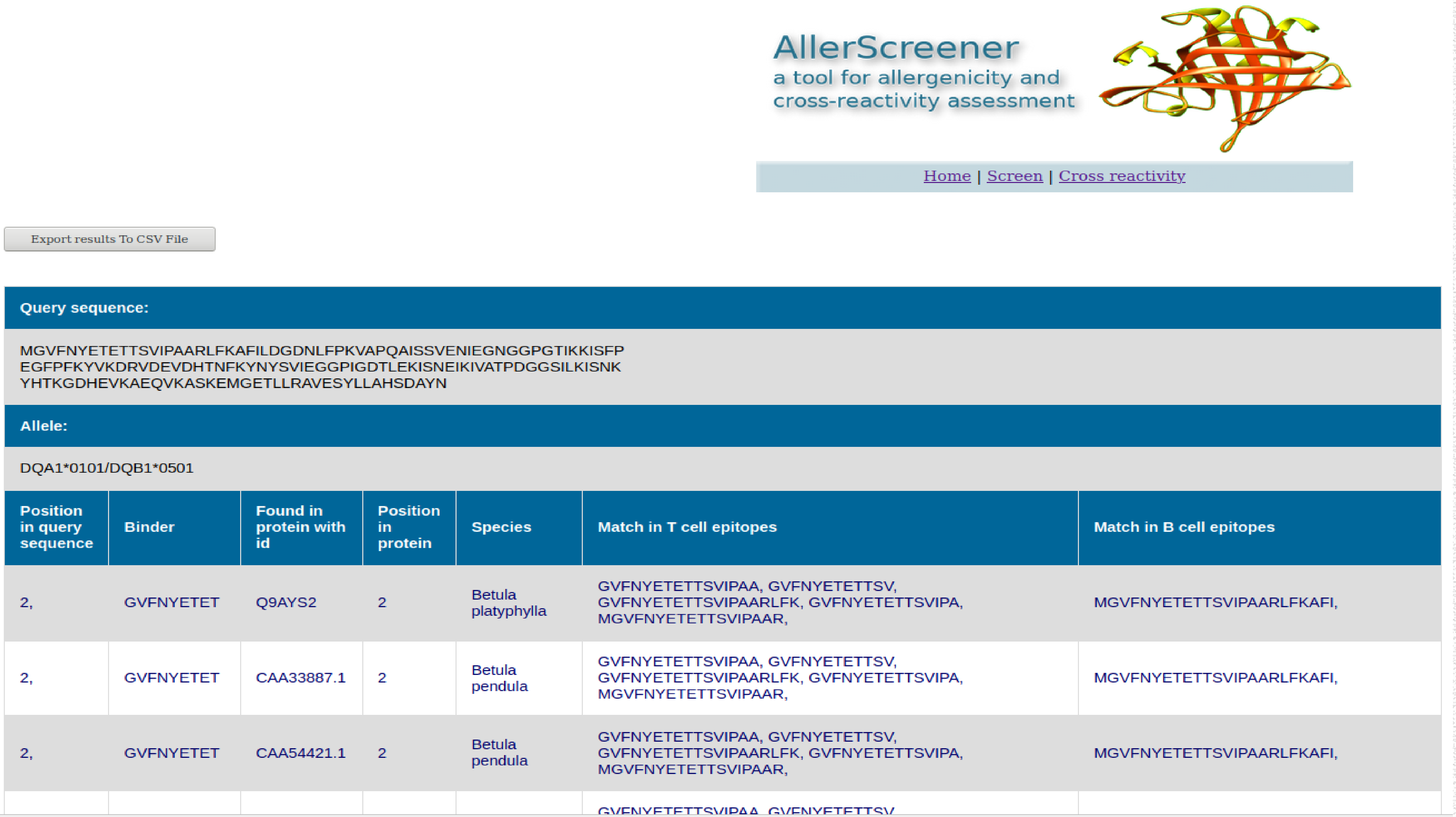
2. Select HLA class II protein.
DRB1*0101

3. Start screening
Drug Design and Bioinformatics Lab

Screening the database:

- The novel protein is chopped into overlapping nonamers.
- Searching for the sequence identity of each nonamer with the predicted binders to the selected allele in the database.
- searching for matches with T-cell and/or B-cell epitopes in the database for each nonamer that matches an HLA binder in the database.

Allergen screening - output



AllerScreener
a tool for allergenicity and
cross-reactivity assessment

Query sequence:

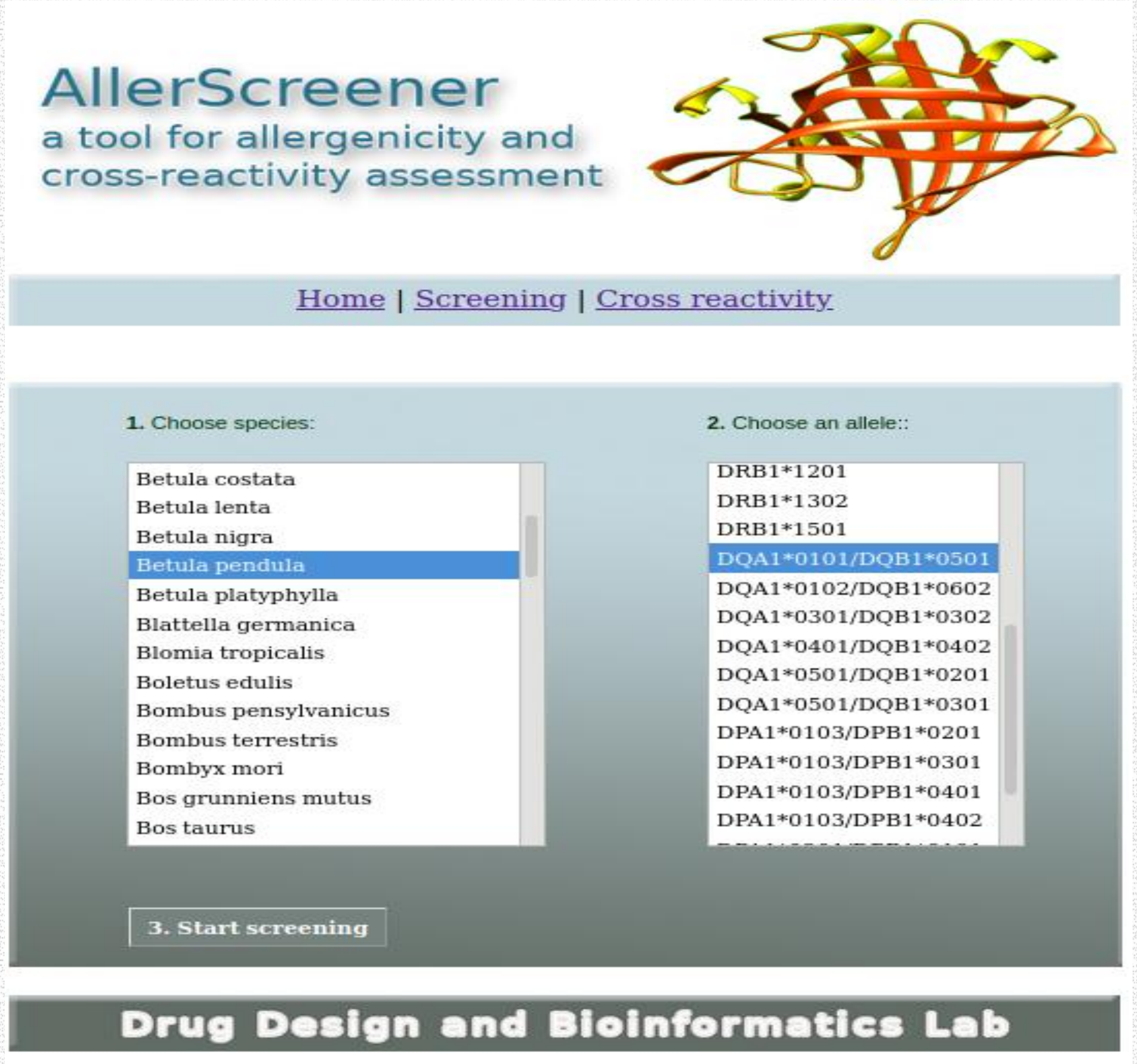
MGVFNYETETTSVIPAARLFKAFILDGDNLFPKVAPQAISSVENIEGNGGPGTIKKISFP
EGFPFKYVKDRVDEVDHTNFKYNYSVIEGGPIGDTLEKISNEIKIVATPDGGSILKISNK
YHTKGDHEVKAEQVKASKEMGETLLRAVESYLLAHSDAYN

Allele:

DQA1*0101/DQB1*0501

Position in query sequence	Binder	Found in protein with id	Position in protein	Species	Match in T cell epitopes	Match in B cell epitopes
2,	GVFNYETET	Q9AYS2	2	Betula platyphylla	GVFNYETETTSVIPAA, GVFNYETETTSV, GVFNYETETTSVIPAARLFK, GVFNYETETTSVIPA, MGVFNYETETTSVIPAAR,	MGVFNYETETTSVIPAARLFKAFI,
2,	GVFNYETET	CAA33887.1	2	Betula pendula	GVFNYETETTSVIPAA, GVFNYETETTSV, GVFNYETETTSVIPAARLFK, GVFNYETETTSVIPA, MGVFNYETETTSVIPAAR,	MGVFNYETETTSVIPAARLFKAFI,
2,	GVFNYETET	CAA54421.1	2	Betula pendula	GVFNYETETTSVIPAA, GVFNYETETTSV, GVFNYETETTSVIPAARLFK, GVFNYETETTSVIPA, MGVFNYETETTSVIPAAR,	MGVFNYETETTSVIPAARLFKAFI,
					GVFNYETETTSVIPAA, GVFNYETETTSV	

AllerScreener for cross-reactivity search



AllerScreener
a tool for allergenicity and
cross-reactivity assessment

Home | Screening | Cross reactivity

1. Choose species:

- Betula costata
- Betula lenta
- Betula nigra
- Betula pendula
- Betula platyphylla
- Blattella germanica
- Blomia tropicalis
- Boletus edulis
- Bombus pensylvanicus
- Bombus terrestris
- Bombyx mori
- Bos grunniens mutus
- Bos taurus

Selected: Betula pendula

2. Choose an allele::

- DRB1*1201
- DRB1*1302
- DRB1*1501
- DQA1*0101/DQB1*0501
- DQA1*0102/DQB1*0602
- DQA1*0301/DQB1*0302
- DQA1*0401/DQB1*0402
- DQA1*0501/DQB1*0201
- DQA1*0501/DQB1*0301
- DPA1*0103/DPB1*0201
- DPA1*0103/DPB1*0301
- DPA1*0103/DPB1*0401
- DPA1*0103/DPB1*0402

Selected: DQA1*0101/DQB1*0501

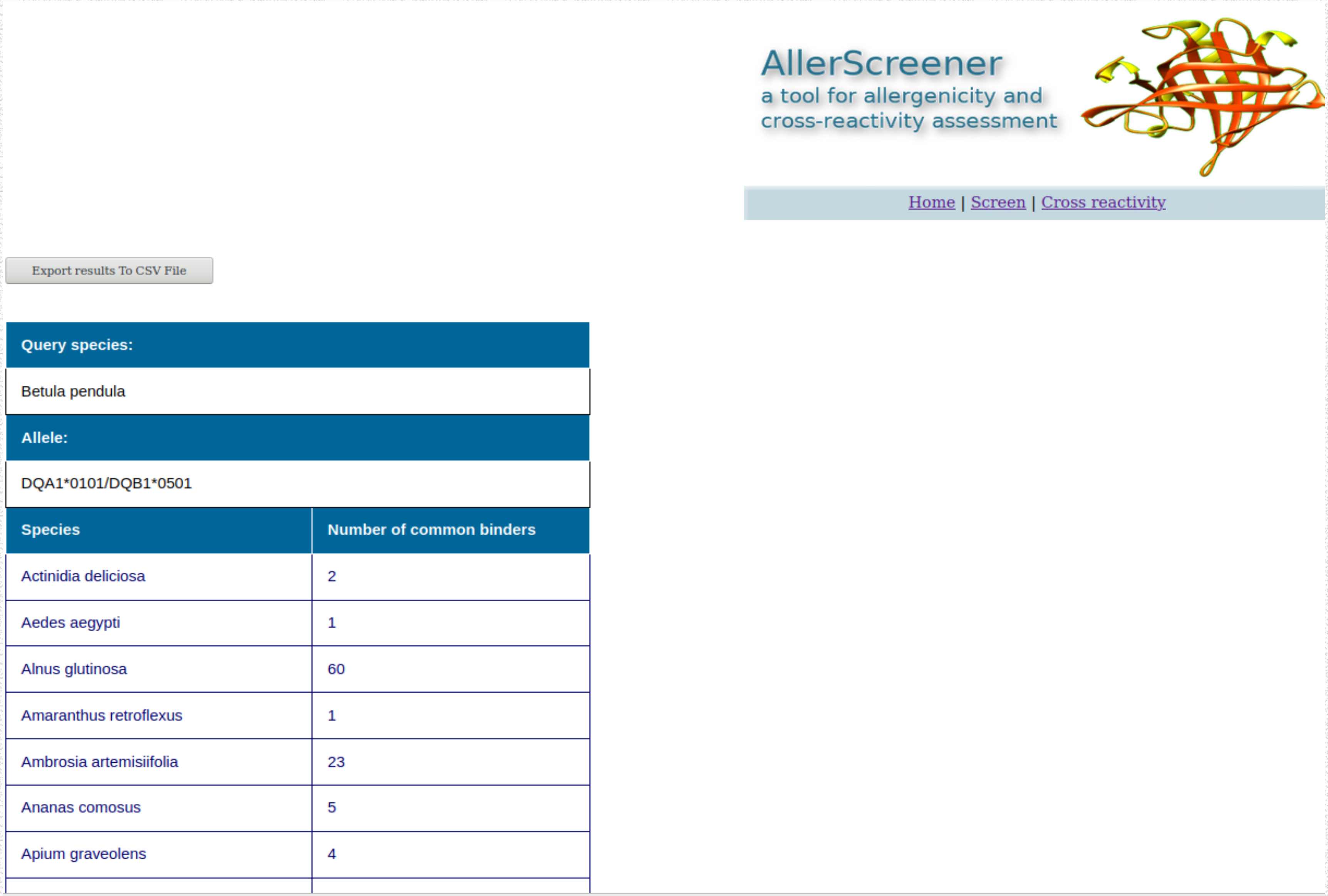
3. Start screening

Drug Design and Bioinformatics Lab

Cross-reactivity search:

- A crossed search for common binders between the query species and all of the species in the database for the chosen allele.

Cross-reactivity search - output



AllerScreener

a tool for allergenicity and cross-reactivity assessment

Query species:	
Betula pendula	
Allele:	
DQA1*0101/DQB1*0501	
Species	**Number of common binders**
Actinidia deliciosa	2
Aedes aegypti	1
Alnus glutinosa	60
Amaranthus retroflexus	1
Ambrosia artemisiifolia	23
Ananas comosus	5
Apium graveolens	4

Future plans

- Update and upgrade the models implemented in the web servers AllerTOP and AllergenFP
- Update and upgrade the models implemented in web servers EpiTOP and EpiDOCK
- Derive models of the GIT digestion and the endosome cleavage of the antigenic proteins
- Increase the functionality of AllerScreener
- Combined application of derived models for protein digestion, endosome cleavage, Allerscreener, AllerTOP and AllergenFP to completely cover both sensitisation and elicitation phase of allergy.

Conclusions

- The up-to-date *in silico* approaches for allergenicity assessment are actually approaches for assessment of IgE cross-reactivity of a new protein and the known allergens.
- Precise and reliable data sets are of utmost importance for training and validation of the *in silico* models.
- A dataset of non-allergenic proteins to be used as a standard negative set would improve the performance of the computational models.
- The application of the standard statistical parameters for classification will help for the comparison and better assessment of the computational methods for allergenicity prediction.
- Models for prediction of binding to HLA proteins could be used as a step in a pipeline of *in silico* tools for modeling the processes in the phases of the allergic reactions.

DRUG DESIGN AND BIOINFORMATICS LAB

FACULTY OF PHARMACY
MEDICAL UNIVERSITY OF SOFIA
BULGARIA

ФАРМАЦЕВТИЧЕН ФАКУЛТЕТ
СОФИЯ

- Prof. Irini Doytchinova, DSci
- Assoc. Prof. Zvetanka Zhivkova, PhD
- Assoc. Prof. Ivan Dimitrov, PhD
- Assist. Prof. Mariyana Atanasova, PhD
- Assist. Prof. Iva Valkova, PhD
- Danislav Spasov, PhD, PostDoc, MSCA Fellow
- Nikolet Doneva, MPharm, PhD student
- Stanislav Sotirov, MPharm, PhD student