Strategie *in silico* di predizione di tossicità applicate alle proteine

Ivano Eberini e Luca Palazzolo

Dipartimento di Scienze Farmacologiche e Biomolecolari “Rodolfo Paoletti”

Università degli Studi di Milano, Milano, Italia

Protein toxicity – Intro

Some proteins can cause adverse effects in humans and animals, via a variety of mechanisms and in a variety of settings (*Dang and Van Damme, 2015; Franceschi et al., 2017; Lucas et al., 2018*).

- In the scientific literature the term 'toxic proteins' generally refers to *proteins of exogenous origin capable of causing adverse effects to human beings or animals in the context of an offence/defence paradigm.*

On the basis of Gene Ontology (GO) definition of toxin activity, toxic proteins (or toxins) can be defined as proteins that *interact selectively with one or more biological molecules in another organism (the "target" organism), initiating pathogenesis (leading to an abnormal, generally detrimental state) in the target organism.*

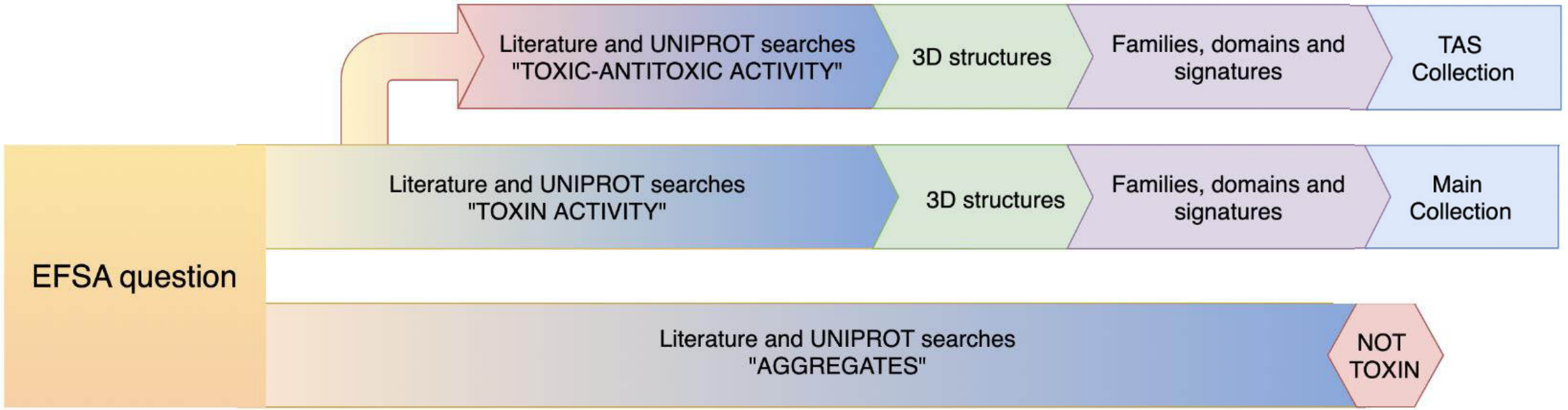
Various plants, animals and bacteria produce toxic proteins to prevail in hostile environments.

The toxic activity of such proteins is achieved via a variety of mechanisms.

For our purpose, it was considered useful to cluster toxins into two groups:

- proteins causing toxic effects *per se*, acting as monomers or homo-multimers; these toxic proteins can be found in animal venoms, in plants and in bacteria.
- toxins acting in the context of toxin-antitoxin systems, i.e., proteins causing a toxic effect only in case of perturbation of the toxin to antitoxin concentration equilibrium; these toxins are found only in bacteria.

Search Pipeline



Searches in UniProtKB database

Question	Which proteins are associated with a well-recognized toxic activity?
Keyword	To select relevant studies concerning proteins with associated toxic activity per se from the comprehensive set of literature previously compiled the following keywords were used in the UniProtKB search: "Toxin activity and Reviewed:Yes" and "Toxin activity and Reviewed:No"
Rationale	The search strategy was based on the term "Toxin activity" described in Gene Ontology (GO) Annotation database, since this term was as the most relevant term to address the terms of reference of this work. This term was used as the UniProtKB search term. The full definition of this term as provided by the GO is reported in the table beside.

Term:	Toxin activity
Synonyms:	toxin receptor binding
Definition:	Interacting selectively with one or more biological molecules in another organism (the 'target' organism), initiating pathogenesis (leading to an abnormal, generally detrimental state) in the target organism. The activity should refer to an evolved function of the active gene product, i.e., one that was selected for. Examples include the activity of botulinum toxin, and snake venom.
Parent terms:	is-a molecular function
Category:	Molecular Function
Id:	GO:0090729

Searches in UniProtKB database - Overview

UniProt string	Reviewed:YES	Reviewed:NO
“Toxin activity”	6,964	47,831

Data on 6,964 proteins with an associated well-recognised toxic activity (“Toxin activity and Reviewed:Yes”) were downloaded from UniProtKB (March 2020). These proteins compose the Main Collection.

Data on 47,831 proteins with an associated well-recognised toxic activity (“Toxin activity and Reviewed:NO”) were downloaded from UniProtKB (March 2020).

Searches in UniProt database – 3D Structures

Method	Number of experimentally-solved structures
X-ray	1441
NMR	586
Model	55
Electron microscopy	31

Out of 6,964 proteins identified in this search, 765 have associated one or more experimentally-derived 3D structure/s.

A total of 5,298 models were downloaded from the Swiss-Model repository and stored into our Collection. There are some toxins with two or more associated models in the SM repository.

From “toxin activity” to “toxin-antitoxin system”

Question	How to extend the toxic protein Main Collection with UniProtKB annotated proteins considering the outcome of the primary search (identification of a new search term)?
Keywords	In order to select relevant studies concerning proteins with some associated toxic effect but not covered by the GO term toxin activity, the following keywords were used “Toxin-antitoxin system and Reviewed:Yes” and “Toxin-antitoxin system and Reviewed:No”.
Rationale	Some proteins identified through the primary search were noted to belong to a toxin-antitoxin system (TAS); however, it was observed that not all of these TAS proteins were associated with the GO term “Toxin activity” in the UniProtKB database but they were linked to toxin activity in the peer-reviewed literature. In fact, the term “toxin activity” in UniProtKB refers to proteins that have a well-recognized toxic activity per se, and that interact primarily with proteins of the target organism. Toxins that belong to TAS primarily interact with their antitoxins and can interact with proteins of a target organism only if the toxin-antitoxin equilibrium is disrupted. Based on the above observations, a pearl growing strategy was applied to identify these additional TAS toxins.

Searches in UniProtKB database – Entries

Search string (UniProt)	Reviewed: Yes	Reviewed: No
Toxin-Antitoxin System	627	155,039

Number of selected toxins for TAS Collection

Papers in the three source literature databases for the TAS Collection

String search	PubMed	WOS	SCOPUS
“toxin-antitoxin system”	410	420	776
toxin-antitoxin (AND) system	1,075	1,260	1,060

Searches in UniProtKB database – 3D Structures

Method	Number of experimentally-solved structures
X-ray	256
NMR	19
Electron microscopy	3

Out of the 627 identified TAS proteins, 114 have associated one or more experimentally-derived 3D structures.

A total of 356 models were downloaded from the Swiss-Model repository and stored into the TAS Collection. Also in this case, there are some toxins with 2 or more associated models in the SM repository.

Families, domain and signatures

Main Collection

Source	Number of families/domains/signatures
PFAM	288
INTERPRO	599
PROSITE	138
CATH-GENE3D	8
SUPFAM	94
PRINTS	66
SMART	62
PANTHER	33
TIGRFAMs	27
PIRSF	25
CDD	35

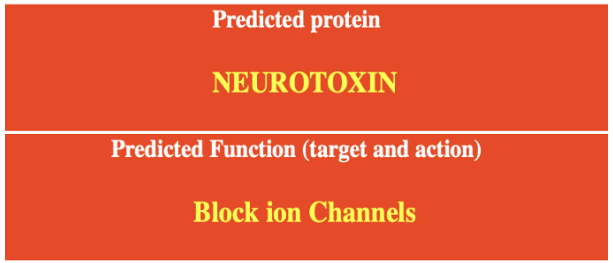
TAS Collection

Source	Number of families/domains/signatures
PFAM	92
INTERPRO	159
PROSITE	11
CATH-GENE3D	1
SUPFAM	21
PRINTS	2
SMART	9
PANTHER	17
TIGRFAMs	21
PIRSF	12
CDD	7

Test of predictive tools

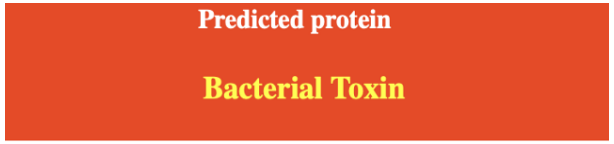
NTXPred

Dataset	Method	Sensitivity	Specificity	Accuracy
TP+TN1	Amino acid	0.95	1	1
TP+TN2		0.95	1	1
TP+TN1	Dipeptide	0.95	1	1
TP+TN2		0.95	1	1
TP+TN1	PSI-BLAST	0.95	1	1
TP+TN2		0.95	1	1

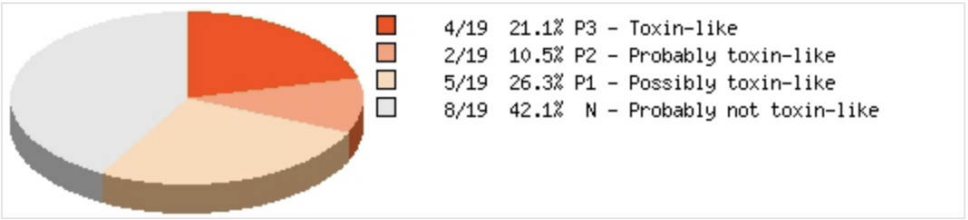


BTXPred

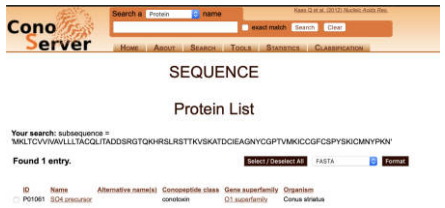
Dataset	Method	Sensitivity	Specificity	Accuracy
TP+TN1	Amino acid (SVM)	0.6	0.05	0.3
TP+TN2		0.6	1	0.8
TP+TN1	Dipeptide (SVM)	0.3	0.25	0.275
TP+TN2		0.3	1	0.65



Clantox



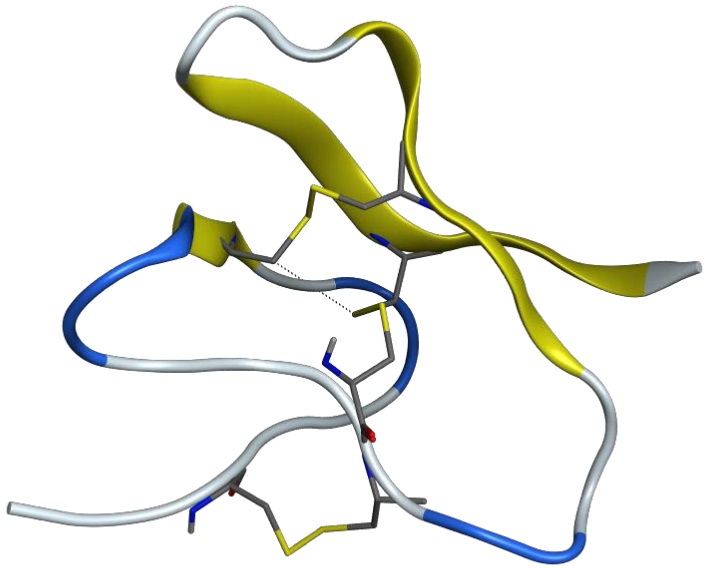
Dataset	Method	Sensitivity	Specificity	Accuracy
TP+TN1	Less conservative	0.2	1	0.6
TP+TN2		0.2	1	0.6
TP+TN1	Conservative	0.5	1	0.76
TP+TN2		0.5	1	0.76
TP+TN1	More conservative	0.6	1	0.8
TP+TN2		0.6	1	0.8



ConoServer

Dataset	Sensitivity	Specificity	Accuracy
TP+TN1	1	1	1
TP+TN2	1	1	1

Test of predictive tools



Knottin

Dataset	Method	Sensitivity	Specificity	Accuracy
TP+TN1	Less conservative	0.5	0.85	0.675
TP+TN2		0.5	1	0.75
TP+TN1	More conservative	0.85	0.85	0.85
TP+TN2		0.85	1	0.925

ToxinPred

Dataset	Sensitivity	Specificity	Accuracy
TP+TN	0.6	0.95	0.78

Overall Conclusions

- An extensive literature and protein database search was carried out.
- All the information was gathered from reference and mainly manually annotated protein databases, considered the golden standard sources by the scientific community.
- Data on protein sequence, structure and activity and key literature entries were retrieved, analysed and collected.
- Two comprehensive Collections of proteins (Main and TAS), associated with toxic effects and related relevant information (knowledgebase), were generated and can be automatically updated by our Toxapex software.
- This updatable knowledgebase is preparatory for the development of a novel risk assessment strategy for poorly characterized proteins.

> RSC Adv. 2020 Jun 4;10(36):21292-21308. doi: 10.1039/d0ra02701d. eCollection 2020 Jun 2.

A joint optimization QSAR model of fathead minnow acute toxicity based on a radial basis function neural network and its consensus modeling

Yukun Wang [1] [2], Xuebo Chen [2]

Affiliations + expand

PMID: 35518745 PMCID: PMC9054390 DOI: 10.1039/d0ra02701d

Free PMC article

J Chem Inf Model. 2021 Feb 22;61(2):653-663. doi: 10.1021/acs.jcim.0c01164. Epub 2021 Feb 3.

Large-Scale Modeling of Multispecies Acute Toxicity End Points Using Consensus of Multitask Deep Learning Methods

Sankalp Jain [1], Vishal B Siramshetty [1], Vinicius M Alves [2], Eugene N Muratov [2], Nicole Kleinstreuer [3 4], Alexander Tropsha [2], Marc C Nicklaus [5], Anton Simeonov [1], Alexey V Zakharov [1]

Affiliations + expand

PMID: 33533614 PMCID: PMC8780008 DOI: 10.1021/acs.jcim.0c01164

Free PMC article

Insieme di tecniche basate su reti neurali artificiali organizzate in diversi strati, dove ogni strato calcola i valori per quello successivo affinché l'informazione venga elaborata in maniera sempre più completa.

> PLoS Comput Biol. 2021 Jul 2;17(7):e1009135. doi: 10.1371/journal.pcbi.1009135.
eCollection 2021 Jul.

Leveraging high-throughput screening data, deep neural networks, and conditional generative adversarial networks to advance predictive toxicology

Adrian J Green [1], Martin J Mohlenkamp [2], Jhuma Das [3], Meenal Chaudhari [4], Lisa Truong [5], Robyn L Tanguay [5], David M Reif [1]

Affiliations + expand

PMID: 34214078 PMCID: PMC8301607 DOI: 10.1371/journal.pcbi.1009135

Free PMC article

Chem Biol Drug Des. 2021 Aug;98(2):248-257. doi: 10.1111/cbdd.13894. Epub 2021 Jun 7.

In silico prediction of drug-induced ototoxicity using machine learning and deep learning methods

Xin Huang [1], Fang Tang [2], Yuqing Hua [1 3], Xiao Li [1 4]

Affiliations + expand

PMID: 34013639 DOI: 10.1111/cbdd.13894

Review > Mol Divers. 2021 Aug;25(3):1409-1424. doi: 10.1007/s11030-021-10239-x.
Epub 2021 Jun 10.

Machine learning models for classification tasks related to drug safety

Anita Rácz [1], Dávid Bajusz [2], Ramón Alain Miranda-Quintana [3], Károly Héberger [4]

Affiliations + expand

PMID: 34110577 PMCID: PMC8342376 DOI: 10.1007/s11030-021-10239-x

Free PMC article

> Sensors (Basel). 2022 Oct 26;22(21):8185. doi: 10.3390/s22218185.

Predicting Chemical Carcinogens Using a Hybrid Neural Network Deep Learning Method

Sarita Limbu [1], Sivanesan Dakshanamurthy [1]

Affiliations + expand

PMID: 36365881 PMCID: PMC9653664 DOI: 10.3390/s22218185

Free PMC article

Towards novel risk assessment procedures

• In parallel to the well-known and validated strategies for risk assessment of chemicals, such as QSAR, the produced knowledgebase may help develop a comprehensive in silico risk assessment strategy for new proteins.

• Procedures could be set and tested with Cooper's statistics, in the same way as to what has already been extensively done in the past for chemicals.

• Homology detection can be used for inferring toxic properties from well characterized entries of our knowledgbase, since homologous proteins share general architecture and functions.

• BLASTing our knowledgebase, instead of the whole UniProtKB database, will increase the probability to identify phylogenetic relationships between the investigated query and a set of known and well-annotated toxins.

• Additional and more sensitive strategies, for detecting distant homology with toxic entries, can be implemented by using multiple alignments/profiles, specific domains and/or molecular signatures and hidden Markov models (HMM).

• Moreover artificial intelligence/machine learning approaches could also be applied combining findings from the methods as the above-mentioned in order to try to increase the accuracy of predictive risk assessment for proteins.