

Agency for Science, Technology and Research SINGAPORE

AllerCatPro -Protein Allergenicity Prediction with 3D Structure Features

- Bioinformatics Institute (BII)
- Co-developed with Krutz *et al.* (P&G) as academic free tool
- Gluten work with TCCC

Dr. Sebastian Maurer-Stroh Executive Director, Bll



The aim – biophysics inspired in silico model of 3D allergen recognition



POWERING DISCOVERIES



DISCOVERIES

POWERING

AllerCatPro v1

Maurer-Stroh S, Krutz NL, Kern PS, Gunalan V, Nguyen MN, Limviphuvadh V, Eisenhaber F, Gerberick GF. **AllerCatProprediction of protein allergenicity potential from the protein sequence.** *Bioinformatics.* 2019 Sep 1;35(17):3020-3027. Cited 40 times in <2 years

Krutz NL et al., Kimber I, Maurer-Stroh S, Gerberick GF. **Determination of the relative allergenic potency of proteins: hurdles and opportunities.** *Crit Rev Toxicol.* 2020 Jul;50(6):521-530.



Combined database of known allergens



Search our 3D

sequence database

Query:

3D Hit:

Allerg:

Conformational epitope (5



Query Protein

(unknown structure)

GVFNYETETTSVIPAARLFKAYILDGD

IgE antibody

Allergen

Highest accuracy on difficult benchmark, compared to Codex rules both have ~100% sensitivity to find known allergens but AllerCatPro has 37-fold higher specificity (less false positives)

3D epitope similarity

GVFNYETETTSVIPAARLFKAYILDGD

GVFNFETETTSVIPAARLFSALILDGD

GL**F**Q**Y**ESDTSSVLPAVKLF**K**A**Y**II**E**SD

Protein hit with known structure

Application 1: New plant product allergenicity screen

1. Use label-free proteomics to identify most abundant proteins in food or plant material for new products



Krutz NL, Winget J, Ryan CA, Wimalasena R, Maurer-Stroh S, Dearman RJ, Kimber I, Gerberick GF. Proteomic and Bioinformatic Analyses for the Identification of Proteins With Low Allergenic Potential for Hazard Assessment. *Toxicol Sci.* 2019 Jul 1;170(1):210-222.

2. Options: Evaluate top 50 protein hits (or above critical concentration) with in silico tool AllerCatPro







(100)

×



Application 2: Safety assessment for novel food protein

RESEARCH ARTICLE

Soy Leghemoglobin

Molecular Nutrition Food Research

Evaluating Potential Risks of Food Allergy and Toxicity of Soy Leghemoglobin Expressed in *Pichia pastoris*

Yuan Jin, Xiaoyun He, Kwame Andoh-Kumi, Rachel Z. Fraser, Mei Lu, and Richard E. Goodman*

Scope: The Soybean (*Clycine max*) leghemoglobin c2 (LegHb) gene was introduced into *Pichia pastoris* yeast for sustainable production of a heme-carrying protein, for organoleptic use in plant-based meat. The potential allergenicity and toxicity of LegHb and 17 *Pichia* host-proteins each representing $\geq 1\%$ of total protein in production batches are evaluated by literature review, bioinformatics sequence comparisons to known allergens or toxins, and in vitro pepsin digestion.

1. Introduction

Hemoglobins (Hbs) are ubiquitous iron binding proteins in nature, present in bacteria, fungi, higher plants, and animals.^[1] Consumption of these proteins serves as an efficient source of bioavailable iron, which is required for ovvcen transport respiration and

State-of-the-art:

11 of 18 tested proteins have linear window matches to allergens. All further evaluated by literature search...

AllerCatPro:

Only **3** of 18 tested proteins have 3D window matches to allergens. Shortlist for further focussed searches...

Novel food protein: e.g. burger patty based on soy leghemoglobin produced in *Pichia*



			All	lerCatPro Results					
1.00	Second Longit U	Lincon albertaria	a d'adama	Sec. 4					
Halan		(prof-th requests)	and the second s	Posto area	N. Mentip. Areast 60 ac	To Mandilay, 30 Apatopar To	Box II spilipe U	-	
Lape C	105	1	1	ILETE CHETH REPARE FUR DAMA CTT & PROVIDE	27.4	818	day is product		New P
C Alges annualgate missione (1815 AA) in 23010000 Aut. LANSTER F	144			No applicant Int (2 - raise three dealer 3 30 Fy					wy falls
Ze Cetetere obgenere official preface (FB 44.15 200 8043 km OVECNET	758			5411 3 FW1 450 C0250	:77.8	- 10.0	data di sectore	100	30asi +971
28 Avenues (2013) (2 (2020) (2 (2 (2 (2 (2 (2 (2 (2 (2 (2 (2 (2 (2	775	1		No applicant to (2 ratio Translati 2 201)				and so the second	- mi felte
3 Torokebbase (5/9 AA) (8 2082/0557 Arx: C40/87968 1	679			No. septilizant Mr. X. value (Proof-001.001)				The second	: 01/145
4 dispand kinasi didi AA Gr2000402 App cARTIDHE F	425			No agelicate by (2 water Presided 1 187)				ALC: NAME	Ho Ma
for Canadase A. (FTE AVE ID (NAROWSE AND APP.) (REAROWSE)	418			Terre III STATUSED GENER			-	unal address	100
Sit. Skanne fryslegeliete belyttergeneers (18975) (654 AA, 51 J3653366) Aut. CAVEGES 1				No. Appendix of the (R. Salar President) 1 (R.1)					-
In Hyperbolical process PHS (628 AV) 10 2002 (V10 Aux COVERTID 1	- 648	- 11		analysis detyclogenesis, always Cirk N 12 (Salingenbar Cirk 107.2.2	12.8	873	distantion of		15000 1-000
8. Weeks whether all and the semigroup was a part of the California Ave. CALIFULT	101			anony-de obly-thogenasa, allegan Cit & 10 (Castopo Ant	76.0	80	des3.0wbc		Magi 1415h
Ta Della annonomiatate deluttatase (141 M/LCB 20203045 Aug. CM17467 4	340			No agrillant Hild Instan Eyenhold (2007)				No. of Concession, Name	641565
The Million of the and an and an and an and a second a second a second and a second				alashi kitata press (Contra aliante) (1981) 91.		94.4	data in plantas	-	20q1-07
Ps Materia delphrogenese (342 AV) (8: 2001/3884 Auc S2012088 1	147			Aug 7 4504 (\$52.2 (\$27.8.1))	101	864	data de administra	and the second value of th	33q1 -97
To Publice protect scheme function (201 Arty St 218822788 Acts DAV/1878 1	100			No applicant life (E-value Proved-Sci 2007)				and the second second	acide.
8 Trise photohais terrema (348 AU-022002090 Aur CATTIED 1	34	1		No 11 PH1 402 GP30	- 121	10.6	dist. 1. for her	out offers	1000
na mganhankan panan panghan (PP1438 (PP1438 CE 19590000 Apr. CEARAGE)	101			Integration (Sugar gives torregated CALTERIES)	- 18.8	957	disk disks		Equi (11)
The Calorado argumentic dominant (TM 146) (32230034030 Aur. CATT(202))	134			194 + 5 4202 4063618	- 10.0	89.2	shin (A. Kontho		Ever.
12 Montunation #These resides (34 My in 2008/JUTH Ave. Conversion)	- 64	1		No Applicated by (\$ -raise Trivid-and 2 MT)				The summer of	84 / MR





- 1. update of the underlying data set
- 2. structure-guided approach to detect Celiac disease peptides
- 3. increased cross-links in the output
- 4. inclusion of experimental epitopes from IEDB both in a linear and 3D context for selected families
- 5. application to cross-reactivity between insect and shellfish allergens

1. update of the underlying data set, 4180=>4313



New: 144 Removed: 11

Database	Number of sequences
Allergome	45
IUIS	40
FARRP	17
UniprotKB + literature	24
Compare	18
Total	144

(100) 100

×

2. structure-guided approach to detect Celiac disease peptides

Old version – 9AA fingerprint



Gluten 3D-Al extension (new)



85%

80%

SEQ1 SEQ2 STR

SAS

SVM

Methods

RF

4aa

SEQ2 STR

SAS

SVM

RF

4aa

9aa

SEQ1

100%

95%

90%

85%

80% 75%

Sensitivity

(100) 100)

×

9aa

3. increased cross-links in the output (Format will still be simplified!)

AllerCatPro Results



/ List of all hits (not just best)

Hit	Protein Name	Species	AllerCatPro ID	UniProt	SUPFAM	Pfam	InterPro	% identity
1	Bet v 2.0101	Betula pendula (European white birch) (Betula verrucosa).	1887	<u>P25816</u>	<u>SSF55770</u>	PF00235	IPR005455 IPR036140 IPR027310	100.0
2	f pdb 1CQA A Chain A, SEQRES	•	-	-	•	-	•	96.2
3	Bet v 2	Betula pendula (European white birch) (Betula verrucosa).	212	<u>A4K9Z8</u>	<u>SSF55770</u>	PF00235	IPR005455 IPR036140 IPR027310	96.2
4	Cor a 2	Corylus avellana (European hazel) (Corylus maxima).	299	<u>A4KA45</u>	<u>SSF55770</u>	<u>PF00235</u>	IPR005455 IPR036140 IPR027310	93.2
5	Cor a 2	Corylus avellana (European hazel) (Corylus maxima).	300	<u>A4KA44</u>	<u>SSF55770</u>	PF00235	IPR005455 IPR036140 IPR027310	91.0
6	Cor a 2	Corylus avellana (European hazel) (Corylus maxima).	302	<u>A4KA39</u>	<u>SSF55770</u>	PF00235	IPR005455 IPR036140 IPR027310	91.0
7	Cor a 2	Corylus avellana (European hazel) (Corylus maxima).	301	<u>A4KA40</u>	<u>SSF55770</u>	PF00235	IPR005455 IPR036140 IPR027310	<mark>91.0</mark>
8	Ric c 8	Ricinus communis (Castor bean).	1410	B9RKF4	<u>SSF55770</u>	PF00235	IPR005455 IPR036140 IPR027310	89.5
9	Gly m 3	Glycine max (Soybean) (Glycine hispida).	477	<u>11K602</u>	<u>SSF55770</u>	PF00235	IPR005455 IPR036140 IPR027310	86.5
10	Mer a 1.0101	Mercurialis annua (Annual mercury).	1908	<u>049894</u>	<u>SSF55770</u>	PF00235	IPR005455 IPR036140 IPR027310	85.0
11	Ole e 2	Olea europaea (Common olive).	835	A4GD52	<u>SSF55770</u>	PF00235	IPR005455 IPR036140 IPR027310	85.8

Cross-links to multiple databases, for protein family (domains), more info in Allergome etc.

(100)

4. inclusion of experimental epitopes from IEDB

AllerCatPro Results

			Gluten	Gluten	Iten								Kn	own allergenic	proteins										
Protein	Sequence Length	allergens (# of Q- repeats)	# of 3x6- mer overlaps	# hits @	Best hit protein name	Species	Allergen info ø	UniProt @	SUPFAM	Pfam @	InterPro @	lgE prevalence Ø	Cross reactivity	Sequence homology	% identity, linear 80 aa window	% identity, 3D epitope	Show 3D epitope	% identity, 3D IEDB	Show 3D IEDB	# IEDB @	# IEDB- 2M ©	Result ©	Comment		
pos- Profilin- Bet.p.	133	0	-	<u>234</u>	Bet v 2.0101	Betula pendula (European white birch) (Betula verrucosa).	<u>AG(5)</u>	<u>P25816</u>	<u>SSF55770</u>	PF00235	IPR005455 IPR036140 IPR027310	<u>247</u>	-	-	100.0	100.0	show in structure	100.0	<u>show in</u> structure	16	23	strong evidence	3Depi >93%		
neg- Profilin- Sac.c.	126	0	-	<u>233</u>	Ama r 2	Amaranthus retroflexus (Redroot amaranth) (Redroot pigweed).	<u>AG(2)</u>	<u>C3W2Q7</u>	<u>SSF55770</u>	PF00235	-	<u>12</u> mendel3.bii.a-star.	edu.sg/METHOD	Che a 2, Cro s 2, Cuc m 2, Hev b 8, Sal k 4 S/allergy/tmp.7	38.8 67. TRI A 12.010	4.4 02 C. P4 Q	show in structure mendel3.	60.0 bii.a-star.edu.sg	show in structure	0 rgy/mp/t	0 mp_767_TR	weak evidence	3Depi <=93%		

AllerCatPro

Spin ON | Spin OFF | Save IMAGE

pon-rotani-adu p. en model: RPT TRI J. 12, 0102, C. P.49233, AFTER, LODE e: 0280, 5315, P44P, GISG, E46E, BOI, In51H, KSRY, ISSD, P54F, E55E, BOH, ILOU, TSO, GIGG, GBOE, YTY, VAY, GYTS, ETRE, GBOD, ISU, KBK with mouse over residue to see its position number. Right-click for more Query: pos-Profilm-Bet p. Reference model: 787, TRL _12, 0102, C, P4923, AFTER_LOOP 30 (E09 epilope: 080); E92, 111, C132, D140, TN, D140, G178, O180, AT Surds, Var M, CL, S228, A228, AE072, E1068, T111, C1140, VARIAB, M171 Howr with mouse over residue to see its position number. Right-click for more options.



For protein families with IEDB info, epitope similarity match over best known B cell epitope with 3D and linear match reported

(100) 1001

×

5. cross-reactivity between insect and shellfish allergens

> Food Chem. 2021 Jun 30;348:129110. doi: 10.1016/j.foodchem.2021.129110. Epub 2021 Jan 19.

Protein extraction protocols for optimal proteome measurement and arginine kinase quantitation from cricket Acheta domesticus for food safety assessment

Utpal Bose ¹, James A Broadbent ¹, Angéla Juhász ², Shaymaviswanathan Karnaneedi ³, Elecia B Johnston ³, Sally Stockwell ¹, Keren Byrne ¹, Vachiranee Limviphuvadh ⁴, Sebastian Maurer-Stroh ⁵, Andreas L Lopata ³, Michelle L Colgrave ⁶

Arginine kinase (AK) phylogenetic tree



Improvement by adding family-specific shellfish allergen epitope score and clinical data guided threshold

		%identity, linear 80 aa	#IEDB	#IEDB-2M	
		window			Threshold
Known allergenic	arginine kinase(Lit v 2) Whiteleg shrimp B0FRF9	100	35	39	
AKs	arginine kinase(Pen m 2) Giant tiger prawn E7CGC2	100	28	37	X
	arginine kinase(Bla g 9.0101) German cockroach Q2HZF2	100	3	23	
Putativa	arginine kinase fragment cricket A0A120MGB4	97.5	3	11	
Cricket	arginine kinase cricket Xbra0010686	96.2	6	33	
AKS	arginine kinase oriental migratory locust A6M9J4	95.0	6	27	
	creatine kinase Human AAC31758.1	86.2	0	0	New
	creatine kinase M-type Cow NP 777198.2	90.0	0	0	
Non-allergenic AK homologs	creatine kinase M-type Chicken NP 990838.1	91.2	0	0	





THANK YOU

www.a-star.edu.sg