上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Evaluation and Integration of Existing Methods for Computational Prediction of Allergens

Jing Wang

ICoVax 2012 Workshop Tongji University October 13, 2012

What's allergen?



Others

Food allergens

Allergens

Venom/salivary

Contact allergens



Aero allergens

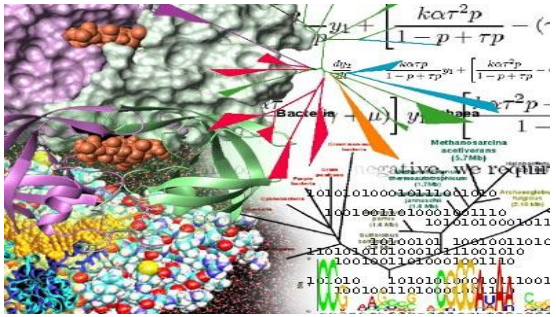

Hazards of allergens

- 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

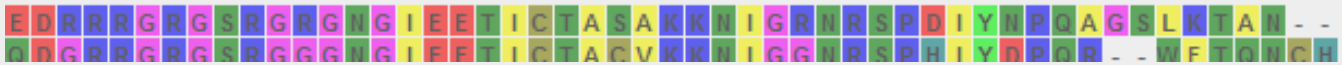
- Bioinformatics prediction methods
  - Sequence-based method *(FAO/WHO, 2001)*
  - Motif-based method *(Stadler, M. B. et al., FASEB J, 2003)*
  - SVM-based method *(Saha S et al., NAR, 2006)*



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

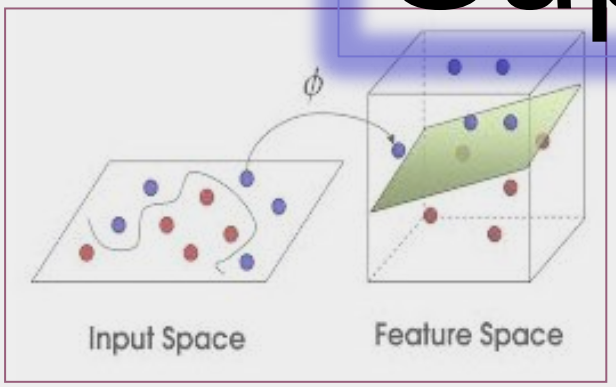
Bioinformatics methods

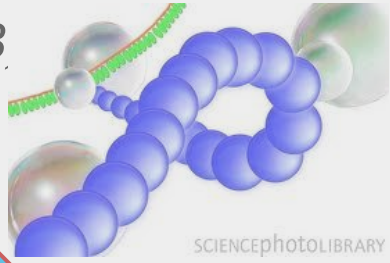
- Bioinformatics prediction methods
  - Sequence-based method *(FAO/WHO, 2001)*

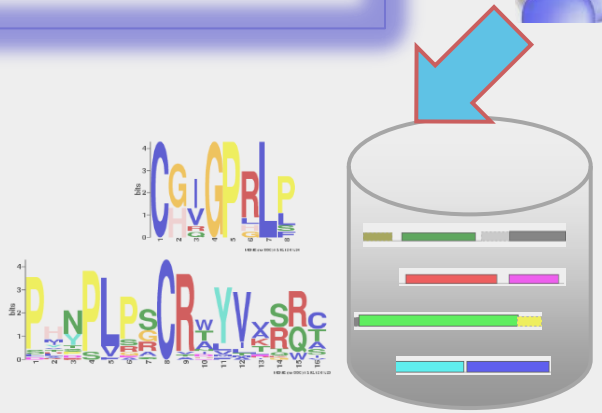


  - Motif-based method *(Stadler, M. B. et al., FASEB J, 2003)*
  - SVM-based method *(Saha S et al., NAR, 2006)*

Superior?







Research Objectives

Methods Evaluation

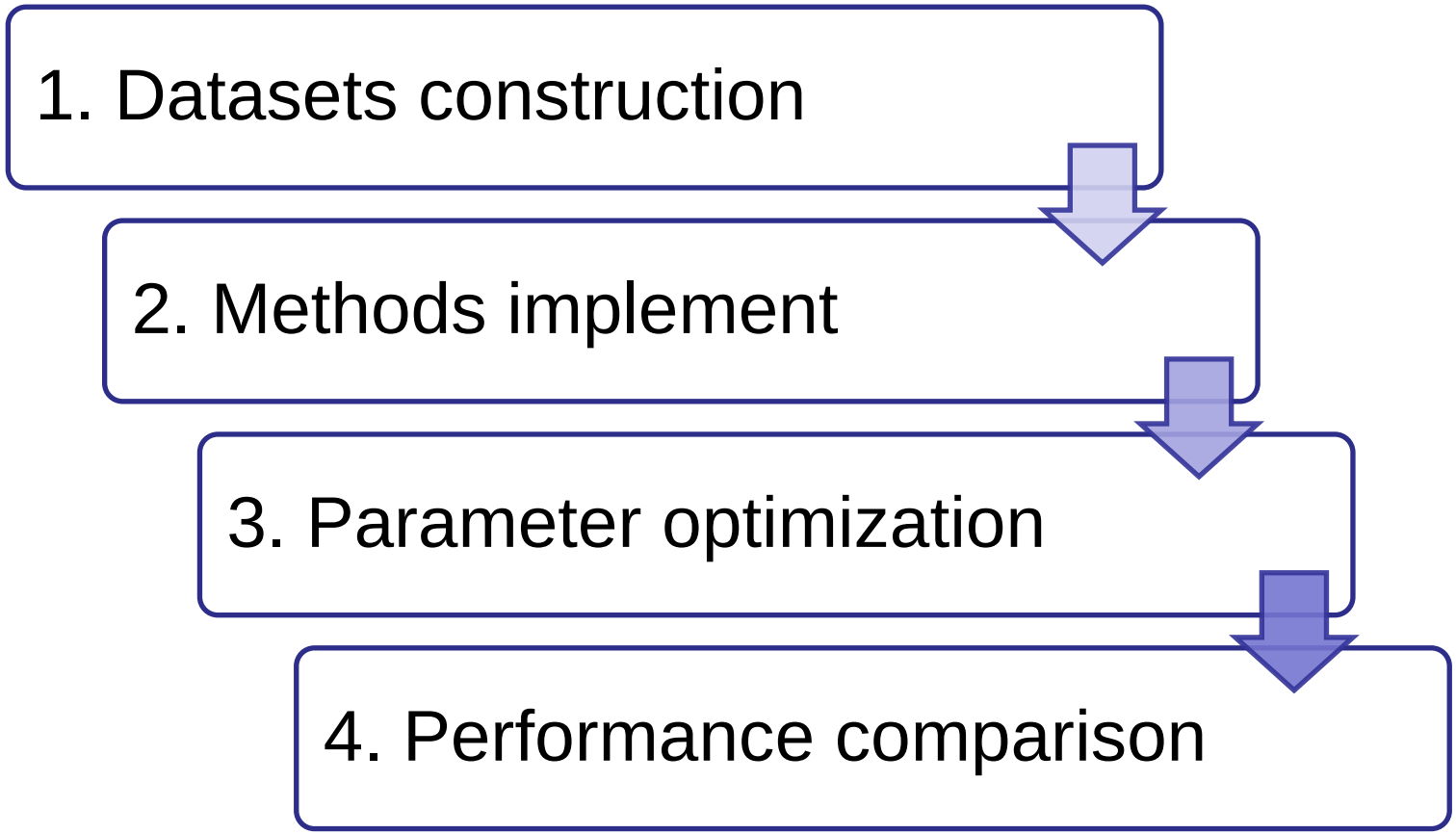
- Compare the performance of a variety of computational methods for allergen prediction
- Find the cons and pros of each method, and perform parameter optimization

Methods integration

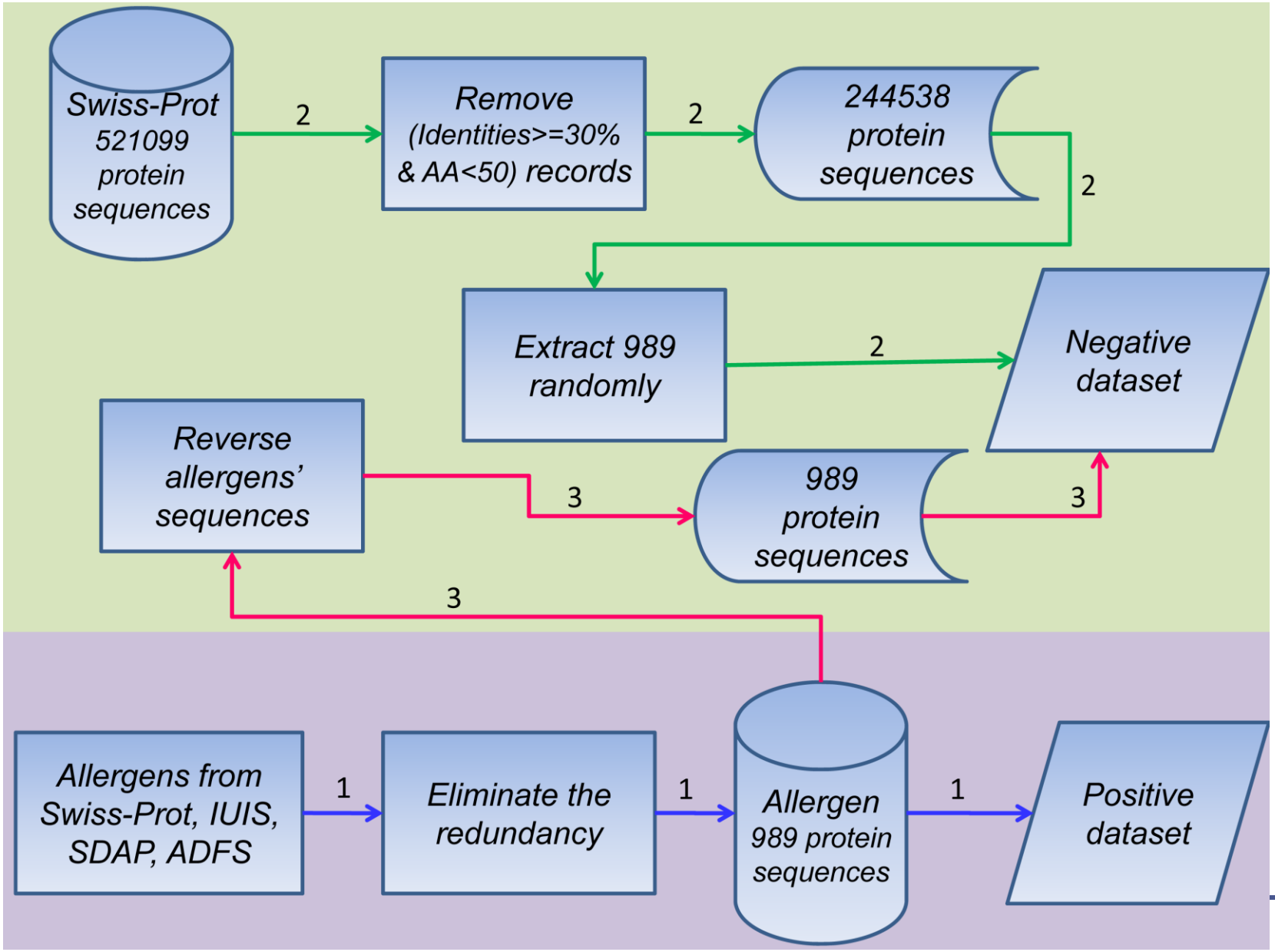
- Database search for all known allergen
- One-stop prediction for the protein allergenicity
  - Single protein prediction
  - Batch prediction

Method Evaluation

- Workflow of the evaluation

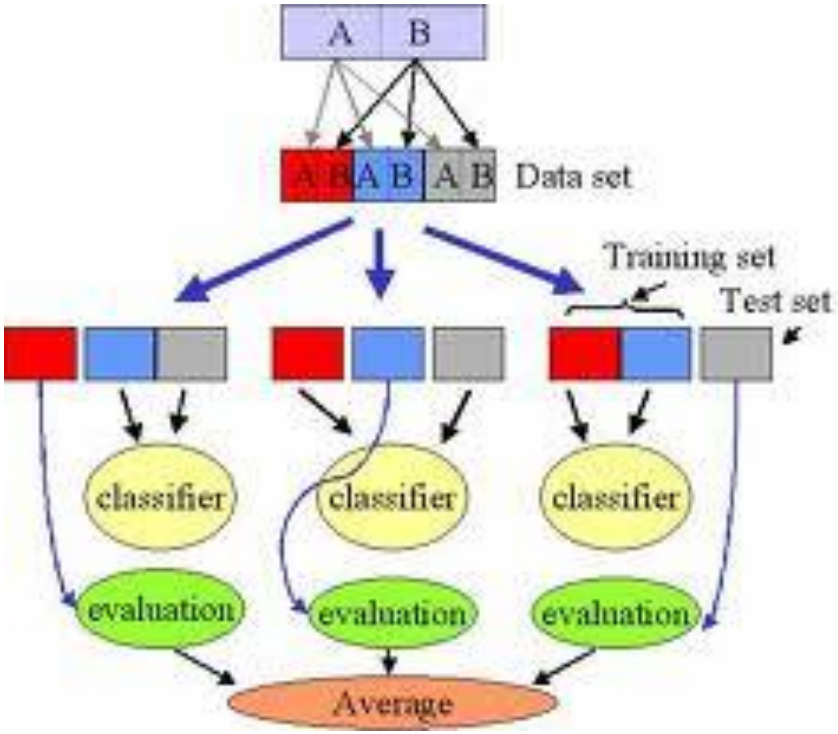


1. Datasets construction



1. Datasets construction

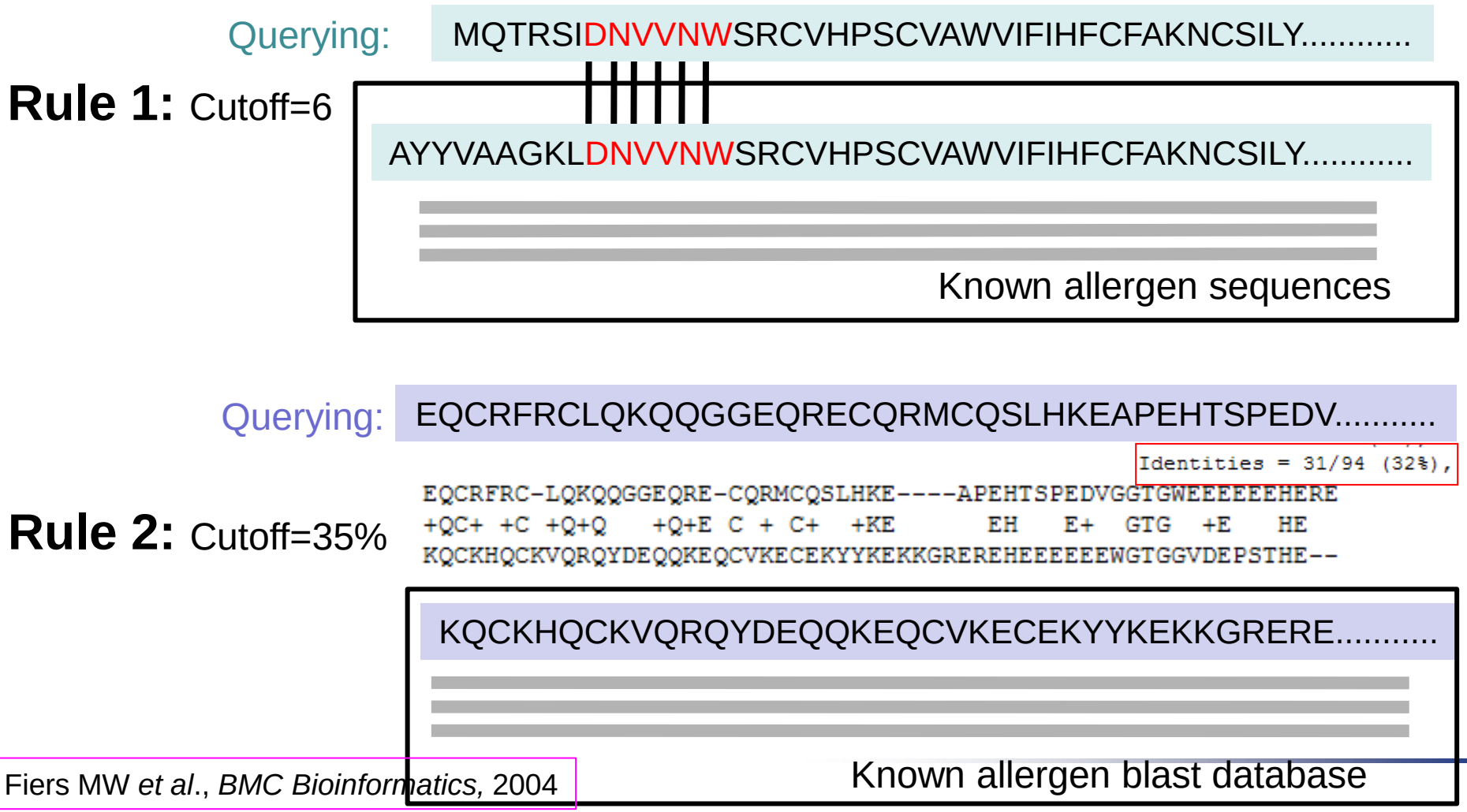
- Ten-fold cross validation



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

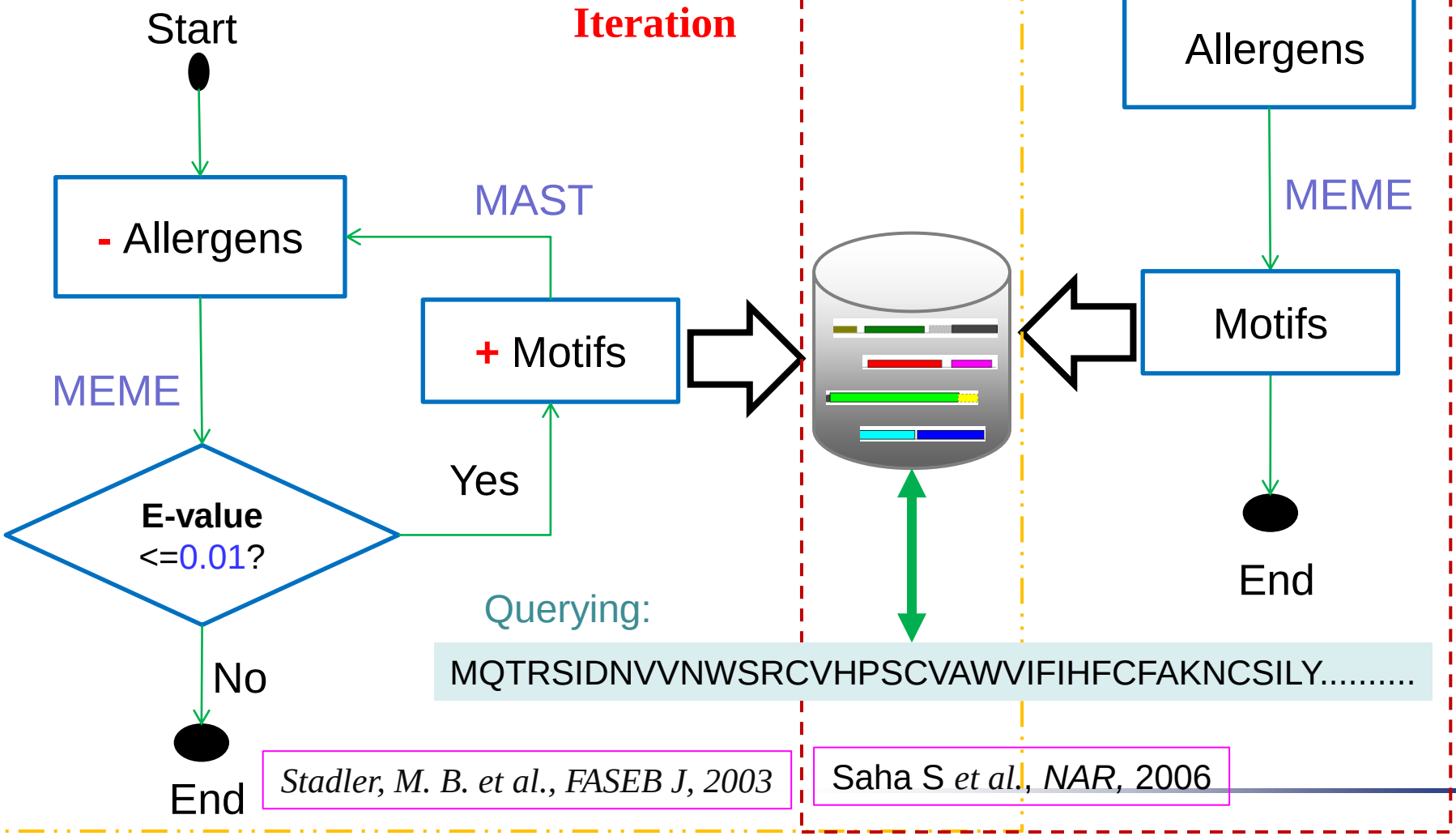
- Sequence-based method (FAO/WHO criteria)



Fiers MW *et al*., *BMC Bioinformatics*, 2004

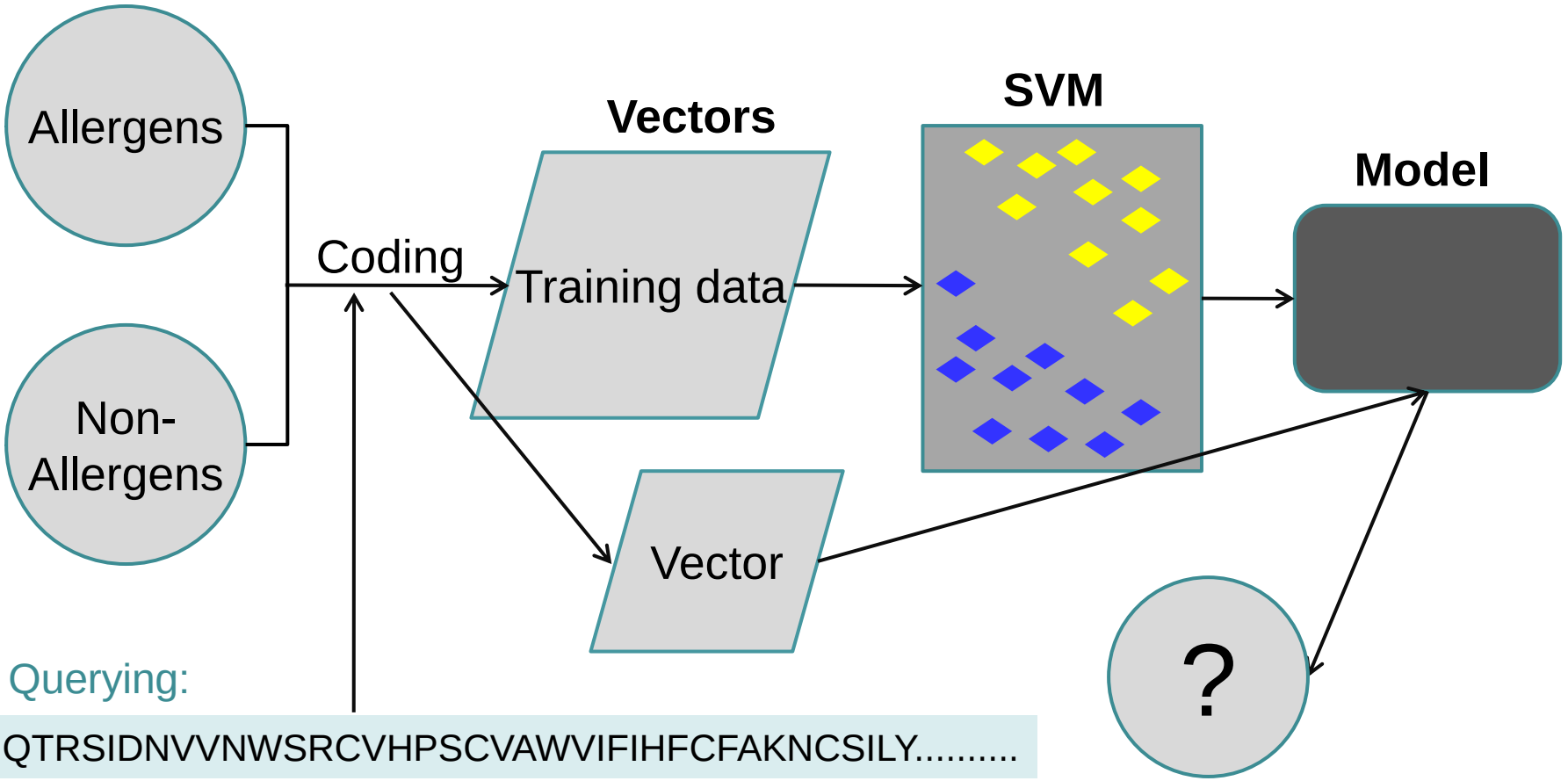
2. Methods implement

Motif-based method



2. Methods implement

- SVM-based method



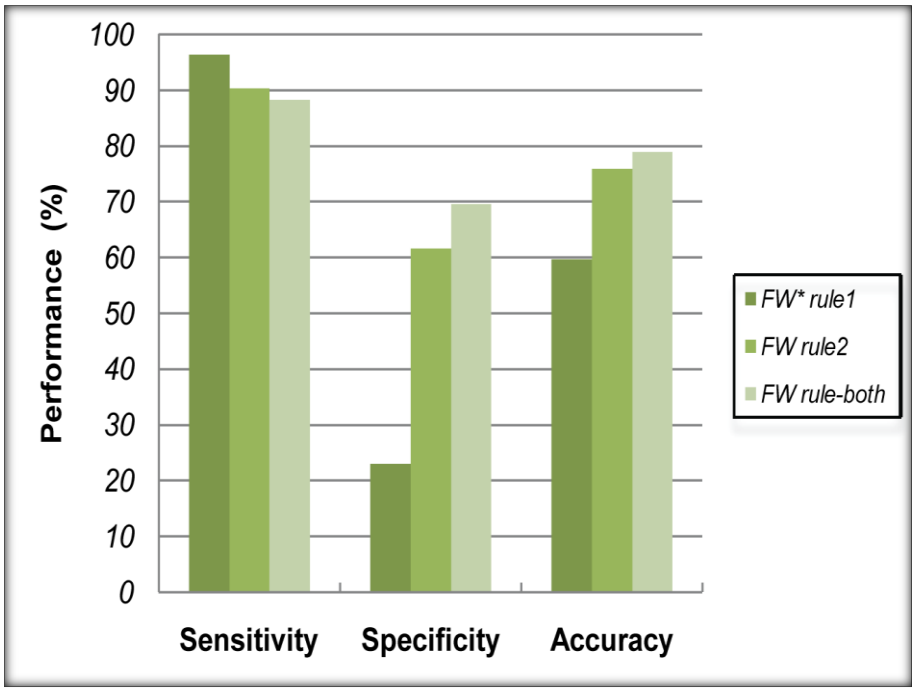
Amino Acid Composition

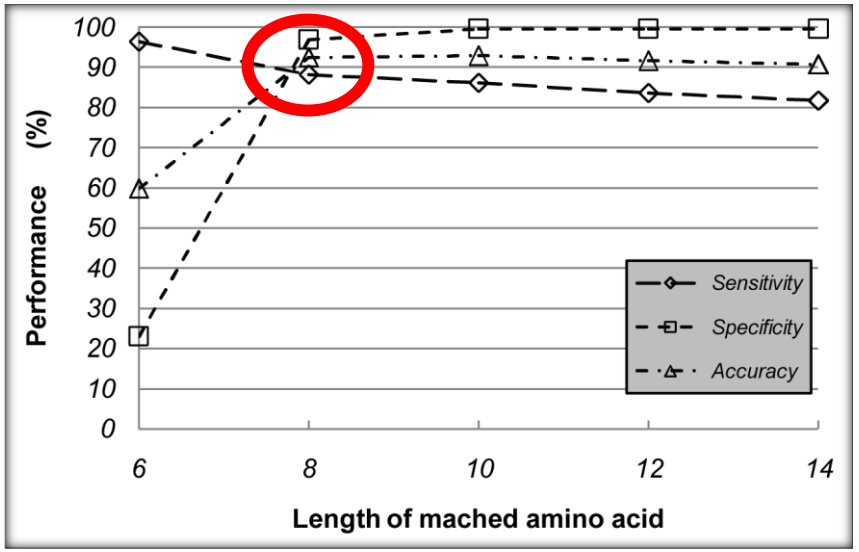
$$\text{Fraction of amino acid } i = \frac{\text{total number of amino acids } (i)}{\text{total number of amino acids in protein}}$$

Saha S et al., *Nucleic Acids Research,* 2006

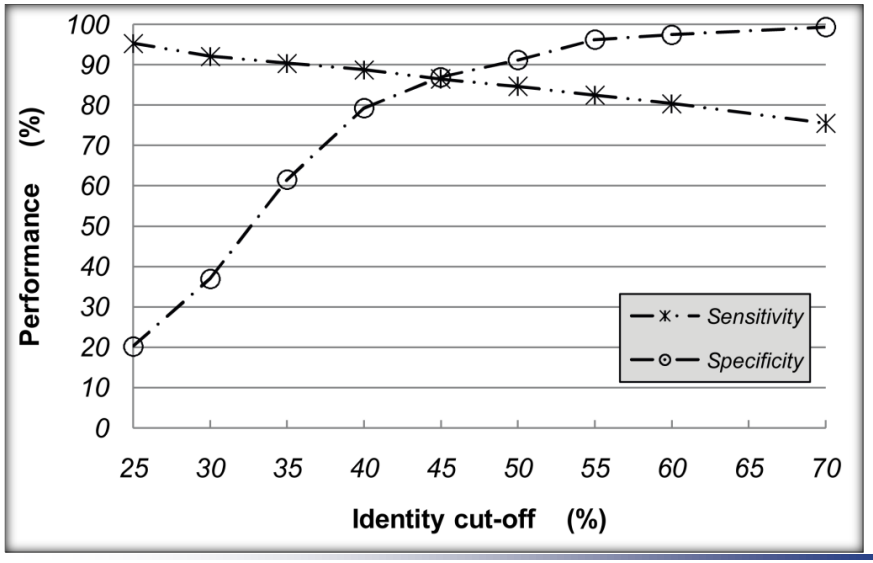
3. Evaluation results

- FAO/WHO





Rule 1



Rule 2

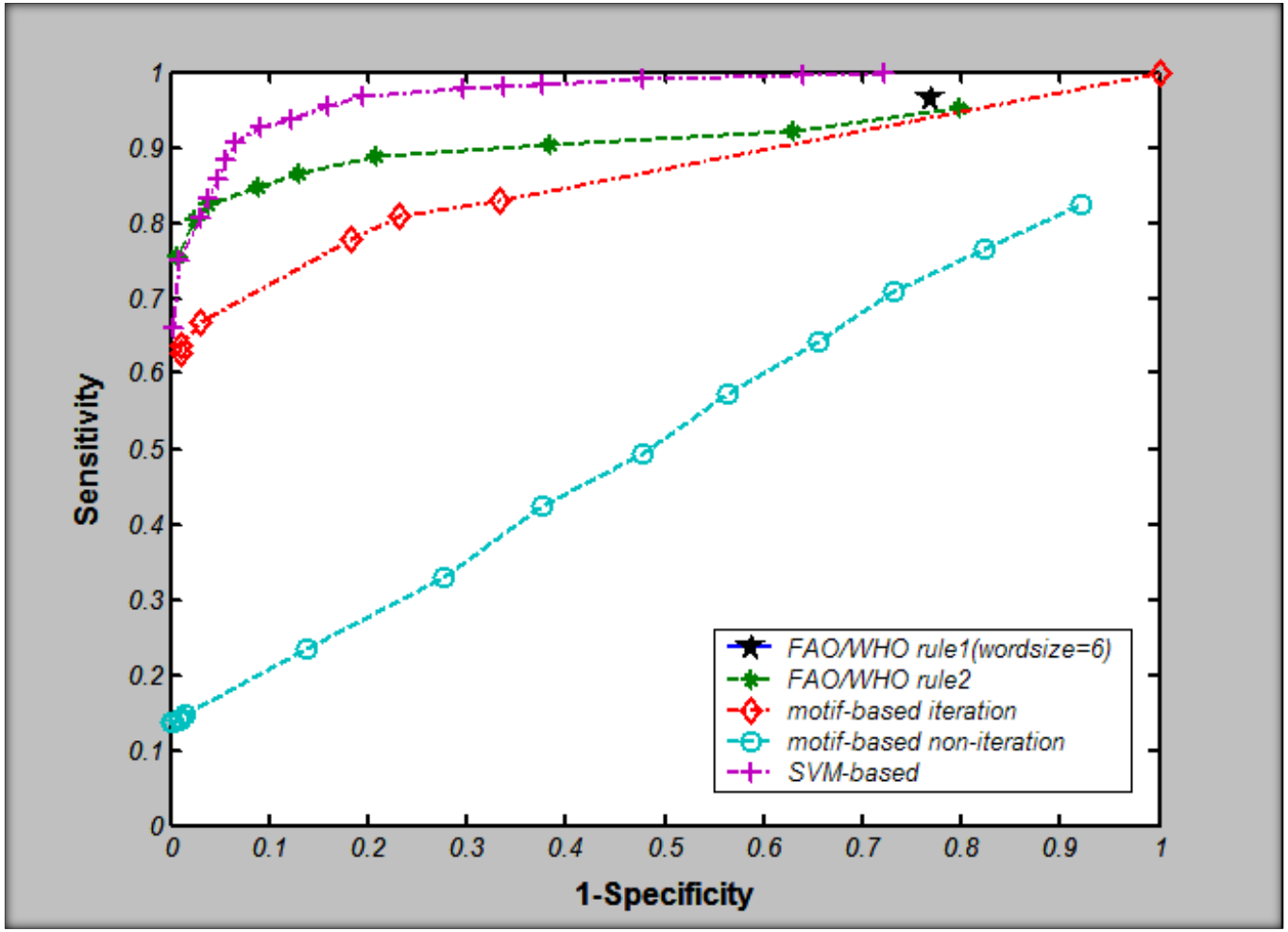
3. Evaluation results

- 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



4. Methods comparison

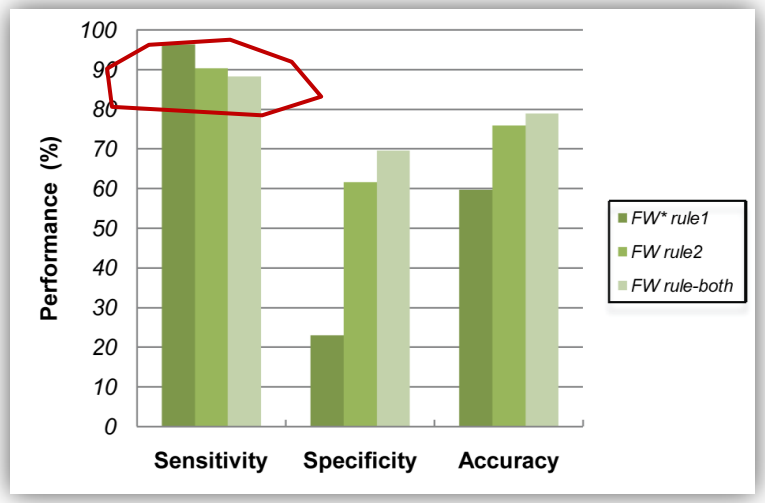
- Methods comparison
  - Time complexity

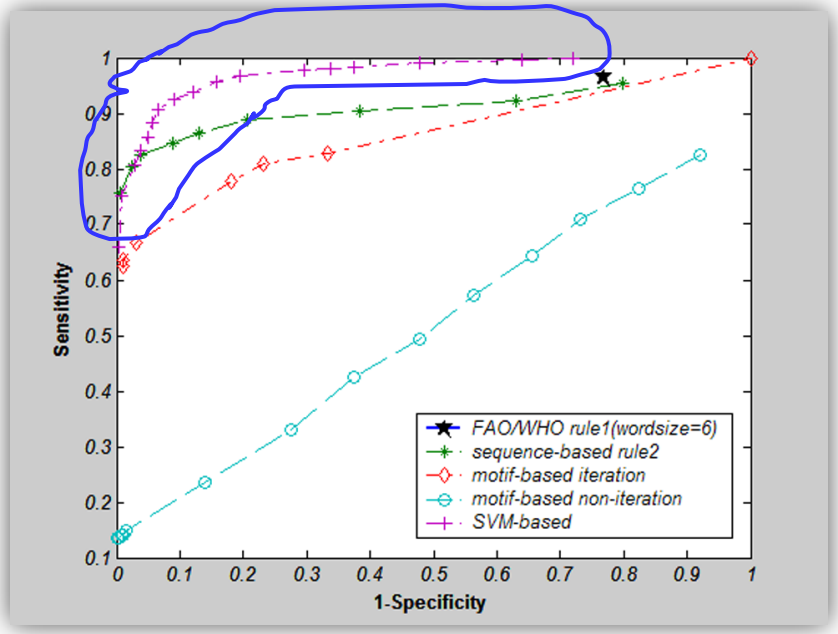
Approaches	F/W rule 1	F/W rule2	Motif-based	SVM-based
Time (ms*)	15940	58640	87	10

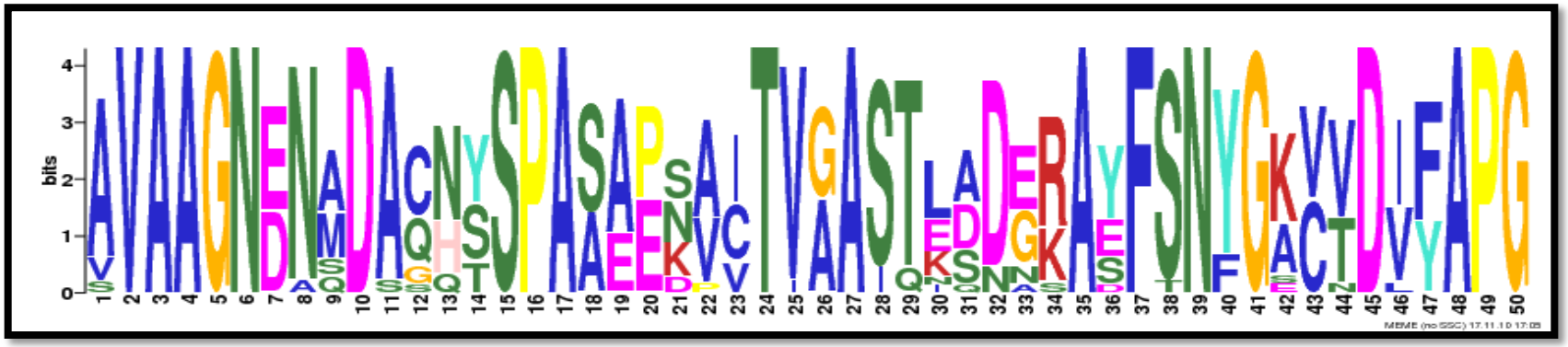
ms means millisecond

Methods integration

- Why integration?







Web-based application

proAP Protein Allergenicity Prediction

Home | Site map | Contact us

Navigation

- Introduction
- Allergen search
- Allergenicity prediction
- Batch prediction

Statistics

- Allergen: 1096
- Allergen category: 13
- Species: 249
- Prediction method: 4

Links

Swiss-Prot/TrEMBL
NCBI-Entrez
SDAP-Structural Database of Allergenic Proteins
BLAST
FASTA
MEME/MAST

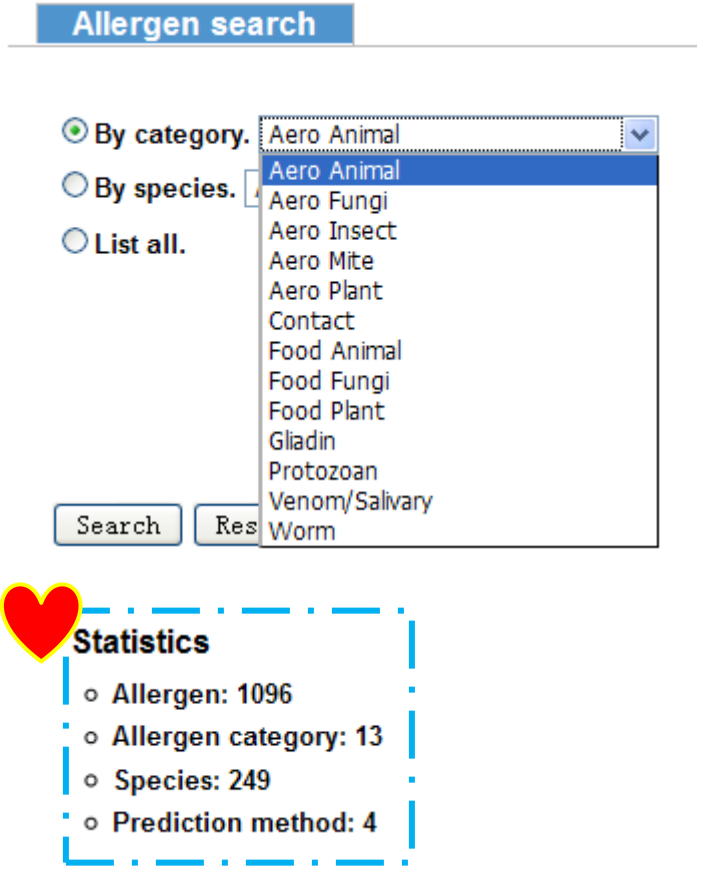
Home

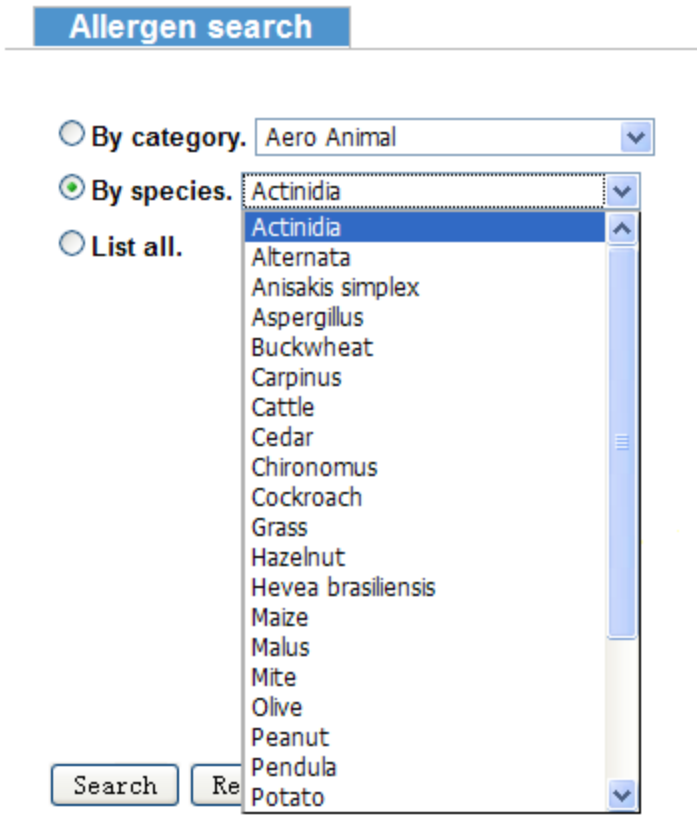
Welcome to proAP !

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

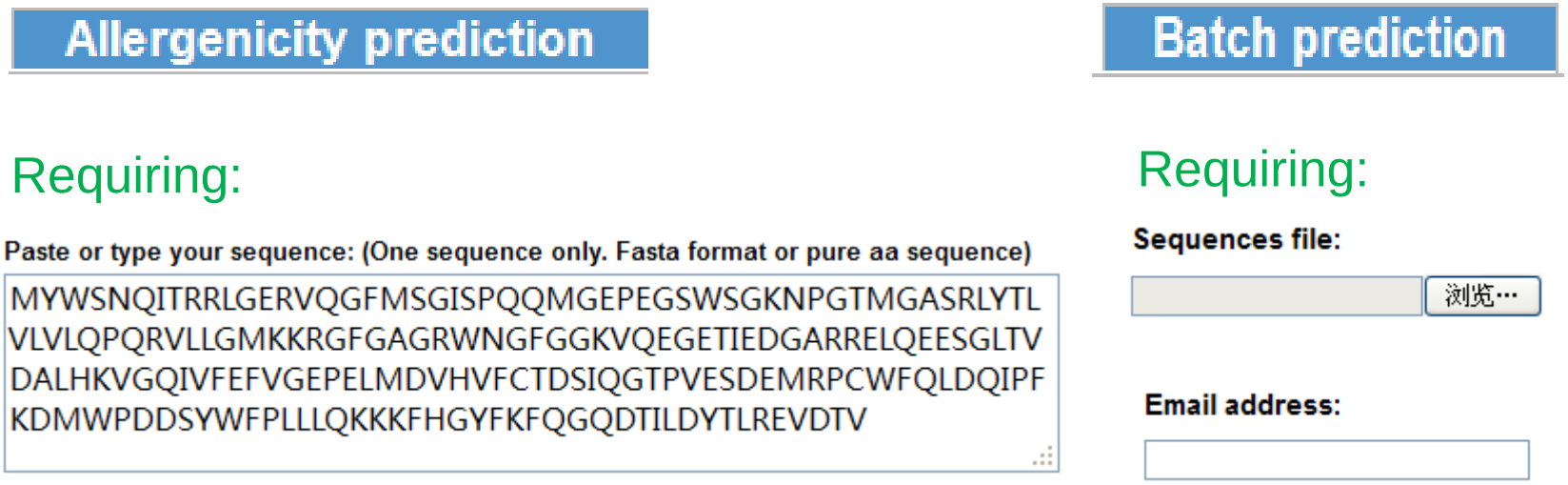
- Allergen search





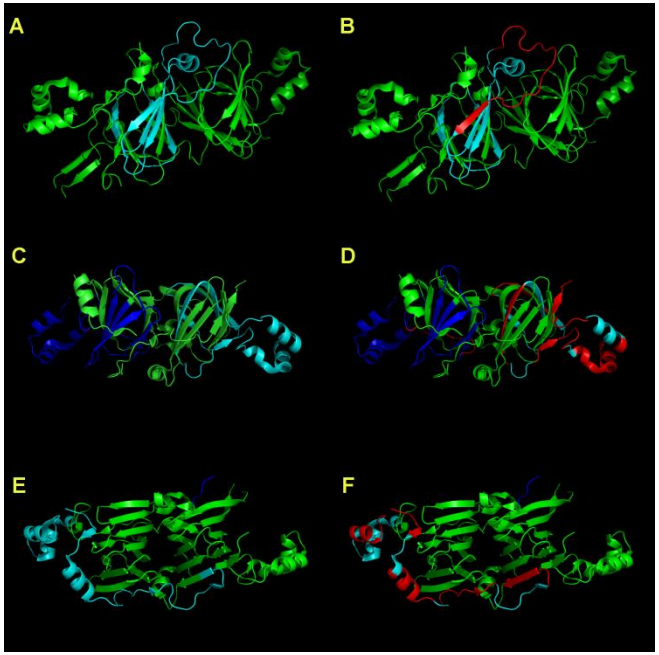
Main modules

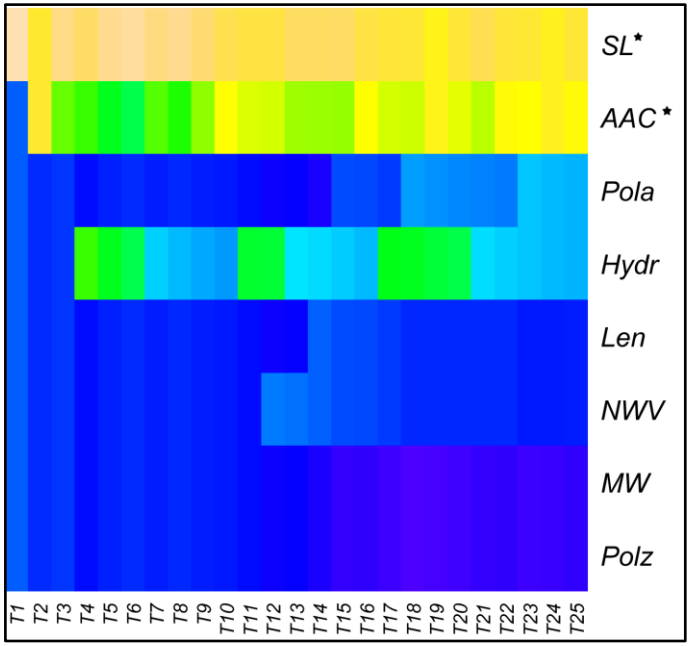
- Allergenicity prediction

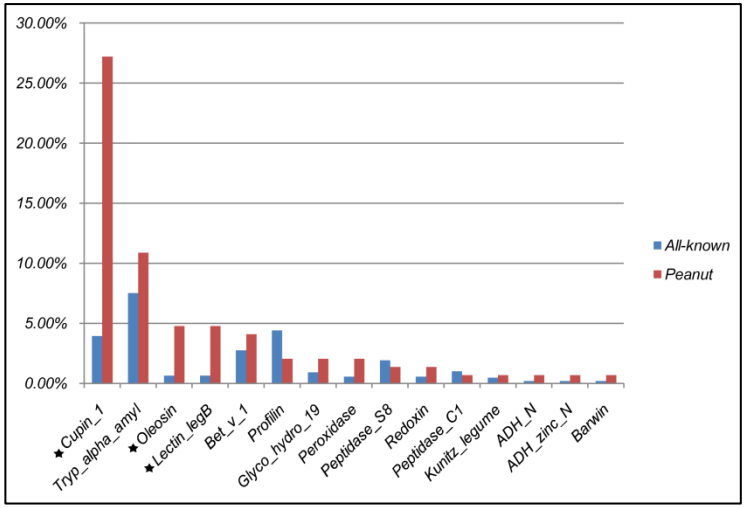


Next step …

- Key features for allergenicity
- Family preference
- Specific structures







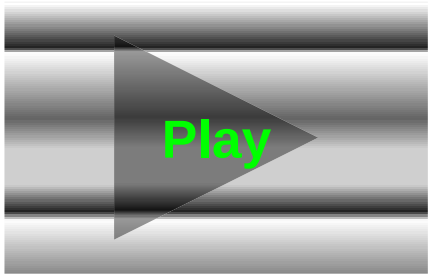
Acknowledgments

- Dr. Jing Li
- Prof. Dabing Zhang
- Yabin Yu
- Yunan Zhao

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

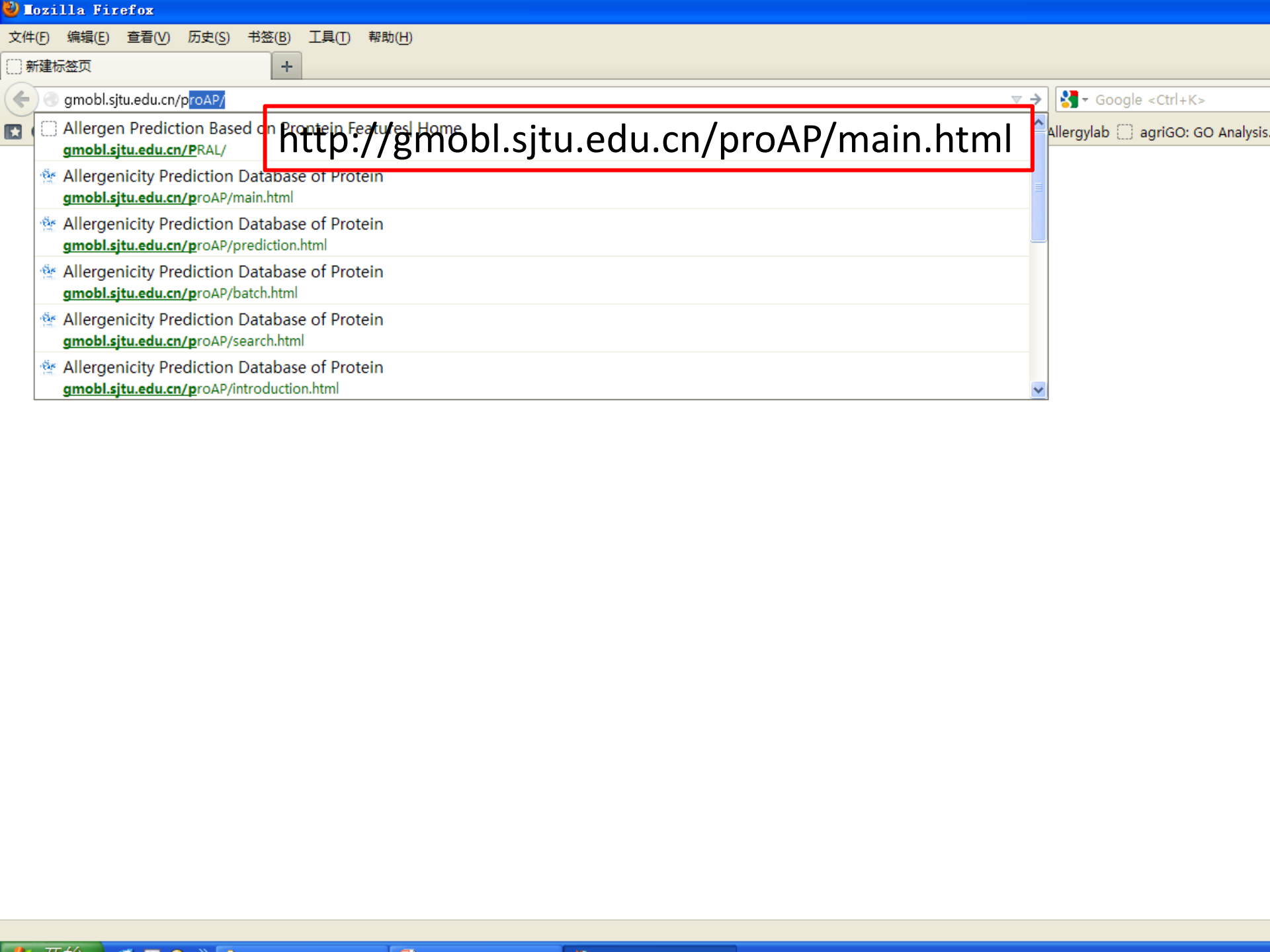
Software Demonstration

- ***proAP*** -- ***Pro***tein Allergenicity ***P***rediction



Thank you


我的文档
Zend Studio - 7.1.2
plsvar.m
简介内容--学生.doc
domainap...
我联系国外博士后职...
Resume(王婧).doc
VBA密码破解.txt
PF00190_...
我的电脑
新建 文本文档.txt
甜菜蚜_半翅目_蚜科_...
clustal-...
Microsoft Word遇到...
在美国做博士后.txt
IMG_0387...
148.jpg
PF00190_...
网上邻居
ipapp.doc
进化树.jpg
foodplan...
writing worksh...
Manual_v...
Resume(E...
2012xing...
1111111.nwk
回收站
安装和用户指南.pdf
IMG_0901...
OS.docx
shanghai jiaotong...
Tra_Gene...
IMG0029A...
2012shen...
Figure2.pdf
ICBC在线银
安装和使用
email.docx
百奥知LIMS
GraphApp...
QMEAN_pl...
IMG_3639...
2012shen...
上海交通大学公开...
Model_1_...
Group时间表.xls
博士等高层次人才...
2012第五届"海内外...
SGM-DY-J...
平均.xlsx
Swiss-pd...使用说明...
080613_S...
Biostat2...
Tra_Expe...
111.txt
Book1.xlsx
上海交通大学+王婧+...
gettemp.gif
暑假班车.txt
学生党支部信息.xls
gettemp2...
Doc1.docx
Administrator
Internet Explorer
迅雷看看播放器
强力卸载电脑上的软件
Mozilla Firefox
位置: C:\Program Files\Mozilla Firefox
Micr... 2007
飞信2012
腾讯QQ
Microsoft Office Excel 2007
Microsoft Office PowerPoint 2007
所有程序(P)
我的文档
我最近的文档(D)
我的电脑
网上邻居
控制面板(C)
网络连接
搜索(S)
运行(R)...
注销(L)
关闭计算机(U)
开始
ICoVax
Demo.pptx


Mozilla Firefox
文件(F) 编辑(E) 查看(V) 历史(S) 书签(B) 工具(T) 帮助(H)
新建标签页
gmobl.sjtu.edu.cn/proAP/
Google <Ctrl+K>
Allergylab
agriGO: GO Analysis
Allergen Prediction Based on Protein Features| Home
gmobl.sjtu.edu.cn/PRAL/
Allergenicity Prediction Database of Protein
gmobl.sjtu.edu.cn/proAP/main.html
Allergenicity Prediction Database of Protein
gmobl.sjtu.edu.cn/proAP/prediction.html
Allergenicity Prediction Database of Protein
gmobl.sjtu.edu.cn/proAP/batch.html
Allergenicity Prediction Database of Protein
gmobl.sjtu.edu.cn/proAP/search.html
Allergenicity Prediction Database of Protein
gmobl.sjtu.edu.cn/proAP/introduction.html
http://gmobl.sjtu.edu.cn/proAP/main.html

Protein Allergenicity Prediction

Navigation

- Introduction
- Allergen search
- Allergenicity prediction
- Batch prediction

Statistics

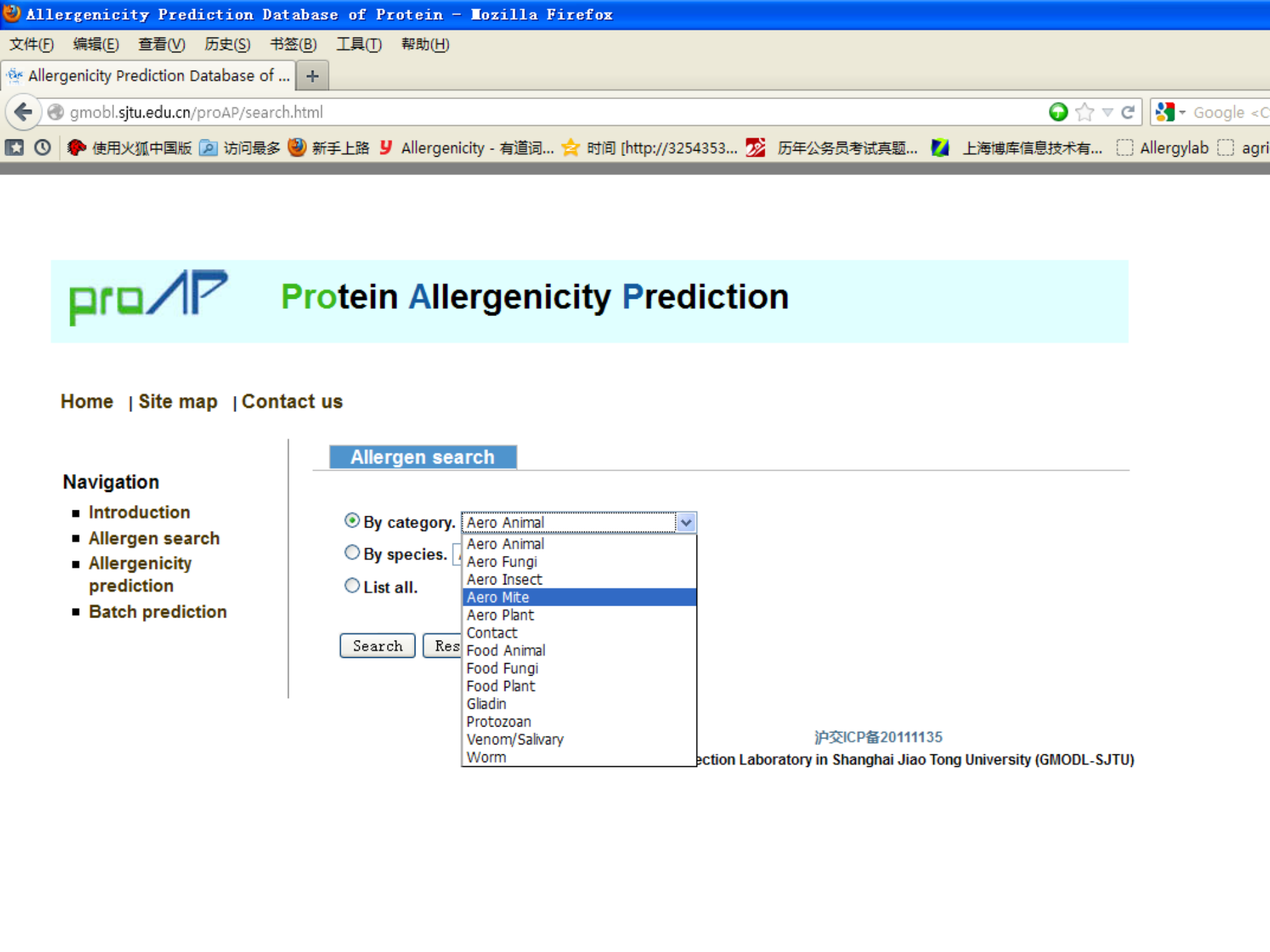
- Allergen: 1096
- Allergen category: 13
- Species: 249
- Prediction method: 4

Links

- Swiss-Prot/TrEMBL
- NCBI-Entrez
- SDAP-Structural Database of Allergenic Proteins
- BLAST
- FASTA
- MEME/MAST

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.


Allergenicity Prediction Database of Protein - Mozilla Firefox
文件(F) 编辑(E) 查看(V) 历史(S) 书签(B) 工具(T) 帮助(H)
Allergenicity Prediction Database of ...
gmobl.sjtu.edu.cn/proAP/search.html
Google
使用火狐中国版 访问最多 新手上路 Allergenicity - 有道词... 时间 [http://3254353... 历年公务员考试真题... 上海博库信息技术有... Allergylab agri
proAP
Protein Allergenicity Prediction
Home | Site map | Contact us
Navigation
Introduction
Allergen search
Allergenicity prediction
Batch prediction
Allergen search
By category.
Aero Animal
By species.
List all.
Search
Res
Aero Animal
Aero Fungi
Aero Insect
Aero Mite
Aero Plant
Contact
Food Animal
Food Fungi
Food Plant
Gliadin
Protozoan
Venom/Salivary
Worm
沪交ICP备20111135
ection Laboratory in Shanghai Jiao Tong University (GMODL-SJTU)

Protein Allergenicity Prediction

Navigation

- Introduction
- Allergen search
- Allergenicity prediction
- Batch prediction

By category.
Aero Animal
By species.
Actinidia
List all.
Search Reset

Search result

Total records: 17 Page: 1/2 First Previous Next Last

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			Prote calciu A7) allerge all (Calciu in 2)(C
2	Can f 1	O18873	Canis lupus familiaris	Canis domesticus,Canis familiaris,Canis canis,Canis lupus familiaris,dogs,dog,	Aero Animal	other		sugar	Major (Allerg (Fla
3	Can f 2	O18874	Canis lupus familiaris	Canis domesticus,Canis familiaris,Canis canis,Canis lupus familiaris,dogs,dog,	Aero Animal	other		sugar	Minor (Allerg (Fla
4	Can f 3	P49822	Canis lupus familiaris	Canis domesticus,Canis familiaris,Canis canis,Canis lupus familiaris,dogs,dog,	Aero Animal	other			Serum (Fla
				Cavia aperea					

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)

- [x] 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.

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)

```
MGIKHCCYILYFTLALVTLVQAGRLGEEVDILPSPNDTRRSLQGCEAHNIIDK
CWRCKPDWAENRQALGD
CAQGFGKATHGGKWGDIYMVTSDQDDDVVNPKEGTLRFGATQDRPLWII
FQRDMIIYLQQEMVVTSDKTI
DGRGAKVFLVYGGITLMNVKNVIIHNIDIHDVRVLPGGRIKSNGGPAIPRHQ
```

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

- [x] FAO/WHO: Amino acids sliding window: 80, sequence identity cutoff >= 45 %
- [ ] 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.

Predict Reset

Protein Allergenicity Prediction

Navigation

- Introduction
- Allergen search
- Allergenicity prediction
- Batch prediction

Prediction results

Query protein:

MQLLLLTVGLALICGLQAQEGNHEEPQGGLEELSGRWHSVALASNKSDLIKPWGHFRVFI
HSMSAKDGNLHGDILIPQDGQCEKVSLTAFKTATSNKFDLEYWGHNDLYLAEVDPKSYLI
LYMINQYNDDTSLVAHLMVRDLSRQQDFLPAFESVCEDIGLHKDQIVVLSDDDRCQGSRD

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			

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

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 canis,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)
>sp|O18874|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%), Simlarity = 100%, Positives = 80/80 (100%)

Query: 1  MQLLLLTVGLALICGLQAQEGNHEEPQGGLEELSGRWHSVALASNKSDLIKPWGHFRVFI 60
          MQLLLLTVGLALICGLQAQEGNHEEPQGGLEELSGRWHSVALASNKSDLIKPWGHFRVFI
Sbjct: 1  MQLLLLTVGLALICGLQAQEGNHEEPQGGLEELSGRWHSVALASNKSDLIKPWGHFRVFI 60

Query: 61 HSMSAKDGNLHGDILIPQDG 80
          HSMSAKDGNLHGDILIPQDG
Sbjct: 61 HSMSAKDGNLHGDILIPQDG 80

query:(2--81)
>sp|O18874|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%), Simlarity = 100%, Positives = 80/80 (100%)

Query: 1  QLLLLTVGLALICGLQAQEGNHEEPQGGLEELSGRWHSVALASNKSDLIKPWGHFRVFIH 60
```

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)

```
LDPAIIGYVAEQENMSASDVVNALNKKSGMLALTGASDMRDVFAKPQENA
VAIKMYVNRV
ADYIAKYLNQLEGNIDGLVFTGGIGENASDCVELFINAVKSLGFATDLKLFVK
YGDSCVV
STPQSKYKIYRVRTNEELMIVEDSIRLTQK
```

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

Navigation

- Introduction
- Allergen search
- Allergenicity prediction
- Batch prediction

Prediction results

Query protein:

MNDNKILVVNAGSSSIKFQLFDYHKKVLAKALCERIFVDGFFKLEFNEQKVEEKVAFPDH
HAAVTHFLNTLKKHKIIQELSDIILVGHRVVQGANYFKDSVIVDAEALAKIKEFIKLAPL
HNKPEADVIEIFFKEVPSAKNVAVFDTTFHTTIPQENYLYAVPRSWEQKHLVRRYGFHGT
SYKFINNYLEKHLNKQNLNLIVCHLGNGASVCAIKNGKSFNTSMGFTPLEGLIMGTRSGD
LDPAIIGYVAEQENMSASDVVNALNKKSGMLALTGASDMRDVFAKPQENAVAIKMYVNRV
ADYIAKYLNQLEGNIDGLVFTGGIGENASDCVELFINAVKSLGFATDLKLFVKYGDSCVV
STPQSKYKIYRVRTNEELMIVEDSIRLTQK

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

--	allergen	allergen	non-allergen

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

Matched records:

No.	Name	UniProtAcc	TaxonomicName	CommonName	Category	Species	Epitope	Sugar
1	Pha a 5	P56166	Phalaris aquatica	Phalaris aquatica L.,Phalaris tuberosa,Phalaris aquatica,canary grass,	Aero Plant	Grass		
2	Par c ?	A2V734	Paralithodes camtschaticus	Paralithodes camtschaticus,Kamchatka crab,red king crab,	Food Animal	other		

Hits Subjects Detail:

>sp|P56166|MPA53_PHAAQ Major pollen allergen Pha a 5.3 OS=Phalaris aquatica PE=1 SV=1
MAVQKYTVALFLAMALVAGPAASYAADAGTPPTPATPAVPGAAAGKATTHEQKLIEDINA
AFKWWPASAPPADKYKTFETAFSKANIAGASTKGLDAAYSVVYNTAAGATPEAKYDSFVT
ALTEALRIMAGTLEVHAVKPATEVPSAKAKILRANSRSSTRSSRFKIAATVATPLSHSTA
ANSAPANDKFTVFEGAFNKAIKERHGGPTETYKFIPSLEAAVKQAYGATVARAPEVKYAV
FEAGLTKAITAMSEAQKVAKPVRLSPQPPQVLPLAAGGAATVAAASDSRGGYKV

>tr|A2V734|A2V734_PARCM Tropomyosin slow-tonic isoform OS=Paralithodes camtschaticus GN=Tm-Pane-tonic PE=2 SV=
MDAIKKKMQAMKLEKDNAMDKADTLEQQNKEANNRAEKTEEEIRLTQKKMQQVENELDVA
QEQLSLANTKLEEKEKALQNAEGEVAALNRRIQLLEEDLERSEERLNTATTKLAEASQAA
DESERMRKVLENRSLSDEERMDALENQLKEARFLAEEADRKYDEVARKLAMVEADLERAE
ERAETGESKIVELEEELRVVGNNLKSLEVSEEKANQREEAYKEQIKTLANKLKAAEARAE
FAERSVQKLQKEVDRLEDELVNEKEKYKNIADEMDQAFSELSGF

Hits Subjects Detail:

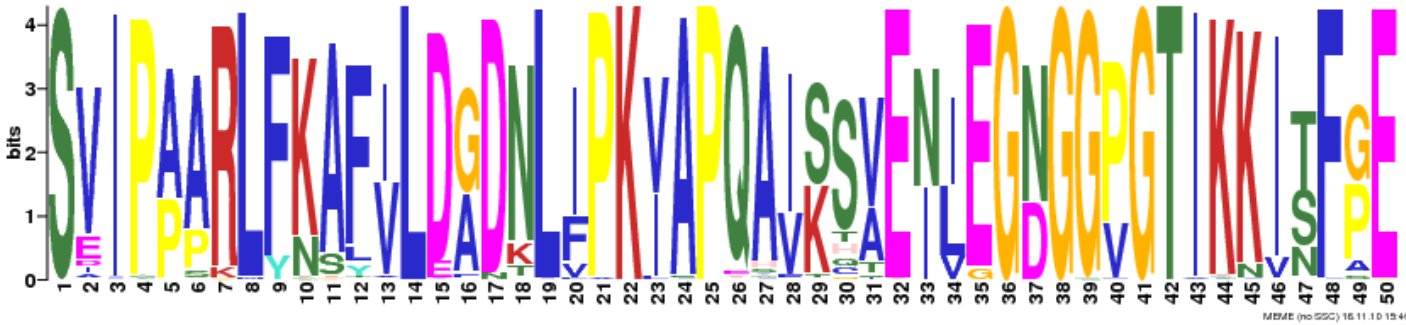
>sp|P56166|MPA53_PHAAQ Major pollen allergen Pha a 5.3 OS=Phalaris aquatica PE=1 SV=1
MAVQKYTVALFLAMALVAGPAASYAADAGTPPTPATPAVPGAAAGKATTHEQKLIEDINA
AFKWWPASAPPADKYKTFETAFSKANIAGASTKGLDAAYSVVYNTAAGATPEAKYDSFVT
ALTEALRIMAGTLEVHAVKPATEVPSAKAKILRANSRSSTRSSRFKIAATVATPLSHSTA
ANSAPANDKFTVFEGAFNKAIKERHGGPTETYKFIPSLEAAVKQAYGATVARAPEVKYAV
FEAGLTKAITAMSEAQKVAKPVRLSPQPPQVLPLAAGGAATVAAASDSRGGYKV

>tr|A2V734|A2V734_PARCM Tropomyosin slow-tonic isoform OS=Paralithodes camtschaticus GN=Tm-Pane-tonic PE=
MDAIKKKMQAMKLEKDNAMDKADTLEQQNKEANNRAEKTEEEIRLTQKKMQQVENELDVA
QEQLSLANTKLEEKEKALQNAEGEVAALNRRIQLLEEDLERSEERLNTATTKLAEASQAA
DESERMRKVLENRSLSDEERMDALENQLKEARFLAEEADRKYDEVARKLAMVEADLERAE
ERAETGESKIVELEEELRVVGNNLKSLEVSEEKANQREEAYKEQIKTLANKLKAAEARAE
FAERSVQKLQKEVDRLEDELVNEKEKYKNIADEMDQAFSELSGF

Motif-based result (allergen)

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

SEQUENCE NAME	DESCRIPTION	E-VALUE	LENGTH
name		0.00048	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:

浏览...

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