

APPROVED: 28 May 2020

doi:10.2903/sp.efsa.2020.EN-1875

Literature search – Exploring *in silico* protein toxicity prediction methods to support the food and feed risk assessment

L. Palazzolo¹, E. Gianazza¹, I. Eberini¹

¹Dipartimento di Scienze Farmacologiche e Biomolecolari, Università degli Studi di Milano, Via Balzaretti 9, 20133 Milano (IT)

Abstract

This report is the outcome of an EFSA procurement (NP/EFSA/GMO/2018/01) reviewing relevant scientific information on *in silico* prediction methods for protein toxicity, that could support the food and feed risk assessment. Several proteins are associated with adverse (toxic) effects in humans and animals, by a variety of mechanisms. These are produced by plants, animals and bacteria to prevail in hostile environments. In the present report, we present an integrated pipeline to perform a comprehensive literature and database search applied to proteins with toxic effects. "Toxin activity" and "toxin-antitoxin system" strings were used as inputs for this pipeline. UniProtKB was considered as the reference database, and only the UniProtKB curator-reviewed proteins were considered in the pipeline. Experimentally-determined structures and homology-based *in silico* 3D models were retrieved from protein structures repositories; family-, domain-, motif- and other molecular signature-related information was also obtained from specific databases which are part of the InterPro consortium. Protein aggregation associated with adverse effects was also investigated using different search strategies. This work can serve as the basis for further exploring novel risk assessment strategies for new proteins using *in silico* predictive methods.

© European Food Safety Authority, 2020

Key words: (protein toxicity, bioinformatics, toxic activity, toxin-antitoxin system, aggregates, predictive toxicity)

Question number: EFSA-Q-2019-00038

Correspondence: gmo_secretariat_applications@efsa.europa.eu

Disclaimer: The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.

Suggested citation: Palazzolo L, Gianazza E and Eberini I, 2020. Literature search – Exploring *in silico* protein toxicity prediction methods to support the food and feed risk assessment. EFSA supporting publication 2020:EN-1875. 89 pp. doi:10.2903/sp.efsa.2020.EN-1875

ISSN: 2397-8325

© European Food Safety Authority, 2020

Summary

Abstract.....	1
Summary	3
1. Introduction.....	5
1.1. Background and Terms of Reference as provided by the requestor	5
1.1.1. Task 1.....	5
1.1.2. Task 2.....	5
1.2. Interpretation of the Terms of Reference as provided by EFSA	5
1.3. Protein toxicity – Background information.....	5
1.3.1. Toxic proteins.....	6
1.3.2. Proteins with putative toxic activity as aggregates.....	8
2. Data and Methodologies	9
2.1. Databases	10
2.1.1. Literature databases.....	10
2.1.2. Protein Databases - Primary structure and function.....	11
2.1.3. Protein databases - Modelling tools	12
2.1.4. Families, domains and signatures.....	13
2.1.5. Toxin predicting tools	15
2.2. Methodologies	16
2.2.1. Compilation of a comprehensive collection of information for toxic proteins	16
2.2.2. Proteins with putative toxic activity as aggregates.....	23
2.3. Three-dimensional structures.....	26
2.4. Families, domains and signatures.....	26
2.5. Automated information retrieval and management tool	26
3. Results	27
3.1. Toxic proteins	27
3.1.1. Main Collection	27
3.1.2. TAS Collection.....	46
3.2. Proteins with putative toxic activity as aggregates.....	60
4. Quality assurance of 3D structures	63
4.1. Protein Data Bank	63
4.1.1. Models downloaded from SM repository	64
5. Evaluation of toxin prediction tools.....	66
5.1. NTXpred.....	67
5.2. BTXpred	70
5.3. Knottin	73
5.4. CLANTOX	76
5.5. ConoServer.....	79
5.6. ToxinPred.....	82
6. Evaluation of other freely available tools.....	83
6.1. Meme.....	83
7. Discussion, conclusions and future perspectives	84
References.....	86

Annex A - Main Collection: Toxin Activity, Reviewed: YES

Annex B - Toxin Activity, Reviewed: NO

Annex C - TAS Collection: Toxin-antitoxin system, Reviewed: YES

Annex D - Toxin-antitoxin system, Reviewed: NO

Annex E - Main and TAS Collections: Pfam

Annex F - Main and TAS Collections: InterPro

Supplementary information:

Main Collection MC_citation_scopes.xlsx

Main Collection MC_citations.xlsx

Main Collection MC_comments.xlsx
Main Collection MC_databases.xlsx
Main Collection MC_go.xlsx
Main collection MC_keywords.xlsx
Main Collection MC_organisms.xlsx
Main Collection MC_pdb.xlsx
Main Collection MC_Pfam_identities.xlsx
Main Collection MC_sequences.xlsx
Main Collection MC_swiss_model.xlsx

TAS Collection TC_citation_scopes.xlsx
TAS Collection TC_citations.xlsx
TAS Collection TC_comments.xlsx
TAS Collection TC_databases.xlsx
TAS Collection TC_go.xlsx
TAS Collection TC_keywords.xlsx
TAS Collection TC_organisms.xlsx
TAS Collection TC_pdb.xlsx
TAS Collection TC_Pfam_identities.xlsx
TAS Collection TC_sequences.xlsx
TAS Collection TC_swiss_model.xlsx

1. Introduction

1.1. Background and Terms of Reference as provided by the requestor

This report is the outcome of a contract titled “Literature search – exploring *in silico* protein toxicity prediction methods to support the food and feed risk assessment” (reference number NP/EFSA/GMO/2018/01), awarded by the European Food Safety Authority (EFSA) to the Università degli Studi di Milano. The following tasks were defined by EFSA.

1.1.1. Task 1

- The contractor will perform a comprehensive literature search to identify and retrieve all related information/data published in peer-reviewed literature, including public protein databases, regarding proteins causing adverse effects (“toxic” proteins) in humans and animals (mammals, fish and birds).
- The contractor will describe the above identified proteins with regards to their biochemical, functional and structural properties and will detail the type of adverse effects occurring in humans and animals (mammals, fish and birds). The pathogenesis leading to the adverse effect(s) will be described, providing information on the underlying molecular mechanism of action(s), when known.
- The contractor will also provide detailed information on any identified molecular signatures (e.g. motifs, domains) associated with the above identified proteins. Moreover, the contractor will describe the role of these signatures in the context of Mode of Action or Adverse Outcome Pathway, whenever available.

1.1.2. Task 2

The contractor will provide a description (e.g. data sources used, curation methods, algorithms employed) and an evaluation (e.g. quality of curation methods, reliability of search options, user-friendliness) of the available *in silico* resources containing information on molecular signatures associated with protein toxicity (e.g. general databases such as InterPro, Protein Data Bank or PFAM databases or more specific tools such as the KNOTTIN database).

1.2. Interpretation of the Terms of Reference as provided by EFSA

Extensive and up-to-date information on proteins, including toxins, with accurate, consistent and rich annotation is contained in the UniProt protein database which can be considered the most comprehensive protein database (Bateman et al., 2017). Therefore, it was decided to base and stem the search strategy to address Task 1 on this database. Together with UniProt, literature databases were also then used to expand the search. An automated tool (Python-based script) was developed to retrieve and manage all the information and to develop an *in silico* pipeline, and it is provided to EFSA.

Moreover, preliminary work and discussion with EFSA identified the need to expand the scope of Task 1 to investigate proteins that may misfold, aggregate and/or polymerize under specific circumstances, causing pathological effects. With this aim and in agreement with EFSA, *aggregates* was identified as a term covering proteins that could be putatively classified as toxic under certain circumstances. Due to the expansion of the original question it was necessary to set new definitions and new methodologies to address this specific aspect (see Section 1.3.2).

1.3. Protein toxicity – Background information

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 a context of an offence/defence paradigm. Some human or animal endogenous proteins are also capable to induce adverse effects following polymerisation or misfolding in particular conditions (i.e. proteins with putative toxic activity as aggregates). These two categories are further described and exemplified in Sections 1.3.1 and 1.3.2 below.

1.3.1. Toxic proteins

Based on Gene Ontology (GO)¹ 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 the purpose of this report 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; these are also referred in the context of this report as proteins with a well-recognised toxic activity;
- toxins acting in the context of toxin-antitoxin systems, i.e. proteins causing a toxic effect only in case of perturbation of the toxin and antitoxin concentration equilibrium. These toxins are found only in bacteria.

Examples of proteins from the two groups are provided below in Section 1.3.1.1 and 1.3.1.2.

1.3.1.1. Proteins causing toxic effects *per se*

In plants toxic proteins have been identified throughout the plant kingdom and have also been found in edible crops; in the literature, they are sometimes defined with the alternative terms "antinutritional factors", or "antinutrients", since their ingestion can lead to interference in the absorption of nutrients.

For example, lectins are a superfamily of proteins selectively binding carbohydrates and functioning as recognition molecules in cell–molecule and cell–cell interactions in a variety of biological systems (Sharon and Lis, 2004; Miyake et al., 2007). They typically agglutinate certain animal cells and/or precipitate glycoconjugates. Lectins have been reported in legumes, tomato, potato, banana and garlic, and the ingestion of legumes containing high levels of certain lectins may be associated with gastrointestinal effects in humans and animals (Noah et al., 1980; Rodhouse et al., 1990; Bardocz et al., 1995; Grant et al., 1995). The LC(50) of the lectin alpha chain of *Dioclea grandiflora* is 2.52 µg/ml against the brine shrimp *A.salina*.

Ribosome inactivating proteins (such as ricin) showing N-glycosidase activity have been identified in several edible plants, including pumpkin, cucumber, beet, and cereals (Dang and Van Damme, 2015; Parisi et al., 2018; Vandenborre et al., 2011). They act as glycosidases that remove a specific adenine residue from an exposed loop of the 28S rRNA (A4324 in mammals), leading to rRNA breakage. As this loop is involved in elongation factor binding, modified ribosomes are catalytically inactive and unable to support protein synthesis. Each protein can inactivate a few thousand ribosomes per minute, faster than the cell can make new ones. Therefore, a single molecule can kill an animal cell.

Protease inhibitors/ α -amylase inhibitors are widely distributed in plants, mainly in storage tissues.

Thionins are cysteine-containing proteins present in a number of monocot and dicot plants causing increased cellular membrane permeability. They possess antifungal activity sensitive to inorganic cations, inducing potential changes in fungal membranes and increased K⁺ efflux and Ca²⁺ uptake.

Cyclotides and other pore-forming toxins are also described in plants.

Toxic proteins of animal origin are typically secreted in animal venoms, such as those of snakes, spiders, scorpions, cone snails, jellyfish, insects, sea anemones, lizards, a few fish and platypuses (Masood et al., 2018; Tasoulis and Isbister, 2017). These can exert their detrimental effects via their enzymatic activity (e.g. acidic phospholipase A2 RV-7of Eastern Russel's viper) or act as neurotoxins (e.g. basic phospholipase A2 ammodytoxin C of Western sand vipera). Some of these poisonous animals use venom injection systems in their predatory behaviour. Poisonous animals that lack venom injection systems, such as certain mammals, toads, ticks and worms, rely on toxins as an effective defence

¹ Gene Ontology is the main database for classifying proteins basing on their function. See Section 2.1.2 for further details.

against predation. A specific section in UniProt is devoted to animal toxin annotation (<https://www.uniprot.org/program/Toxins>). Modes of action (MoA) of these animal toxins are very diversified. As an example of MoA, we report that of phospholipase A2 P-elapitoxin-Aa1a alpha chain of *Acanthophis antarcticus*, whose heterotrimer inhibits nerve-evoked twitch contractions but not responses to cholinergic agonists acetylcholine and carbachol and to depolarizing agonist KCl.

Among **bacterial** toxic proteins there are tetanus, botulinum, and diphtheria toxins, which are relevant in human and animal pathology. Some bacterial toxic proteins are entomopathogenic, showing specificity to target species. *Bacillus thuringiensis* is a Gram positive, spore-forming bacterium that synthesizes parasporal crystalline inclusions containing Cry and Cyt proteins that have been successfully used as bioinsecticides (Yuan et al., 2019). It expresses the crystal protein Cry that promotes colloid osmotic lysis by binding to the midgut epithelial cells of insects.

1.3.1.2. Toxin-antitoxin systems

Toxin-antitoxin (TA) systems are molecular modules that are ubiquitous in free-living prokaryotes. Diverse in structure and function, TA loci comprise two genes, usually encoded in a single operon, either in the bacterial chromosome or in plasmids: one of them encodes for a stable toxin whose overexpression kills the cell or causes growth stasis; the other gene encodes for an unstable antitoxin that counteracts toxin action. The systematic balancing of the levels of expression and the interactions between toxins and antitoxins allows a number of adaptive outcomes, including stress tolerance, virulence, phage defence, and biofilm formation. Under basal conditions, the two components together do not cause any harm; accordingly, they are not associated with 'toxic activity' and deserve to be investigated and referred to separately from toxins causing toxic effects per se described above such as those of animal or vegetal origin already commented on. At the state-of-the-art, six types of TA systems are known (Durand et al., 2012 and Figure 1):

- Type I TA system. Toxin synthesis is inhibited by antisense RNA that forms base pairs with toxin mRNA. In this type of system, the toxin generally affects cellular membrane integrity, in turn impairing essential cellular functions such as cell division.
- Type II TA system. Both the toxin and antitoxin are proteins. The toxin activity is inhibited by the antitoxin by forming the TA complex. Under unfavourable conditions, cellular proteases are activated and degrade the antitoxin, which releases the toxin.
- Type III TA system. The toxin protein is bound to the antitoxin RNA, forming the RNA pseudoknot–toxin complex. When complexed, the toxin is inactivated.
- Type IV TA system. Both the toxin and antitoxin are proteins: the protein toxin destabilizes filamentous cytoskeleton proteins and inhibits cell division, whereas the antitoxin protein stabilizes filamentous cytoskeleton proteins.
- Type V TA system. The toxin peptide is involved in membrane lysis while the antitoxin is a ribonuclease specific for the toxin mRNA.
- Type VI TA system. Both the toxin and antitoxin are proteins. In the absence of the antitoxin, the toxin binds to the sliding clamp and inhibits DNA replication.

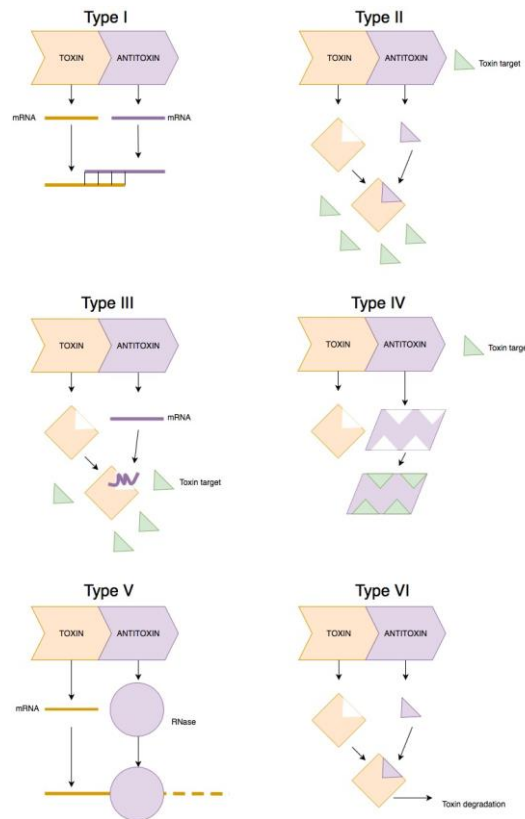


Figure 1: Models representing the interaction between toxins and antitoxins in the different types of toxin-antitoxin (TA) systems.

1.3.2. Proteins with putative toxic activity as aggregates

In the scientific literature the term 'toxic proteins' is also used for endogenous proteins owing a cellular biological role distinct from the offence/defence paradigm typical of exogenous toxins. This is due to the fact that these proteins may form structured or unstructured aggregates in tissues showing pathological conditions. The aggregation event may occur under specific conditions (including aging) either with the wild-type form or with pathological genetic variants. It is noteworthy that extensive investigations conducted in humans on these phenomena did not allow to establish a clear cause-effect relationship between the aggregation processes and the pathological conditions in all cases (Forloni and Balducci, 2018).

In general, these proteins show an abnormal folding behaviour, failing to correctly fold into stable 3D structures; this is often associated with sequence mutations. This is for instance the case of the Z variant of the plasma protease inhibitor alpha-1-antitrypsin, that – due to a charge inversion mutation, Glu342Lys – forms polymeric assemblies in the endoplasmic reticulum of hepatocytes instead of proceeding through the secretion pathway and be released into the plasma and systemic circulation. This leads to the accumulation of insoluble deposits of aggregated proteins in the liver, with functional tissue substituted by cirrhotic nodules; in parallel, a decreased plasmatic level of alpha-1-antitrypsin diminishes antiprotease protection in various tissues, with subsequent disease conditions (e.g. pulmonary emphysema) (Gooptu and Lomas, 2009; Huang et al., 2016; Lucas et al., 2018).

This is also the case of transthyretin, a tetrameric protein, composed of subunits in alpha/beta secondary structure. This protein is mainly synthesised in the liver and it is primarily responsible for the transport of thyroxine and of retinol through its complex with retinol binding protein, to different parts of the body and brain. Some genetic variants of transthyretin form unstable quaternary assemblies and, after dissociation, each subunit rearranges to an all-beta structure able to extensively interact with other similar monomers, eventually resulting in insoluble amyloid fibrils mainly in the nervous system and in

the heart. Such event may occur even with non-mutated transthyretin, in elderly people (senile systemic amyloidosis) (Benson, 2012; Saraiva, 2001).

The causes of aggregation into amyloid fibrils for some proteins in their wild-type structure seem to be related to their abnormally high concentration in tissues, often secondary to a pre-existing disease (high synthesis, reduced catabolism or disposal); and to the declining ability of aging tissues to either promote refolding of misfolded proteins (through the chaperones) or to dispose of them (through the proteasome) (Comenzo, 2006; Iadanza et al., 2018; Nativi-Nicolau and Maurer, 2018; Soto and Pritzkow, 2018). Examples of all these types of pathology-association have been documented; some are listed hereafter. Among the proteins with a tendency to form amyloid deposits, other than alpha-1-antitrypsin and transthyretin, are for instance atrial natriuretic factor (typical of senile amyloidosis of heart atria), beta-amyloid from amyloid precursor protein (typical of Alzheimer's disease), immunoglobulin light chains (in cases of multiple myeloma), serum amyloid A (in the course of inflammation), prolactin (in patients with prolactinoma), calcitonin (in patients with medullary carcinoma of the thyroid), beta-2-microglobulin (in dialysis patients), amylin (in diabetic patients), mutants of apolipoprotein A-I, fibrinogen alpha-chain or lysozyme (in familial renal amyloidosis) (Bratosiewicz-Wasik et al., 2004; Nativi-Nicolau and Maurer, 2018).

Another protein that belongs to the category of endogenous compounds described as "toxic" is the prion (PrP) (Bratosiewicz-Wasik et al., 2004; Soto and Pritzkow, 2018). PrPC is a constitutive protein present in the cell membranes. Familial, sporadic or acquired prion diseases are described. Familial and sporadic prion diseases are associated with the presence of mutated PrP. Acquired prion disease is associated with "infection" via ingestion. The infectious isoform of PrP, known as PrP^{Sc}, is able to convert PrPC proteins into PrP^{Sc} by changing their conformation from the normal alpha-helix structure to a higher proportion of beta-sheet. Aggregations of these abnormal isoforms form highly structured amyloid fibres, which accumulate to form plaques. The significance of PrPs has been highlighted following the emergence of bovine spongiform encephalopathy (BSE) as a major cattle disease in the United Kingdom. The onset of this disease was attributed to the feeding of cattle with meat- and bone-meal prepared from the carcasses of scrapie-infected sheep. Scrapie is also caused by prion proteins, as is the human equivalent - variant Creutzfeldt-Jakob disease (vCJD). The incidence of vCJD in humans has been linked to the consumption of BSE-infected beef. This association drove to extensive and stringent legislation in the European Union concerning the use of specified animal products in livestock feeding (Bratosiewicz-Wasik et al., 2004; Costanzo and Zurzolo, 2013). In the present work, the endogenous proteins above described are grouped under the heading of *proteins with putative toxic activity as aggregates*.

2. Data and Methodologies

The primary search for **toxic proteins** was based on the GO definition of "**toxin activity**", i.e. 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. The activity should refer to an evolved function of the active gene product, i.e. one that was selected for*. This search was then expanded into a secondary search using the "**toxin-antitoxin system**" (TAS) string. The above two searches were integrated with searches on 3D structures and families, domains and signatures of the identified proteins, giving rise to the Main Collection and to the Toxin-Antitoxin System (TAS) Collection, respectively. A Python-based software was created to manage all the information retrieved on the identified toxic proteins and associated literature. The *in silico* pipeline developed to address the EFSA Tasks is presented in Figure 2.

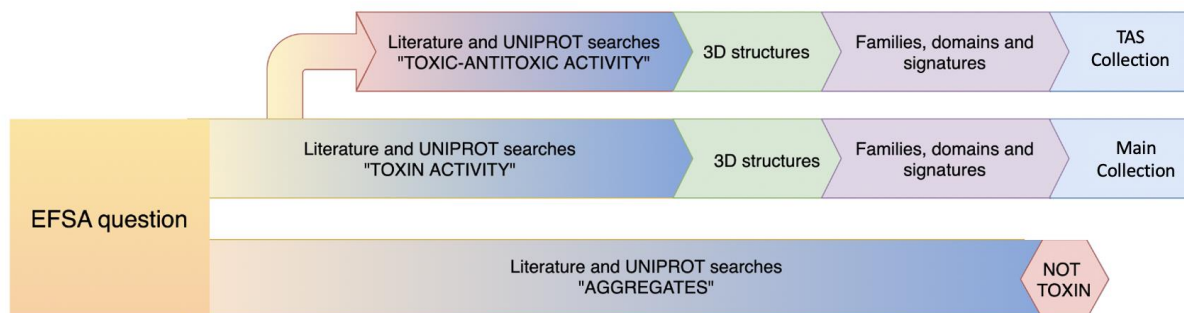


Figure 2: *In silico* pipeline devised to fulfil the EFSA request.

Furthermore, the main freely available on-line tools for predicting protein toxicity were carefully tested to evaluate their performance.

In addition, based on a preliminary work searching Pubmed with the search string “protein toxicity” and discussion with EFSA, **proteins with putative toxic activity as aggregates** were identified. These were treated separately, and specific definitions and methodologies to address this specific issue were set (Section 2.2.2). Details on the databases used and on the searching strategies are described in Sections 2.1 -2.5 below.

2.1. Databases

To perform a comprehensive search and to retrieve all the related information, peer-reviewed literature databases (Section 2.1.1) and protein databases containing information on protein primary structure and function (Section 2.1.2) were consulted. These databases were used for all the searches on toxic proteins (i.e. proteins showing *toxin activity* or belonging to *toxin-antitoxin system*) and on aggregates. To integrate the outcome of the above searches with information on the 3D structures and families, domains and signatures², additional databases were consulted (Sections 2.1.3 and 2.1.4). These databases contain information on toxic proteins as defined above and therefore were used only for *toxin activity* and *toxin-antitoxin system* proteins. Further details on these databases can be found in the respective web pages (see links in paragraphs below).

Section 2.1.5 briefly describes the toxin prediction tools available and the tests conducted to evaluate their performance in predicting the toxicity of unknown proteins.

2.1.1. Literature databases

The ‘Web of Science’, ‘Scopus’ and ‘Pubmed’ databases (Table 1) were used to search for publications (both primary research and review articles) providing information on the review questions.

Table 1: Information sources used in the carried out extensive literature searches.

Source	Link
Web of Science™	www.webofknowledge.com
Scopus	www.scopus.com
PubMed	www.ncbi.nlm.nih.gov/pubmed/

These databases are regarded as comprehensive sources of information by the scientific community (Figure 3) and were thus used to search for both toxic proteins and aggregates.

² In the present report the term “signature” is used to describe a specific sequence pattern, motif or fingerprint, according to EMBL definition. Signatures inform on conserved residues that may be essential for stability or function of the protein.

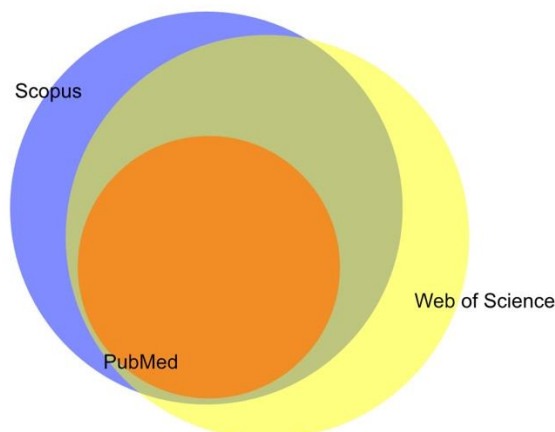


Figure 3: Literature databases sources diagram.

Below a brief description extracted from the webpage of each source used for the literature search.

WEB OF SCIENCE™ CORE COLLECTION (Clarivate Analytics)

Web of Science™ Core Collection provides multidisciplinary content covering over 12,000 of the highest impact journals worldwide (including open access journals) and over 160,000 conference proceedings. The collection features current and retrospective coverage in the sciences, social sciences, arts, and humanities, with coverage dating back to 1900.

SCOPUS (Elsevier)

More than 60 million records are indexed in Scopus, including over 21,500 peer-reviewed journals, of which more than 4,200 are full open access. Scopus indexes also articles-in-press (i.e., articles that have been accepted for publication) from more than 5,000 international publishers.

PUBMED (NCBI)

PubMed comprises over 25 million citations for biomedical literature from MEDLINE, life science journals, and online books. PubMed citations and abstracts include the fields of biomedicine and health, covering portions of the life sciences, behavioural sciences, chemical sciences, and bioengineering. PubMed also provides access to additional relevant web sites and links to the other NCBI molecular biology resources.

2.1.2. Protein Databases - Primary structure and function

The most important protein information databases UniProt and GO (Table 2) were used to search for keywords feeding the review questions. These two databases were used to search for both toxic proteins and aggregates. In fact, UniProt is the gold standard source of information for proteins, while GO is the main database for classifying proteins based on their function. These two databases are interconnected, since UniProt contains the GO classification but it does not contain the GO definitions.

Table 2: Information sources used in the carried out extensive primary structure searches.

Source	Link
UNIPROT	www.uniprot.org
GENE ONTOLOGY	www.geneontology.org

The paragraphs below provide a brief description of these databases, extracted from the Uniprot and GO webpages. Further details on these databases can be found in their respective webpage (see links in Table 2).

UNIPROT

Uniprot (Consortium, 2008) is a freely accessible database whose mission is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information, with many entries being derived from genome sequencing projects. UniProt is

composed of three core (sub)databases: UniProtKB (with sub-parts Swiss-Prot and TrEMBL), UniParc and UniRef. In particular, the Swiss-Prot repository contains manually annotated and reviewed records with information extracted from literature and curator-evaluated computational analysis.

A schematic representation of the Uniprot database is shown in Figure 4.

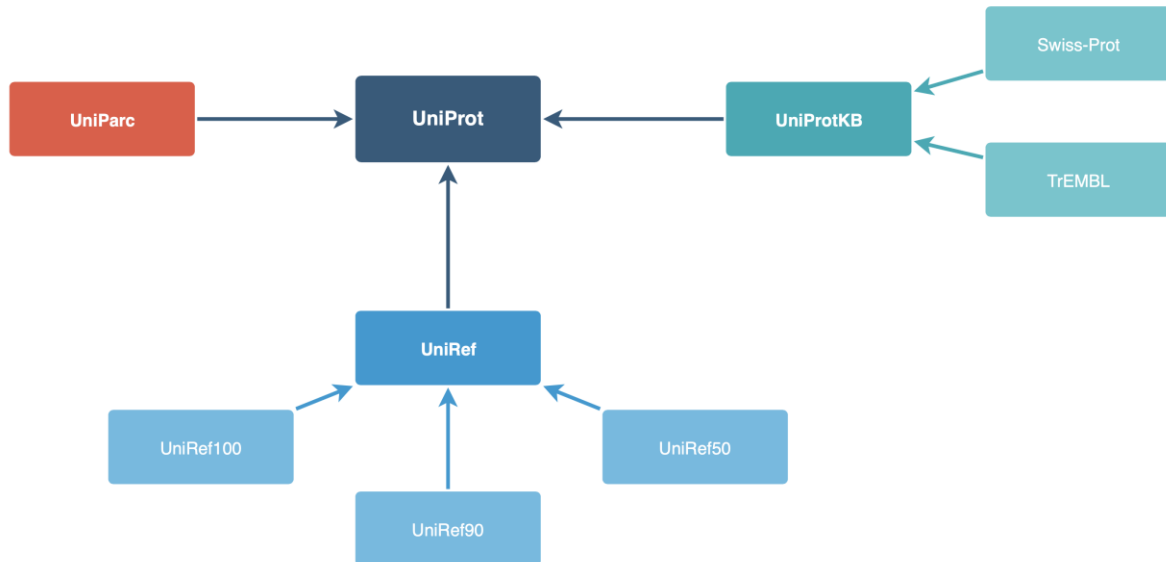


Figure 4: Schematic representation of the UniProt database.

GENE ONTOLOGY

The Gene Ontology (GO) knowledgebase (Ashburner et al., 2000; Carbon et al., 2019) is the world's largest source of information on the functions of genes. The mission of the GO Consortium is to develop a comprehensive, computational model of biological systems, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life. The ontology covers three domains: biological processes, molecular functions and cellular components.

A schematic representation of the GO database is shown in Figure 5.

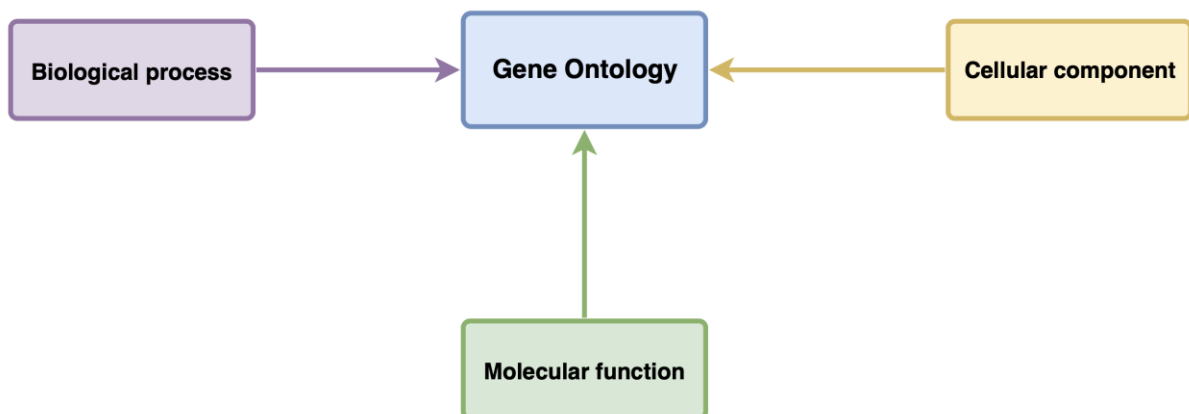


Figure 5: Schematic representation of the GO database.

2.1.3. Protein databases - Modelling tools

The databases containing 3D protein structures (either experimentally derived or predicted based on homology modelling) that were used in this report are listed in Table 3. The 'Protein Data Bank (PDB)' is the reference database for all experimentally determined structures, whereas the 'Swiss-Model'

repository is a reference database for predicted models. In addition, 'CPH MODEL' is one of the most relevant, freely available tools for peptide and protein modelling. As indicated above, these 3D-structure oriented databases were used to search only for toxic proteins (*toxin activity* and *toxin-antitoxin system* searches). A brief description of these databases is provided below; this is extracted from the respective webpages, where further information can be found (see links in Table 3).

Table 3: Information sources used for 3D structure searches.

Source	Link
Protein Data Bank	www.rcsb.org
Swiss-Model	www.swissmodel.expasy.org
CPH MODEL	www.cbs.dtu.dk/services/CPHmodels/

PROTEIN DATA BANK

The PDB archive (Berman et al., 2002; Burley et al., 2019) is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy are used to determine the location of each atom relative to each other in the molecule. This information is deposited, annotated and publicly released into the archive by the wwPDB.

SWISS-MODEL

SWISS-MODEL (Benkert et al., 2011; Bienert et al., 2017; Waterhouse et al., 2018) is a structural bioinformatics web-server dedicated to homology modelling of 3D protein structures. Homology modelling is currently the most accurate method to generate reliable three-dimensional protein structure models and is routinely used in many practical applications. Homology (or comparative) modelling methods make use of experimental protein structures ("templates") to build models for evolutionary related proteins ("targets").

CPH MODEL SERVER

CPH models 3.2 (Nielsen et al., 2010) is a protein homology modelling server. The template recognition is based on profile-profile alignment guided by secondary structure and exposure predictions. The CPH models is an easy to use web server for comparative protein homology modelling.

2.1.4. Families, domains and signatures

Protein-families, domains- and signatures-related databases used in this report are listed in Table 4 and are described below. All the listed databases are interconnected under the InterPro consortium, which classifies proteins into families, domains and other important motifs/sites (e.g. binding sites) and provides protein functional information for each entry based on those classifications. In fact, although the databases are independent of each other, the InterPro consortium is a collector of this information, especially with regards to domains. InterPro provides information both original and derived from the Consortium members. At the state of the art, InterPro is the most important database (Li et al., 2015) to classify protein domains and signatures while Pfam is the most important database for protein family classification (Finn et al., 2016). InterPro and Pfam are overlapping databases since a Pfam member (a family of proteins) can contain one or more InterPro entries (domains). To get the broadest view, we used both the databases, although some of the data are recurrent between the two of them. As above indicated, these databases were used only for toxic proteins (*toxin activity* and *toxin-antitoxin system* searches). Further details on these databases can be found in their respective web page (see links in Table 4).

Table 4: Information sources used to carry out the domains, families and signatures searches.

Source	Link
Interpro	https://www.ebi.ac.uk/interpro/beta/
Pfam	http://pfam.xfam.org
PROSITE	https://prosite.expasy.org
CATH-GENE3D	http://www.cathdb.info
SUPFAM	http://supfam.org

PRINTS	http://130.88.97.239/PRINTS/index.php
SMART	http://smart.embl-heidelberg.de
PANTHER	http://pantherdb.org
TIGRFAMS	https://www.jcvi.org/research/tigrfams
PIRSF	https://proteininformationresource.org/pirwww/dbinfo/pirsf.shtml
CDD	https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

INTERPRO

InterPro (Goujon et al., 2010; Li et al., 2015) is a database of protein families, domains and functional sites, in which identifiable features found in known proteins can be applied to new protein sequences in order to functionally characterize them. The contents of *InterPro* consist of diagnostic signatures and the proteins that they significantly match. Additional information such as a description, consistent names and Gene Ontology (GO) terms, are associated with each entry.

In particular, *InterPro* integrates predictive signatures representing functional sites into a single resource from a number of family and domain databases: Gene3D, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY and TIGRFAMS. *InterPro* uses a pattern-based approach to classify the signatures and the information it contains was curator-evaluated.

PFAM

Pfam (Finn et al., 2016) is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models. *Pfam* is produced at the European Bioinformatics Institute using a sequence database called *Pfamseq*, which is based on UniProtKB.

PROSITE

PROSITE (de Castro et al., 2006; Hulo et al., 2008; Sigrist et al., 2013, 2002) is a protein database that consists of entries describing the protein families, domains and functional sites as well as amino acid patterns and profiles in them. *PROSITE*'s uses include identifying possible functions of newly discovered proteins and analysis of known proteins for previously undetermined activity. Properties from well-studied genes can be propagated to biologically related organisms and, for different or poorly known genes, biochemical functions can be predicted from similarities. *PROSITE* uses tools for protein sequence analysis and motif detection.

CATH-GENE3D

CATH-Gene3D (Dawson et al., 2017) describes protein families and domain architectures in complete genomes. Protein families are formed using a Markov clustering algorithm, followed by multi-linkage clustering according to sequence identity. Mapping of predicted structure and sequence domains is undertaken using hidden Markov models libraries representing *CATH* and *Pfam* domains.

SUPFAM

SUPERFAMILY (Wilson et al., 2009) is a database of structural and functional annotation for all proteins and genomes. It classifies amino acid sequences into known structural domains. Domains are functional, structural, and evolutionary units that form proteins. Domains of common ancestry are grouped into superfamilies. Superfamilies are groups of proteins which have structural evidence to support a common evolutionary ancestor but may not have detectable sequence homology.

PRINTS

The *PRINTS* (Attwood et al., 1994) database houses a collection of protein family "fingerprints". It provides both a detailed annotation resource for protein families and a diagnostic tool for newly determined sequences. A fingerprint is a group of conserved motifs taken from a multiple sequence alignment - together, the motifs form a characteristic signature for the aligned protein family. The diagnostic strength of fingerprints consists in their ability to distinguish sequence differences at the clan, superfamily, family and subfamily levels. This allows functional diagnoses of uncharacterized sequences, allowing discrimination between family members based on the ligands they bind or the proteins with which they interact.

SMART

SMART (Simple Modular Architecture Research Tool) (Carnate and Ed, 2008) is a biological database that is used in the identification and analysis of protein domains within protein sequences. *SMART* uses

profile-hidden Markov models built from multiple sequence alignments to detect protein domains in protein sequences. The most recent release of SMART contains 1,204 domain models.

PANTHER

PANTHER (Protein ANalysis THrough Evolutionary Relationships) (Thomas et al., 2003) is a large collection of protein families that have been subdivided into functionally related subfamilies. The subfamilies model the divergence of specific functions within protein families, allowing more accurate association with function as well as inference of amino acids important for functional specificity. Hidden Markov models (HMMs) are built for each family and subfamily for classifying additional protein sequences.

TIGRFAMs

TIGRFAMs (Haft et al., 2013) is a database of protein families designed to support manual and automated genome annotation. Each entry includes a multiple sequence alignment and hidden Markov model (HMM) built from the alignment. Sequences that score above the defined cut-offs of a given TIGRFAMs HMM are assigned to that protein family and may be assigned the corresponding annotations. TIGRFAMs uses the HMMER package; the current version has 4488 models.

PIRSF

The primary PIRSF (Nikolskaya et al., 2006) classification unit is the homeomorphic family, whose members are both homologous (evolved from a common ancestor) and homeomorphic (sharing full-length sequence similarity and a common domain architecture).

CDD

CDD (Conserved Domain Database) (Marchler-Bauer et al., 2017) consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes domains, which use 3D-structure information to define domain boundaries and provide insights into sequence, structure, and function relationships.

2.1.5. Toxin predicting tools

Using the keywords from previous searches and adding “tool” and/or “prediction” terms, we selected the main freely available on-line tools (Table 5) from both a web search and the available scientific literature. Deprecated or not documented tools were discarded. These tools were carefully tested using both Main and TAS Collections in order to evaluate their performance in toxins prediction. A brief description of these tools is provided below; this is extracted from the respective webpages, where further information can be found (see links in Table 5).

Table 5: Toxin predictors - Selected tools

Source	Link
NTXpred	http://crdd.osdd.net/raghava/ntxpred/
BTXpred	http://crdd.osdd.net/raghava/btxpred/
KNOTTIN	http://www.dsimb.inserm.fr/KNOTTIN/
CLANTOX	http://www.clantox.cs.huji.ac.il
CONOSERVER	http://www.conoserver.org
ToxinPred	https://webs.iitd.edu.in/raghava/toxinpred/index.html

NTXpred

The aim of NTXpred server is to predict neurotoxins and their probable function from primary amino acid sequence using SVM based on composition and PSI-Blast. Neurotoxins are key players in science and medicine and are used in ion channels and receptor studies, drug discovery, and formulation of insecticides.

BTXpred

The aim of BTXpred server is to predict bacterial toxins and their function from primary amino acid sequence using SVM, HMM and PSI-Blast. Bacterial toxins play a major role in causing disease and are responsible for the majority of symptoms and lesions during infections.

KNOTTIN

This tool predicts whether a protein is a knottin, using the primary (Knotter1D) or the 3D (Knotter3D) structures of the query. It is also able to build knottin models (Knotter1D3D).

CLANTOX

ClanTox is a classifier of animal toxins that, given a protein sequence, tries to predict whether it represents a toxin or a toxin-like protein. The output of ClanTox is reported as a pair of numbers: the mean score and the standard deviation.

CONOSERVER

CONOSERVER is a database specializing in sequence and structure of conopeptides, which are peptides expressed by carnivorous marine cone snails. Conopeptides are classified into disulfide rich (conotoxins) and several classes of disulfide poor peptides. The database uses three different schemes to classify conotoxins.

ToxinPred

*ToxinPred is an *in silico* method, developed to predict and design toxic/non-toxic peptides. The main dataset used in this method consists of 1805 toxic peptides (≤ 35 residues).*

2.2. Methodologies

The strategy applied for both the literature and protein databases searches described in this report followed three key principles:

1. methodological rigour and coherence in the retrieval and selection of studies;
2. reproducibility;
3. transparency.

To ensure that these principles were implemented in the searches, the methods and techniques described in the EFSA guidance on application of systematic review methodology to food and feed safety assessments to support decision-making (EFSA, 2010) were applied. The most up-to-date searches were conducted in March 2020.

Section 2.2.1 describes the key steps of the search strategy to identify toxic proteins to populate both the Main and TAS collections. For these proteins a Python script was produced to automatically download all the available structures and models, to collect all the respective UniProtKB-based information and to cluster the identified proteins. This Python script together with the associated user guide and an illustrative video are provided to EFSA as an Annex to the final report.

Section 2.2.2 describes the approach to the search for proteins with putative toxic activity as aggregates. This differs from the previous searches, since aggregation-prone proteins cannot be classified as toxins *per se* based on scientific evidence; moreover, no 3D structures were collected for these entries.

For both approaches, the starting point was a UniProtKB search; the search terms were defined based on the GO database terminology and subsequently refined (see details below). Three different literature-containing databases were queried to retrieve relevant information (Table 1).

All the related information from UniProtKB (e.g. keywords, GOs, primary structure-related information, organism of origin) was extracted and stored in separate lists (the respective Excel tables containing these lists supplement this report).

2.2.1. Compilation of a comprehensive collection of information for toxic proteins

Six steps were undertaken to compile a comprehensive collection of information on toxic proteins (Figure 6).

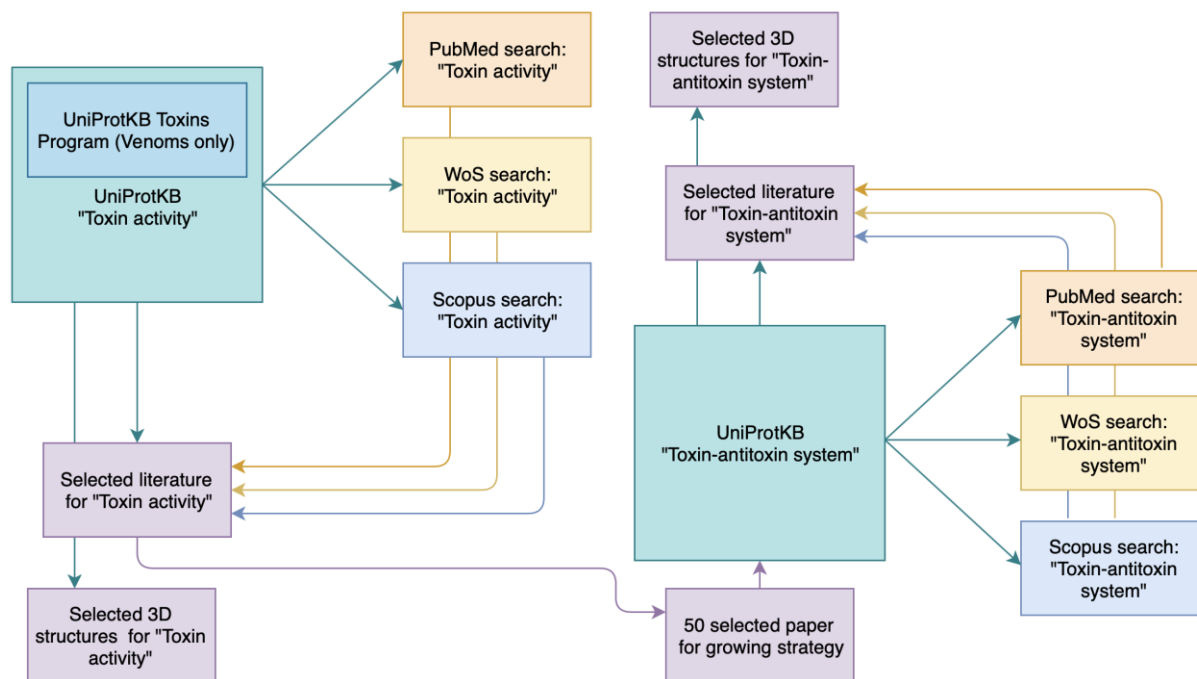


Figure 6: Schematic representation of the comprehensive protein toxin search.

These steps included a primary search to identify toxic proteins in UniprotKB (Step 1) and in literature databases (Step 2); a secondary search expanding on toxin-antitoxin systems in UniprotKB (Step 3); a consistency check between the papers identified in the above searches and those from literature databases, using the same keywords (Step 4); a robustness check testing different search terms (step 5); and finally data extrapolation and clustering (Step 6). A detailed description of Steps 1-6 is provided below.

Step 1: Searches in UniProt database

Strategy: To answer the question “Which proteins are associated with a well-recognized toxic activity (i.e. cause toxic effects *per se*?”), we made reference to the term “toxin activity” found in GO: proteins showing toxin activity are those 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. The activity should refer to an evolved function of the active gene product, i.e. one that was selected for.* The main search for toxic proteins was carried out querying the UniProtKB database. Papers linked to the retrieved Uniprot entries were selected as relevant literature and included in the Main Collection. Full details of the search are reported in Box 1 below.

Box 1

<u>Question</u>	Which proteins are associated with a well-recognized toxic activity (i.e. cause toxic effects <i>per se</i>)?
<u>Keyword</u>	To select relevant studies concerning proteins with associated toxic activity <i>per se</i> 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”
<u>Rationale</u>	The search strategy was based on the term “Toxin activity” described in GO Annotation database (Table 6), 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, as mentioned above. The full definition of this term as provided by the GO is reported in the table below.
--	---

Table 6: “Toxin activity” term in Gene Ontology

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:	<i>is-a</i> molecular function
Category:	Molecular Function
Id:	GO:0090729

Results: The proteins identified from “Toxin activity and Reviewed:Yes” string search constitute the Main Collection (Table 7). As expected, the GO “toxin activity” term was present in each entry and was used as the main term for the Main Collection search. For the Main Collection we selected only the annotated proteins (Annex A), with the associated literature already revised by UniProtKB curators. Non-annotated protein entries and associated literature were stored into a separate database (Annex B). We do not consider the non-annotated proteins as part of the Main Collection. With this strategy we extended the collection of protein toxins already identified in the UniprotKB “Toxin annotation project” for toxins of animal origin to also include bacterial and plant toxins while verifying at the same time that all the animal toxins annotated within the UniprotKB project were included in the Main Collection.

Table 7: Number of selected toxins for Main Collection (March 2020)

<i>UniProt string</i>	<i>Reviewed:YES</i>	<i>Reviewed:NO</i>
“Toxin activity”	6,964	47,831

Step 2: Literature database search

Literature sources (Scopus, Web of Science and PubMed) were also queried using the “toxin activity” as the search term, with the aim of verifying that all the papers selected from Step 1 are also retrievable in these literature databases. For each entry, the article DOI (Digital Object Identifier) was used as the unique identifier and authors, title, journal and source were also considered. Table 7 lists the number of papers found for each search string using PubMed, Web of Science and Scopus, respectively. Results are reported in Table 8. The identified papers were then cross-checked (please, see the Step 4).

Table 8: Papers in the three source literature databases for the Main Collection

<i>String search</i>	<i>PubMed</i>	<i>WOS</i>	<i>SCOPUS</i>
“Toxin activity”	570	529	709
toxin activity	77,726	34,935	50,214

Step 3: Citation pearl growing strategy using publications known to be landmark publications in the field

Strategy: Using a pearl growing strategy we extended our search to proteins relevant in the context of this work but not associated to the GO term ‘*toxic activity*’. From the publications in the Main Collection, 50 highly representative papers, considered as landmark publications in the field based on both citations and number of toxins discussed, were selected and reviewed (Table 9). The selection of the papers was based on the number of toxins cited (avoiding repetition of organism of origin or specific protein), and the presence of a description of toxic effects/pathways. Among these, five papers concerning *Mycobacterium tuberculosis* and *Escherichia coli* were included since these bacteria resulted associated with the highest number of bacterial toxic proteins (see the “Organism” Excel file of the TAS Collection for details on the toxins from these bacteria). Based on these papers, a new keyword was identified as

relevant (see Box 2 and Table 10) and an additional search was run in the UniProtKB database, giving rise to the TAS Collection. The new search covers additional bacterial toxins, that are not comprehensively present in the Main Collection. Cold shock-like protein CspD from *E.coli* and Ribonuclease VapC2 from *M. tuberculosis* are the only two proteins to have associated both GO terms "Toxin activity" and "Toxin-antitoxin" system.

Box 2

Question	How to extend the toxic protein Main Collection with UniProtKB annotated proteins taking into account 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 <i>toxin activity</i> , 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 <i>per se</i> , and that interact primarily with proteins of the target organism. Toxins that belong to TAS primarily interact with their antitoxins and are able to 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.

Table 9: List of the 50 selected papers.

UniProt	Pubmed ID	Title
POC1N6	11158371	Mechanisms for evolving hypervariability: the case of conopeptides.
P59939	12459475	A large number of novel Erg toxin-like genes and ERG K ⁺ -channels blocking peptides from scorpions of the genus <i>Centruroides</i> .
P60210	15025998	Proteomics of the venom from the Amazonian scorpion <i>Tityus cambridgei</i> and the role of prolines on mass spectrometry analysis of toxins.
PODJM6	15032748	Molecular evolution and structure-function relationships of crotoxin-like and asparagine-6-containing phospholipases A2 in pit viper venoms.
POCF14	15688451	A novel strategy for the identification of toxin-like structures in spider venom.
A8S6B3	16261251	Identification and analysis of venom gland-specific genes from the coastal taipan (<i>Oxyuranus scutellatus</i>) and related species.
O76199	16278100	Comparison of the partial proteomes of the venoms of Brazilian spiders of the genus <i>Phoneutria</i> .
Q09GJ9	16292255	Early evolution of the venom system in lizards and snakes.
POCI13	16477526	Genes expressed in a turrid venom duct: divergence and similarity to conotoxins.
Q1A3Q1	16908117	Diversity and evolution of conotoxins based on gene expression profiling of <i>Conus litteratus</i> .
POC8B6	17320133	Venomic analyses of <i>Scolopendra viridicornis nigra</i> and <i>Scolopendra angulata</i> (Centipede, Scolopendromorpha): shedding light on venoms from a neglected group.
A0SE59	17400270	From the identification of gene organization of alpha conotoxins to the cloning of novel toxins.
B1P1A4	17476710	Proteomic and peptidomic analysis of the venom from Chinese tarantula <i>Chilobrachys jingzhao</i> .
A7X3M3	17855442	Evolution of an arsenal: structural and functional diversification of the venom system in the advanced snakes (<i>Caenophidia</i>).

A7SCE5	18222944	Concerted evolution of sea anemone neurotoxin genes is revealed through analysis of the <i>Nematostella vectensis</i> genome.
B1P1A0	18581053	Molecular diversity and evolution of cystine knot toxins of the tarantula <i>Chilobrachys jingzhao</i> .
B2ZBB8	18923708	Discovery of a distinct superfamily of Kunitz-type toxin (KTT) from tarantulas.
COJAQ5	19042943	Molecular evolution, functional variation, and proposed nomenclature of the gene family that includes sphingomyelinase D in sicariid spider venoms.
B4XSY4	19059426	C-type lectin protein isoforms of <i>Macrovipera lebetina</i> : cDNA cloning and genetic diversity.
B9W5G6	19268680	Molecular mechanism of pore formation by actinoporins.
P00624	19371136	Exploring the venom proteome of the western diamondback rattlesnake, <i>Crotalus atrox</i> , via snake venomomics and combinatorial peptide ligand library approaches.
POC1N6	19380747	Rapid sensitive analysis of cysteine rich peptide venom components.
P0DL61	19627569	Comprehensive EST analysis of the symbiotic sea anemone, <i>Anemonia viridis</i> .
B6DCJ0	19875276	Transcriptome analysis of the venom glands of the Chinese wolf spider <i>Lycosa singoriensis</i> .
D2Y1X6	20192277	Molecular diversification of peptide toxins from the tarantula <i>Haplopelma hainanum</i> (<i>Ornithoctonus hainana</i>) venom based on transcriptomic, peptidomic, and genomic analyses.
A6YR40	20363338	Evolution of <i>Conus</i> peptide toxins: analysis of <i>Conus californicus</i> Reeve, 1844.
C6ZH27	20663230	Comparative venom gland transcriptome analysis of the scorpion <i>Lychas mucronatus</i> reveals intraspecific toxic gene diversity and new venomous components.
D2Y100	21172372	Diversity of conotoxin types from <i>Conus californicus</i> reflects a diversity of prey types and a novel evolutionary history.
C1J5M6	21266071	Characterization of the <i>Conus bullatus</i> genome and its venom-duct transcriptome.
P0DMX6	21281459	The mining of toxin-like polypeptides from EST database by single residue distribution analysis.
C6JUP1	21515432	Snake venomomics and venom gland transcriptomic analysis of Brazilian coral snakes, <i>Micrurus altirostris</i> and <i>M. corallinus</i> .
C9X4J8	22355312	Identification and phylogenetic analysis of <i>Tityus pachyurus</i> and <i>Tityus obscurus</i> novel putative Na ⁺ -channel scorpion toxins.
I6R1R5	22595790	Chemical punch packed in venoms makes centipedes excellent predators.
B3EWS5	22672445	Multicomponent venom of the spider <i>Cupiennius salei</i> : a bioanalytical investigation applying different strategies.
A7LCJ2	22683676	Development of a rational nomenclature for naming peptide and protein toxins from sea anemones.
POC1N6	22709442	Constrained de novo sequencing of conotoxins.
P01022	22869554	Peptidomics of three <i>Bothrops</i> snake venoms: insights into the molecular diversification of proteomes and peptidomes.
A0A0R4I951	23148443	Venomomic and transcriptomic analysis of centipede <i>Scolopendra subspinipes dehaani</i> .
F8S0Y4	24231107	Linking the transcriptome and proteome to characterize the venom of the eastern diamondback rattlesnake (<i>Crotalus adamanteus</i>).
A3R0T9	24297900	The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system.
W4VRU3	24351713	A proteomics and transcriptomics investigation of the venom from the barychelid spider <i>Trittame loki</i> (brush-foot trapdoor).
A0A023IWD9	24613547	The molecular diversity of toxin gene families in lethal <i>Amanita</i> mushrooms.
A0A023VZR2	24847043	Clawing through evolution: toxin diversification and convergence in the ancient lineage <i>Chilopoda</i> (centipedes).
P0DN44	26025559	Molecular diversity and gene evolution of the venom arsenal of <i>Terebridae</i> predatory marine snails.
B3EWF2	27287558	<i>Lachesana tarabaevi</i> , an expert in membrane-active toxins.

P9WFB9	15718296	Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes.
P69348	19028895	Influence of operator site geometry on transcriptional control by the YefM-YoeB toxin-antitoxin complex
Q47149	17263853	Escherichia coli dinJ-yafQ genes act as a toxin-antitoxin module
O07227	20011113	Comprehensive functional analysis of <i>Mycobacterium tuberculosis</i> toxin-antitoxin systems: implications for pathogenesis, stress responses, and evolution.
Q47150	19707553	A differential effect of <i>E. coli</i> toxin-antitoxin systems on cell death in liquid media and biofilm formation.

Table 10: Definition of toxin-antitoxin systems (Unterholzner et al., 2013)

Term:	Toxin-antitoxin systems (TAS)
Definition:	TAS were defined as follows (Unterholzner et al., 2013): "TAS are small genetic elements composed of a toxin gene and its cognate antitoxin. The toxins of all known TAS are proteins while the antitoxins are either proteins or non-coding RNAs. Based on the molecular nature of the antitoxin and its mode of interaction with the toxin, the TAS modules are currently grouped into six classes. In general, the toxin is more stable than the antitoxin but the latter is expressed to a higher level. If supply of the antitoxin stops, for instance under special growth conditions or by plasmid loss in case of plasmid-encoded TAS, the antitoxin is rapidly degraded and can no longer counteract the toxin. Consequently, the toxin becomes activated and can act on its cellular targets. Typically, TAS toxins act on crucial cellular processes including translation, replication, cytoskeleton formation, membrane integrity, and cell wall biosynthesis. TAS and their components are also versatile tools for a multitude of purposes in basic research and biotechnology. Currently, TAS are frequently used for selection in cloning and for single protein expression in living bacterial cells. Since several TAS toxins exhibit activity in yeast and mammalian cells they may be useful for applications in eukaryotic systems. TAS modules are also considered as promising targets for the development of antibacterial drugs and their potential to combat viral infection may aid in controlling infectious diseases."

Results: To populate the TAS Collection we selected only the annotated proteins reviewed by UniProtKB curators (Annex C), while non-annotated proteins and associated literature were stored into a separate database (Annex D). Unreviewed proteins are not part of the TAS Collection. Following the Step 2 procedure, we queried the literature sources (Scopus, Web of Science and PubMed) using the "toxin-antitoxin system" as the search term. Table 11 lists the number of selected toxins from UniProtKB that populated the TAS Collection, while Table 12 lists the number of papers found for each search string using PubMed, Web of Science and Scopus, respectively.

Table 11: Number of selected toxins for TAS Collection (March 2020)

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

Table 12: Papers in the three source literature databases for the TAS Collection

String search	PubMed	WOS	SCOPUS
"Toxin-antitoxin system"	365	420	776
Toxin-antitoxin system	1,169	1,260	1,060

Step 4: Literature search consistency

In order to cross-check literature information identified through the UniProtKB and literature database (PubMed, WOS and Scopus) searches for both Main and TAS Collections, we verified that all the papers found from the UniProtKB searches (Step 1 & 3) could also be found in either PubMed, WOS or Scopus

searches (Step 2 & 3). Table 13 reports the number of selected papers associated to each collection. Duplicate papers were removed.

Table 13: Number of selected papers (March 2020)

Search string	Reviewed: Yes	Reviewed: No	Collection
"Toxin activity"	5430	2143	Main
"Toxin-AntiToxin System"	482	1970	TAS

Step 5: Exclusion of other search terms

We also verified the robustness of the selected search strings (Step 1 and Step 2) in covering all the proteins marked as toxic in the UniProtKB database, confirming that by using different word combinations (e.g. toxic and proteins, toxic and animals, toxic and enzymes) searching UniProtKB it would not result in the identification of additional proteins classifiable as toxins, as shown in Table 14.

Table 14: Exclusion of search terms evaluation

Search string	Findings
"Toxic" and its combination with other words	Using this string leads to the retrieval of toxic proteins <i>per se</i> , which are included in 'Toxin Activity'; and proteins not relevant to the scope of the search e.g. Q9LPV4 – Protein detoxification 31(DTX31_ARATH)
"Toxicity" and its combination with other word	Proteins with associated toxicity are included in 'Toxin Activity', but some of them are not related with the aim of the search e.g. P33302 - Pleiotropic ABC efflux transporter of multiple drugs (PDR5_YEAST)
"Pathogen" and its combination with other word	Proteins with associated toxicity are included in 'Toxin Activity', but some of them are not related with the aim of the search e.g. Q94C26 - Cysteine-rich and transmembrane domain-containing protein PCC1 (PCC1_ARATH)
"Toxin"	It includes Toxin Activity, Toxin-antitoxin but also some other proteins that are not toxic <i>per se</i> e.g. O00476 - Sodium-dependent phosphate transport protein 4 (SLC17A3), associated with the toxin transmembrane transporter activity (GO: 0019534)

Step 6: Data extrapolation and clusters

Information from UniProtKB for both Collections (Main and TAS) was stored into different Excel files, for possible further analyses (see Table 15). The proteins in these collections were clustered based on the annotated UniprotKB flags.

Three different flags were used to group the proteins:

- Keyword, that contains all the UniProt-related keywords;
- GO, that contains all the GO-related labels; and
- Organism, that contains all the organisms that express the selected proteins.

Table 15 lists the Excel files and Annexes delivered with this Final Report as part of results.

Table 15: Files and Annexes delivered with the Report

Collection	File³/Annex	Contents
Main Collection	MC_citation_scopes.xlsx	Citations with statistics of UniProtKB scopes
Main Collection	MC_citations.xlsx	Citations
Main Collection	MC_comments.xlsx	UniProtKB comments
Main Collection	MC_databases.xlsx	Families, domains and signatures classification according to databases
Main Collection	MC_go.xlsx	GO terms
Main Collection	MC_keywords.xlsx	UniProtKB keywords

³ Supplementary files in Folder#1

Main Collection	MC_organisms.xlsx	Organisms from UniProtKb
Main Collection	MC_pdb.xlsx	Available experimentally solved structures (from RCSB Protein Data Bank)
Main Collection	MC_Pfam_identities.xlsx	Identities in databases with respect to Pfam
Main Collection	MC_sequences.xlsx	Toxin sequences
Main Collection	MC_swiss_models.xlsx	Models from the Swiss Model Repository
TAS Collection	TC_citation_scopes.xlsx	Citations with statistics of UniProtKB scopes
TAS Collection	TC_citations.xlsx	Citations
TAS Collection	TC_comments.xlsx	UniProtKB comments
TAS Collection	TC_databases.xlsx	Families, domains and signatures classification according to databases
TAS Collection	TC_go.xlsx	GO terms
TAS Collection	TC_keywords.xlsx	UniProtKB keywords
TAS Collection	TC_organisms.xlsx	Organisms from UniProtKb
TAS Collection	TC_pdb.xlsx	Available experimentally solved structures (from RCSB Protein Data Bank)
TAS Collection	TC_Pfam_identities.xlsx	Identities in databases with respect to Pfam
TAS Collection	TC_sequences.xlsx	Toxin sequences
TAS Collection	TC_swiss_models.xlsx	Models from the Swiss Model Repository
TAS Collection	TC_citation_scopes.xlsx	Citations with statistics of UniProtKB scopes
Main Collection	Annex A	Results for Toxin Activity, Reviewed:YES
	Annex B	Results for Toxin Activity, Reviewed:NO
TAS Collection	Annex C	Toxin-antitoxin system, Reviewed:YES
	Annex D	Toxin-antitoxin system, Reviewed:NO
Main and TAS Collections	Annex E	Pfam results
Main and TAS Collections	Annex F	Interpro results

2.2.2. Proteins with putative toxic activity as aggregates

This section illustrates the search strategy that has been introduced in this project to expand the scope of Task 1, in order to also investigate proteins that may misfold and/or polymerize under specific circumstances (“aggregates”) causing pathological effects, as requested by EFSA and specified in section 1.2. These proteins are physiological components of the organism and do not fit with the definition in the GO “toxin activity” (see Table 6), since they act in the same organisms that express them. Consistently with these observations, none of the protein aggregates was retrieved with the searches carried out during the previous steps of this work. Accordingly, a dedicated search strategy was set up (Figure 7).

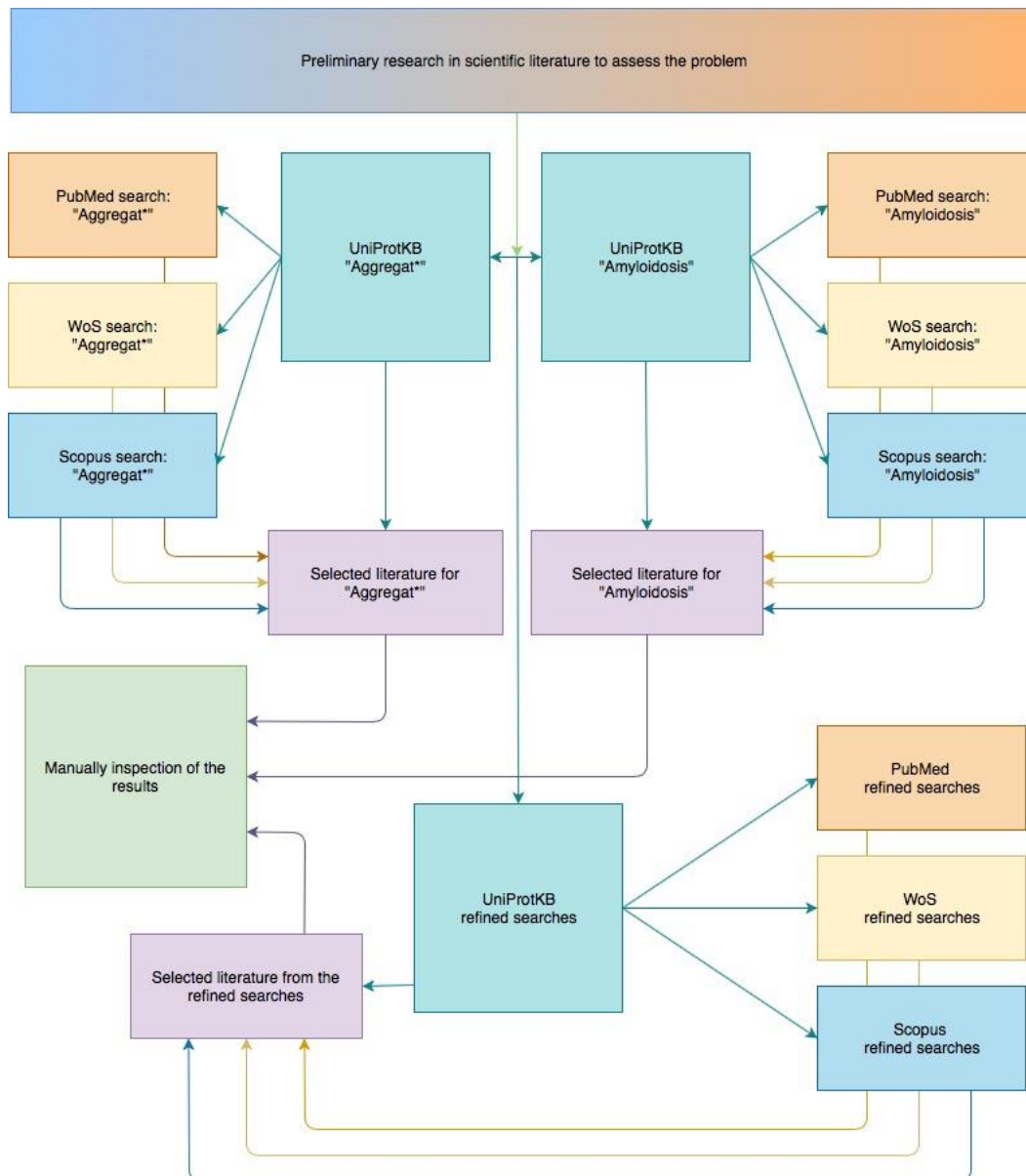


Figure 7: Schematic representation of the *aggregates* search.

A preliminary search in PubMed using the search string "protein toxicity" yielded a little more than 150 papers (full details are not reported in this report). Browsing the retrieved items it was noted that in scientific literature the term 'protein toxicity' is used not only to describe, in its proper sense, the noxious effects of exogenous proteins – toxin and TAS dealt with in the previous sections of this report – but also, arguably stretching the meaning of the term, to include the noxious effects of some endogenous proteins. Indeed, a few proteins – non-toxic *per se* but rather involved in some essential physiological function(s) in humans/animals – may become misfolded and/or aggregated under specific circumstances, with negative effects that appear to exceed the expected loss-of-function.

Details on the strategy followed are presented in Box 3.

Box 3

Question	Which proteins form in vivo aggregates connected with human or animal pathology?
Keywords	The keywords used for search in UniProt were amyloidosis and aggregat* (to include both aggregate and aggregation); the latter was used twice, first limiting the search to human proteins, then to non-human proteins.
Rationale	Scientific reports retrieved in literature databanks with the search term 'toxic protein' identify the insoluble deposits connected with human pathology, in most cases, with structured protein aggregates, formed in a process defined amyloidosis; in other cases, with unstructured aggregates.

Results: The 50 most cited papers for each query in the bibliographical databases were carefully evaluated to understand if protein aggregation could be considered as a toxic pathway and proteins involved in this pathological mechanism could be classified as toxins. Tables 16 and 17 summarize the results. As a result, we also found that the word "toxic" is improperly used in the literature to describe the effects of protein aggregates since none of these proteins have associated the "toxin" label in UniProtKB.

Table 16: Search terms statistics in bibliographic databases

<i>Search string</i>	<i>PubMed</i>	<i>Scopus</i>	<i>Web of Science</i>
"amyloidosis"	31,053	39,517	25,668
"protein aggregation" OR "protein aggregate"	83,328	40,198	11,312

Table 17: Results of aggregates-related searches

<i>Search</i>	<i>Findings</i>
Amyloidosis	In PubMed, the query « "amyloidosis" » retrieved little more than 30,000 papers, including little less than 5,000 reviews, defining it as the most specific and at the same time the furthest reaching search term to retrieve aggregation-prone molecules from the literature. Since the other two sources, i.e. Scopus and Web of Science, produced the same results, PubMed was selected as the reference source of information. We searched the UniProt database with the string: amyloidosis AND reviewed:yes, and retrieved 71 entries; we reviewed the list by inspecting the content of each entry. <i>Amyloidosis</i> is not associable to toxins.
Protein aggregates	To consider the situations in which proteins may assemble in a disordered/less-organised form than amyloid fibers, an additional search was conducted. The search string « "protein aggregation" OR "protein aggregate" » in PubMed yielded over 80,000 entries. The other two sources, i.e. Scopus and Web of Science, produced the same results. <i>Protein aggregates</i> are not associable to toxins.
Aggregation of human proteins	We thus set a search with main keyword 'aggregat*' to include both aggregate (noun and verb) and aggregation. As we did not want to duplicate findings, we included in the search string the words 'NOT amyloidosis' in order to disregard all proteins already identified by the 'aggregation search'; the difference in number, however, was negligible. As we meant to monitor first proteins relevant to human pathologies, and on the basis of the results of the previous search for amyloidosis, we initially restricted the search to our species by including in the search string the words AND organism:"Homo sapiens (Human) [9606]". Accordingly, the search string for our second interrogation of UniProt for endogenous proteins run as: « aggregat* NOT platelet NOT amyloidosis AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]" ». Same as for above, the results of the search were manually inspected. <i>Aggregation of human proteins</i> is not associable to toxins.

Aggregation of animal proteins	While the overall scenario was already clear, we repeated anyhow the search excluding human proteins by using the search string « aggregat* NOT platelet NOT amyloidosis NOT human AND reviewed:yes » in UniProt. We then individually inspected the over 3,500 entries to include only animals, ending up with a selection of 548 items. <i>Aggregation of animal proteins</i> is not associable to toxins.
---------------------------------------	---

2.3. Three-dimensional structures

The RCSB Protein Data Bank (PDB) was queried to retrieve all the related experimentally-determined structures to integrate the entries in both the Main and the TAS collections. The experimentally-determined structure selected for each protein was based on both primary structure coverage and experimental resolution (weighing these two terms 80% and 20%, respectively). All the selected structures were structurally-refined and energy-minimised using, respectively, the MOE software and Amber10:ETH as forcefield. Gaps in the secondary structures were filled through the 'loop modelling' procedure in which structure-related errors were also fixed, hydrogens were added while water and hetero-molecules were removed. By these steps, we refined the 3D structures retrieved from the PDB, generating an energy-minimized state for each structure.

The Swiss-model (SM) repository was queried to collect all the available 3D predicted models of proteins for which no experimentally-determined structure was available. Proteins found neither in PDB nor in the SM repository and with primary structure longer than 30 amino acids were modelled by us using the SM modelling tool; 30 amino acids is the lower limit for modelling in SM. For each SM-generated model, two different parameters (GMQE and QMEAN, Benkert et al., 2011) were carefully assessed in order to verify the quality of the structure. GMQE (Global Model Quality Estimation) is a quality estimation that combines properties from the target-template alignment and the template search method. The resulting GMQE score is expressed as a number between 0 and 1, reflecting the expected accuracy of the built model. The closer the GMQE score is to 1, the higher the reliability. On the other hand, QMEAN is a composite estimator based on different geometrical properties and provides both global (i.e. for the entire structure) and local (i.e. per residue) absolute quality estimates on the basis of one single model (Benkert et al., 2011).

Using CPH model, we also generated 3D models for peptides with primary structure between 20 and 30 amino acids, while peptides shorter than 20 amino acids were not modelled, since the accuracy is expected to be low due to the intrinsic disorder in the folding of very short sequences (Ohtake et al., 2013).

Consistently with the experimentally-determined 3D structures, all the SM and CPH protein/peptide models were also structure-prepared and energy-minimized using MOE software and Amber10:ETH as forcefield.

2.4. Families, domains and signatures

Redundancy among sources was investigated leading us to conclude that Pfam can be considered the main reference source for protein family classification, while InterPro can be considered the most complete database for both domains and signatures clustering. Although these two databases are interconnected, they provide the users with different classifications of the proteins, useful to perform different kinds of analysis or search. For this reason, the two databases cannot be considered equivalent. For each family in Pfam and domain/signature in Interpro, a brief description is reported in the Results section (Section 3), while the comprehensive results are listed in Annex E and Annex F, respectively. Moreover, a reference protein was selected, on the basis of the root mean square deviation (RMSD) of the positions of its alpha-carbons vs the protein family, and carefully checked, reporting in the Results section all the associated domains found in order to have a reference structure of the family, like a centroid of a cluster.

2.5. Automated information retrieval and management tool

A Python-based automated tool (script) was created in order to retrieve and manage all the information on the identified toxic proteins and associated literature. This tool queries the UniProtKB database and

interprets the .xml response by building an object-oriented Python-based database. From this object-oriented database many Excel-compatible tables are generated reporting and summarizing all the entries retrieved by the query. For each retrieved UniProtKB entry, a folder is created and filled with the sequence, the structure and other relevant information. For each entry the best structure is obtained first by trying to download the PDB structure, then, if none is available, by attempting to download the best SM model for the sequence; it is however also possible to add a user-made model. After these steps, the available structures are refined by automatically calling the external program MOE and by applying a standard protocol for processing. MOE is not included into the automated tool that is provided to EFSA, so the refinement option is not available.

3. Results

The searches conducted in this work according to the methodology described in Section 2 resulted in the definition of:

- two exhaustive collections of information on **toxic proteins**, based on UniProtKB database: the Main collection, containing information on proteins toxic *per se*, and the TAS collection, informing on proteins belonging to toxin-antitoxin systems (TAS). All the information retrieved on the identified toxic proteins and associated literature are managed by an *ad hoc* Python-based software. The results are presented in detail in Section 3.1;
- a compilation of information on **protein aggregates** associated with pathological conditions, with results discussed separately in Section 3.2.

The *in silico* pipeline developed to support the search is summarised in Figure 8. Furthermore, the results of the quality assurance of 3D structures of toxic proteins and of the testing of the main freely available on-line tools for predicting protein toxicity are presented in Section 4 and 5 respectively.

This report provides information on all toxic proteins currently known based on the available information and the state-of-the-art of scientific research in this field. It cannot be excluded that some additional toxic proteins not yet characterised as toxins do exist.

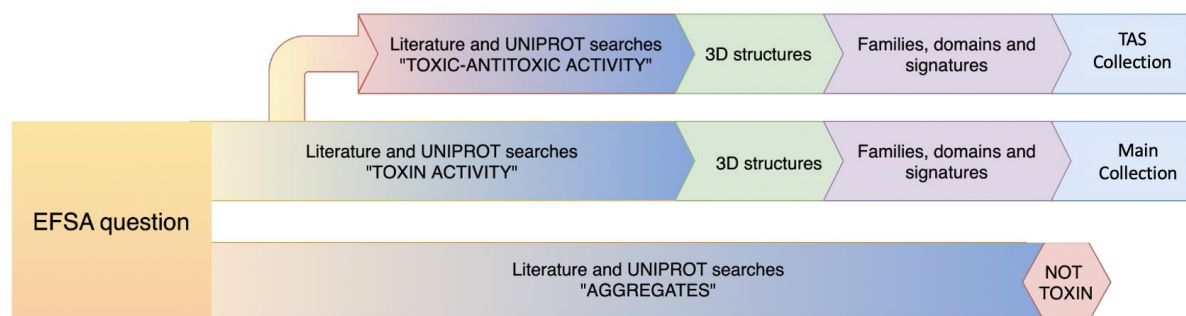


Figure 8: *In silico* pipeline devised to fulfil the EFSA request.

3.1. Toxic proteins

This report is structured in order to provide information on 1) the number of proteins fitting the definition of proteins causing toxic effects *per se* or belonging to TAS, respectively; 2) the primary structure length of these proteins (expressed as distributions); 3) the organisms that produce them, 4) the associated GO terms (Molecular Function, Cellular Component and Biological Process) and 5) keywords.

The proposed structure has been chosen since matching the gold standard database UniProtKB (and related databases). This can facilitate possible further bioinformatics analysis in this area. Further detailed information on the toxicity of each identified protein is reported in Annexes A-D, as detailed in the sections below.

3.1.1. Main Collection

As detailed in Section 2.2.1, proteins causing toxic effects *per se* are identified in UniProtKB by the keyword "toxin activity". Full details are reported in the Main Collection related annexes (see Table 15).

3.1.1.1. Number of proteins identified

Data on 6,964 proteins with an associated well-recognised toxic activity (i.e. causing toxic effects *per se*, Keyword "Toxin activity and Reviewed:Yes") were downloaded from UniProtKB (March 2020). Among them, 4,225 have a precursor sequence (i.e. a sequence containing a peptide that inhibits the toxic function until it is removed), while 1,157 are labelled as fragments (the whole primary structure is still unknown). Full details are included in Excel files (see Table 15); an example is presented in Table 18.

Table 18: Example of primary structures table (see MC_sequences.xlsx)

Uniprot	Fragment	Precursor	Structure	Length	Experimental determination of the structure	Best pdb
P62377			Swiss Model	62	0	
P25681			Swiss Model	61	0	
P0CG02			Swiss Model	60	0	
Q98962		true	Swiss Model	81	0	
Q9W716		true	Swiss Model	83	0	
Q53B57		true	Swiss Model	91	0	
B2KKV7	single	true	Swiss Model	302	0	
P01452			PDB	60	100%	1cdt
...

Fragment: is used to inform that the protein is only a fragment of the primary sequence

Precursor: informs if the toxin has a precursor

Structure: source of download for the selected structure

Length: number of amino acids in the primary structure

Experimental determination of the structure: informs on the percentage of the primary structure coverage of the experimentally-derived 3D model (with respect to the entire length of primary structure)

Best pdb: the protein data bank code of the best resolved 3D structure

3.1.1.2. Sequence length

As shown in Figures 9 and 10, ~ 90% of the proteins are shorter than 300 amino acids whereas ~65% are shorter than 100 amino acids. Only 465 proteins (~7%) are shorter than 20 amino acids and were not modelled.

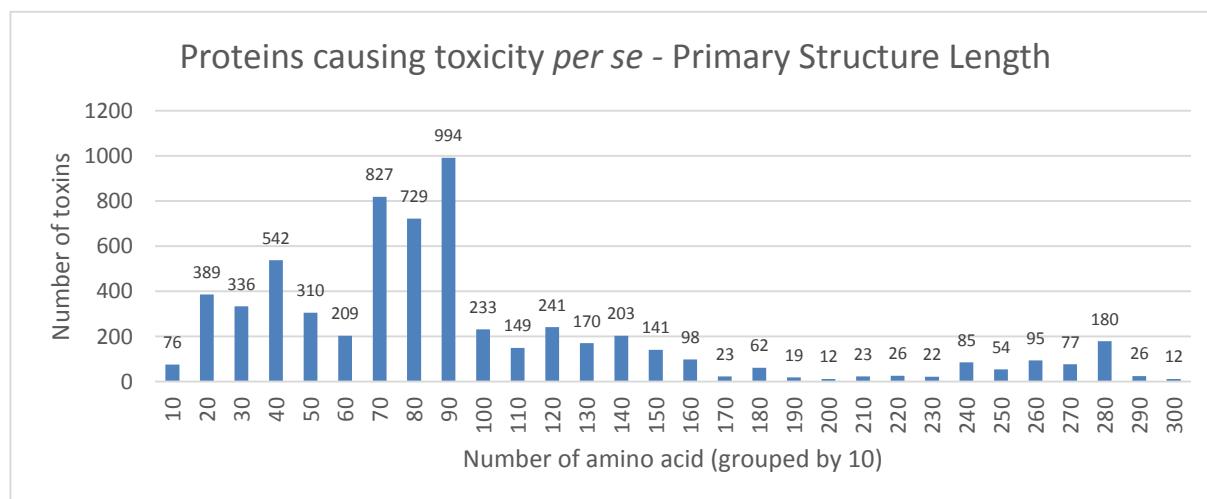


Figure 9: Histogram of the primary structure length for proteins causing toxicity *per se* - Main Collection (up to 300 amino acids in length).

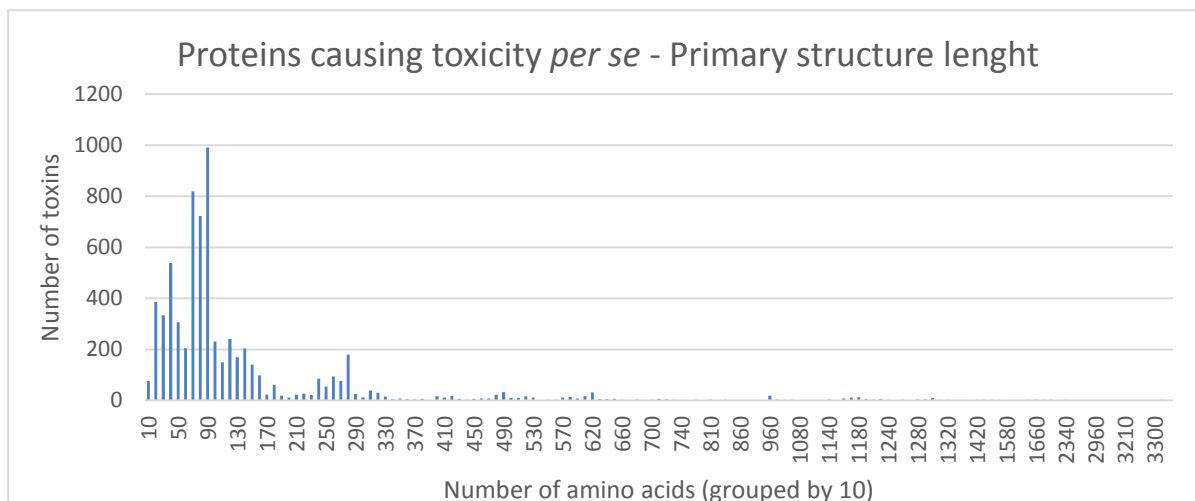


Figure 10: Histogram of the primary structure length proteins causing toxicity *per se*.

3.1.1.3. Organisms

Reviewing the information on the organisms that express the selected toxins, we found that the most representative ones (i.e. those expressing the highest number of toxins) belong to the animal kingdom. The most recurrent organism is *Cyriopagopus hainanus* (292 proteins), followed by *Lycosa singoriensis* (217 proteins), *Californiconus californicus* (118 proteins), *Chilobrachys guangxiensis* (104 proteins) and *Conus textile* (100 proteins). In the following lines, a brief description is reported for the five most representative organisms:

- i) *Cyriopagopus hainanus* is a spider species from the family Theraphosidae (tarantulas), found in China; it is also known as the "Chinese bird spider". Its venom contains different proteins (e.g. hainantoxins) able to inhibit tetrodotoxin-sensitive voltage-gated sodium channels, causing blockage of neuromuscular transmission;
- ii) *Lycosa singoriensis* is a hunting spider with a widespread distribution in Northwest China. The venom gland of these spiders can secrete abundant and complex neurotoxins;
- iii) *Californiconus californicus*, common name the Californian cone, is a species of small, predatory sea snail, a marine gastropod mollusc in the family Conidae, the cone snails, secreting a conotoxin;
- iv) *Chilobrachys guangxiensis* is a species of Araneae in the family Theraphosidae, found in China; it is also known as the "Chinese fawn tarantula"; it secretes proteins inhibiting voltage-gated potassium channels;
- v) *Conus textile Linnaeus* is a marine gastropod mollusc of the Conidae family, widespread in the Red Sea and in the coral reefs of South-east Asia and Australia. Its poison, containing neurotoxins belonging to the family of the counterfeiters, is dangerous and often even deadly for humans.

The five most representative organisms from Bacteria, Plantae and Fungi kingdoms are also highlighted:

Bacteria:

- i) *Bacillus thuringiensis* is a Gram-positive, soil-dwelling bacterium, commonly used as a pesticide. It also occurs naturally in the gut of caterpillars of various types of moths and butterflies, as well on aquatic environments, leaf surfaces, insect-rich environments, animal faeces and grain-storage facilities. During sporulation, it produces crystal proteins (Cry proteins), among which there are δ -endotoxins that promote colloid-osmotic lysis by binding to the midgut epithelial cells of many lepidopteran larvae;
- ii) *Escherichia coli* is the most widely studied prokaryotic model organism, and an important species in the fields of biotechnology and microbiology, where it has served as the host organism for the majority of work with recombinant DNA. It expresses hemolysin, which has some hemolytic activity towards mammalian cells;

- iii) *Mannheimia haemolytica* is an opportunistic bacterium able to gain access to the lungs when the host defences are compromised by infection due to respiratory viruses or mycoplasma, such as parainfluenza, IBR or environmental stress. It expresses leukotoxins, exotoxins that attack host leukocytes and especially polymorphonuclear cells, causing cell rupture;
- iv) *Staphylococcus aureus* is a facultative anaerobe, Gram-positive bacterium and a usual member of the microbiota of the body: it can be found on human skin, nose, armpit, groin, and other areas. Staphylococcal enterotoxins cause the intoxication staphylococcal food poisoning syndrome. In particular *S. aureus* expresses beta-hemolysin, a phospholipase C with specific activity toward sphingomyelins;
- v) *Streptococcus pyogenes* is a Gram-positive and catalase-negative bacterium that causes several diseases in human beings such as pharyngitis, skin and throat infections and acute rheumatic fever. It expresses streptolysin, a sulfhydryl-activated toxin that causes cytolysis by forming pores in cholesterol-containing host membranes. It is commonly used in bionanotechnology and genome editing, such as CRISPR.

Plantae:

- i) *Abrus precatorius*, commonly known as jequirity bean or rosary pea, is a weedy and invasive plant typical of temperate and tropical regions, such as Asia and Australia, which contains one of the most lethal toxins, abrin, a toxalbumin that inhibits protein synthesis, causing cell death;
- ii) *Arabidopsis thaliana* is a small, diploid and short-lived flowering plant native to Eurasia and Africa, member of the mustard family (Brassicaceae), including cultivated species such as cabbage and radish, which is a powerful model for the study of growth and development processes in plants, due to its relatively small genome of approximately 135 megabase pairs and 5 chromosomes. *A. thaliana* expresses thionins, small plant proteins that are toxic to animal cells. They seem to exert their toxic effect at the level of the cell membrane;
- iii) *Hordeum vulgare* (barley) is a diploid photoautotroph herbaceous species of plant of the family of True grasses, with 14 chromosomes and a self-supporting growth habit. It's a major cereal grain with simple, broad leaves and yellow flowers. Barley has been largely used as animal fodder, as a source of fermentable material for beer and certain distilled beverages, and as a component of various foods. It expresses thionins that are small plant proteins toxic to animal cells. They seem to exert their toxic effect at the level of the cell membrane.

Fungi:

- i) *Amanita bisporigera* is a deadly poisonous fungus from the family of the Amanitaceae. The mushroom has a smooth white cap, which can reach up to 10 cm in diameter, and a stipe, up to 14 cm long. It contains amatoxins, cyclic peptides that inhibit the enzyme RNA polymerase II and interfere with various cellular functions. Symptoms of poisoning appear 6 to 24 hours after consumption, followed by a period of apparent improvement, and includes liver and kidney failure, which lead to death after four days or more. The DNA of *A. bisporigera* has been partially sequenced, and the genes responsible for the production of amatoxins have been identified;
- ii) *Amanita exitialis* is a deadly poisonous mushroom belonging to the large family of Amanitae. The highest content of toxic peptides are found in the cap and these peptides have been widely used in biological research as chemical agents to inhibit RNA polymerase II, an enzyme essential for protein synthesis;
- iii) *Amanita fuliginea* is a lethal poisonous mushroom responsible for most of the fatal mushroom poisoning incidents in Southern China. The toxic activity is due to several toxic peptides: α -amanitin, β -amanitin, amaninamide, phallocin, phallacidin, phallisacin, desoxoviroidin, and an additional uncharacterized phallotoxin;
- iv) *Amanita phalloides* is one of the most poisonous mushrooms known. It has been linked in the majority of human deaths from mushroom poisoning: half a mushroom contains enough

toxin to kill an adult human. Also known as the death cap, it is distributed worldwide, especially across Europe. The class of toxins found in these mushrooms, amatoxins, can cause hepatic and renal failure and are thermostable, so their toxic effects are not reduced by cooking;

- v) *Amanita rimosa* grows in a broad-leaved forest dominated by Fagaceae and this mushroom is only known from the type locality (China, Hunan Province, Yizhang County, Mangshan) at present. A remarkable feature of *Amanita rimosa* is its splitting cap surface, which is a result of the mushroom's having a slightly gelatinized upper layer of the cap's skin that includes abundant inflated cells. The major toxins of *A. rimosa* belong to the bicyclic octapeptides amatoxins, which act by binding non-competitively to RNA polymerase II, greatly slowing the elongation of transcripts from target promoters.

Figure 11 histogram summarises the information about the number of proteins in the Main Collection on the basis of the organism producing them, while Table 19 exemplifies data summarised in the dedicated Organisms.xlsx file.

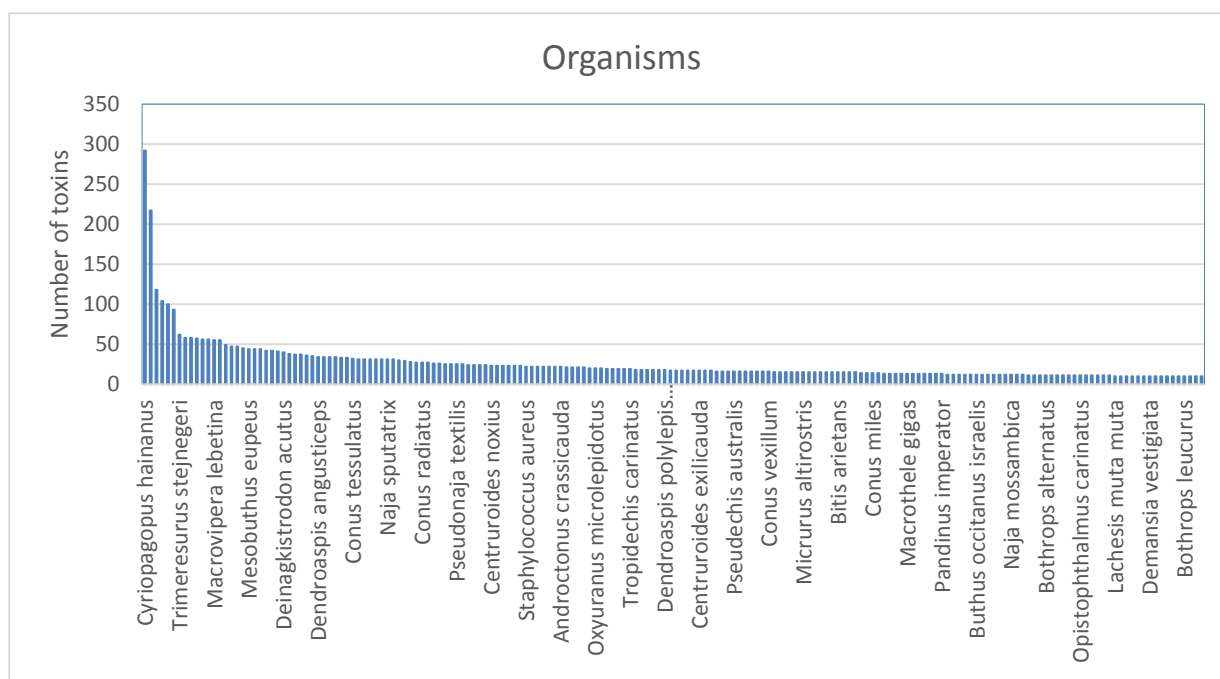


Figure 11: Histogram of organism source of the proteins toxic *per se*

Table 19: Extract of organisms table (MC_organisms.xlsx)

UniProt	Organism
A0A1Q7CZY2	Actinobacteria bacterium 13_2_20CM_2_66_6
A0A2S1FHS7	Polaromonas sp. E10S
A0A351WVN6	Geobacter sp.
A0A1F9GT57	Deltaproteobacteria bacterium RIFCSPHIGHO2_12_FULL_43_9
A0A1M3BB20	Solirubrobacterales bacterium 70-9
A0A3X1TZV0	Salmonella enterica
...	...

UniProt: the UniProtKB identifier of the protein/peptide; Organism: organism that express the toxin.

3.1.1.4. Gene ontology (GO) terms

The GO terms (see Section 2.1.2) cover three domains: biological processes, molecular functions and cellular components. This information is important to understand the mechanism of action of the toxins and associated pathogenesis. GOs for the identified proteins toxic *per se* (Main Collection) were

analysed according to these three categories (see Figures 12-14 reporting the frequencies of the individual GO in the Main collection of proteins). More than one GO can be associated with each toxin, while 18 retrieved proteins were found not to be associated with this GO term “toxin activity”. After careful inspection, it was observed that all these proteins are flaggable as toxins that belong to a toxin-antitoxin system. Table 20 exemplifies data summarised in dedicated go.xlsx files.

Table 20: Extract of GO terms table (MC_go.xlsx)

UniProt	ID	GO
A0A1Q7CZY2	Go:0000287	Magnesium ion binding
A0A1Q7CZY2	Go:0004540	Ribonuclease activity
A0A1Q7CZY2	Go:0090729	Toxin activity
A0A2S1FHS7	Go:0000287	Magnesium ion binding
A0A2S1FHS7	Go:0004540	Ribonuclease activity
...

UniProt: the UniProtKB identifier of the protein/peptide; ID: the Gene Ontology unique identifier; GO: the name of the GO.

In the following sections, the most recurrent GOs for the identified proteins are reported:

Molecular Function (Figure 12):

- i) *toxin activity: interacts 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;*
- ii) *ion channel inhibitor activity: stops, prevents, or reduces the activity of an ion channel;*
- iii) *metal ion binding: interacts selectively and non-covalently with any metal ion;*
- iv) *sodium channel inhibitor activity: stops, prevents, or reduces the activity of a sodium channel;*
- v) *calcium ion binding: interacts selectively and non-covalently with calcium ions (Ca²⁺).*

Except for toxin activity, it is interesting to point out that the most frequently found molecular functions of toxic proteins are related to channel inhibition activity and to ion binding.

Biological process (Figure 13):

- i) *defence response: reactions, triggered in response to the presence of a foreign body or the occurrence of an injury, which result in restriction of damage to the organism attacked or prevention/recovery from the infection caused by the attack;*
- ii) *hemolysis in other organism: the cytolytic destruction of red blood cells, with the release of intracellular hemoglobin, in one organism by another;*
- iii) *lipid catabolic process: the chemical reactions and pathways resulting in the breakdown of lipids, compounds soluble in an organic solvent but not, or sparingly, in an aqueous solvent;*
- iv) *arachidonic acid secretion: the controlled release of arachidonic acid from a cell or a tissue;*
- v) *phospholipid metabolic process: the chemical reactions and pathways involving phospholipids, any lipid containing phosphoric acid as a mono- or diester.*

Cellular component (Figure 14):

- i) *extracellular region: the space external to the outermost structure of a cell. For cells without external protective or external encapsulating structures this refers to space outside of the plasma membrane. This term covers the host cell environment outside an intracellular parasite.*

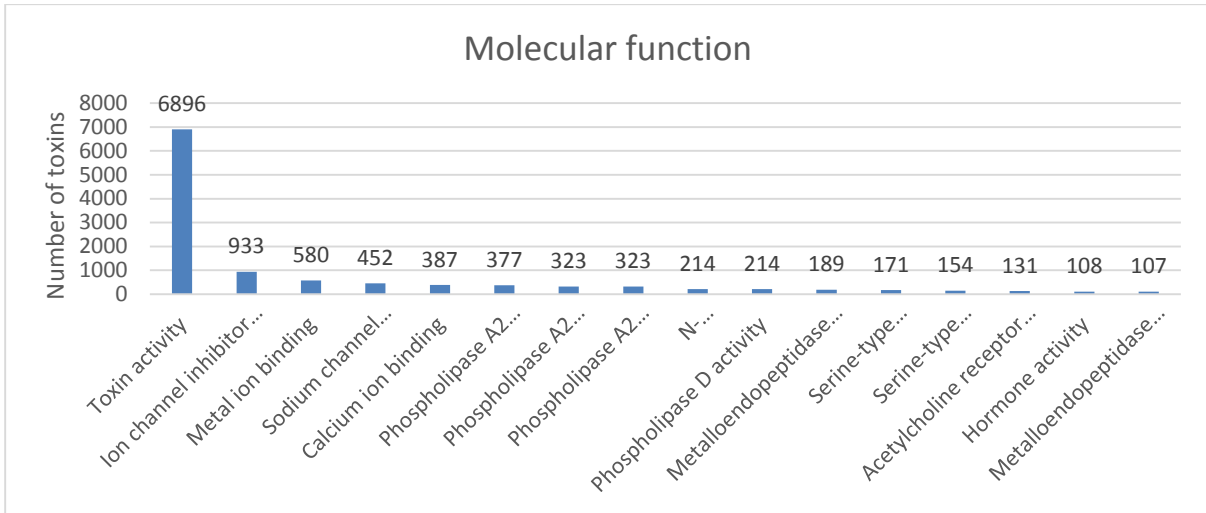


Figure 12: Frequency of the Molecular Function GO terms

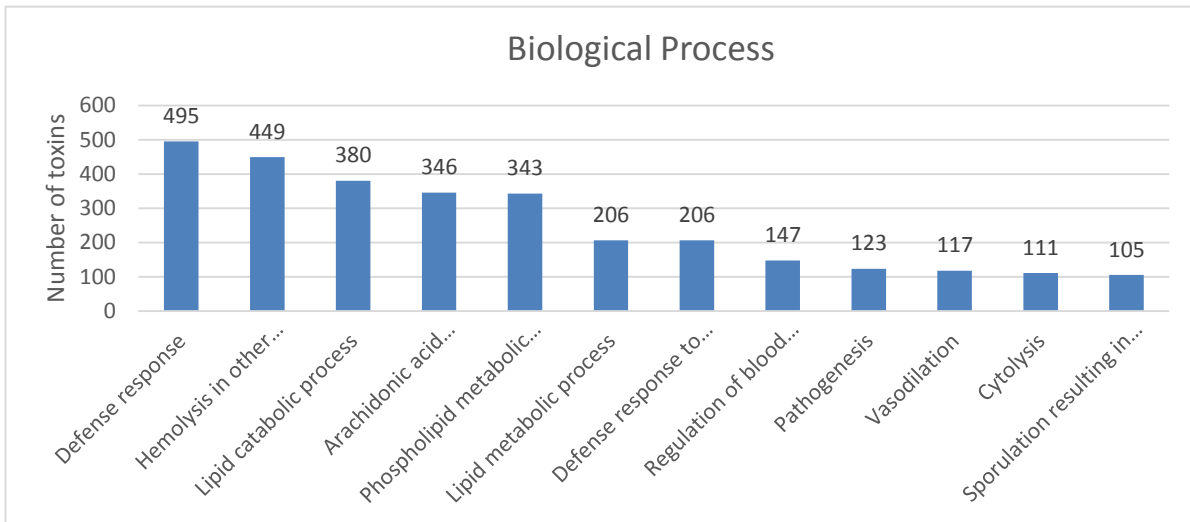


Figure 13: Frequency of the Biological Process GO terms

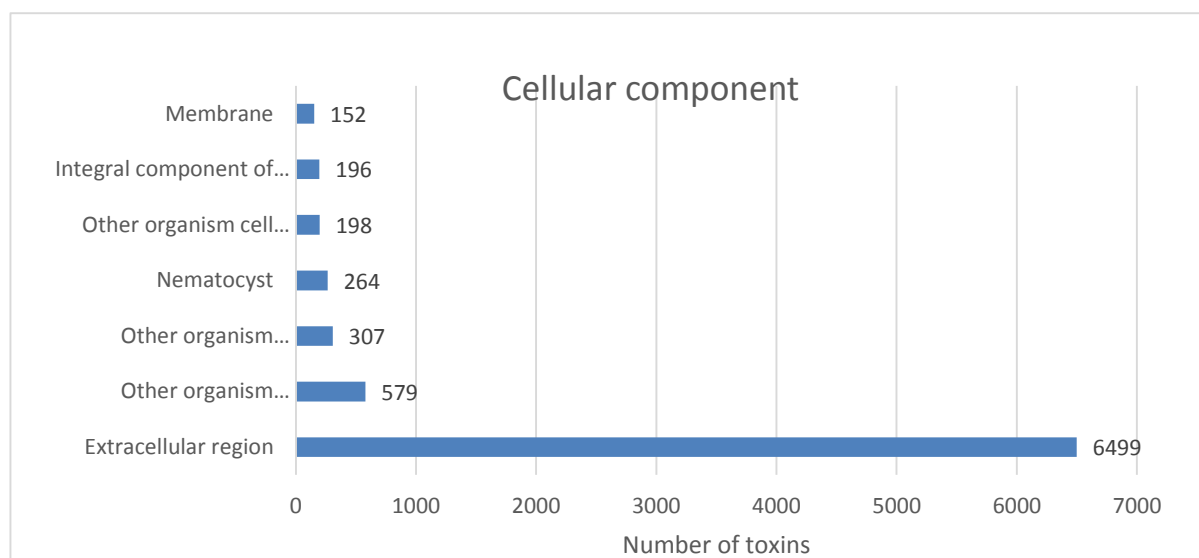


Figure 14: Frequency of the Cellular Component GO terms

Defence response is the most frequently found biological process while Extracellular region is the most frequently found cellular component, in compliance with the GO-based definition of toxin.

3.1.1.5. Keywords

For each protein from the Main Collection, keywords were extracted from UniProtKB (unique ID and category) (see Table 21). UniProtKB keywords constitute a controlled vocabulary with a hierarchical structure, different from GO. Keywords are native of UniprotKB and summarize the content of a UniProtKB entry, facilitating the search for proteins of interest. Keywords were assigned by UniProtKB curators and are classified in 10 categories:

- Biological process
- Cellular component
- Coding sequence diversity
- Developmental stage
- Disease
- Domain
- Ligand
- Molecular function
- Post-translational modification (PTM)
- Technical term

Table 21: Extract of keywords table (MC_keywords.xlsx)

Uniprot	ID	Keyword	Category
A0A1Q7CZY2	KW-0378	Hydrolase	Molecular function
A0A1Q7CZY2	KW-0460	Magnesium	Ligand
A0A1Q7CZY2	KW-0479	Metal-binding	Ligand
A0A1Q7CZY2	KW-0540	Nuclease	Molecular function
A0A1Q7CZY2	KW-0800	Toxin	Molecular function
A0A1Q7CZY2	KW-1277	Toxin-antitoxin system	Biological process
A0A2S1FHS7	KW-0378	Hydrolase	Molecular function
A0A2S1FHS7	KW-0460	Magnesium	Ligand
...

UniProt: the UniProtKB identifier of the protein/peptide; ID: the keyword unique identifier; Keyword: the keyword; Category: the category in which the keyword is categorized.

An entry often contains several keywords. Within a category, the keywords are stored in alphabetical order. Some keywords are also related to GO. Figure 15 reports the count of all the keywords, not divided into categories. The following sections describe in more detail the main keywords, according to their frequency:

- i) Toxin
Definition: Naturally-produced poisonous protein that damages or kills other cells, or the producing cells themselves in some cases in bacteria. Toxins are produced by venomous and poisonous animals, some plants, some fungi and some pathogenic bacteria. Animal toxins (mostly from snakes, scorpions, spiders, sea anemones and cone snails) are generally secreted in the venom of the animal.
Category: Molecular function
GO: toxin activity [GO:0090729]
- ii) Secreted
Definition: Protein secreted into the cell surroundings.
Category: Cellular component
GO: extracellular region [GO:0005576]
- iii) Disulfide bond
Definition: Protein, which is modified by the formation of a bond between the thiol groups of two peptidyl-cysteine residues. The process of chemical oxidation that forms interchain disulfide bonds can produce stable, covalently linked protein dimers, multimers or complexes, whereas intrachain disulfide bonds can contribute to protein folding and stability. Depending on the protein environment, some disulfide bonds are more labile, forming transient redox-active disulfide bonds that are alternately reduced and oxidized in the course of an enzymatic reaction.
Category: PTM
- iv) Signal
Definition: Protein, which has a signal sequence, a peptide usually present at the N-terminus of proteins and which is destined to be either secreted or part of membrane components. The signal sequence (usually 20-30 amino acids long) interacts with the signal recognition particle and directs the ribosome to the endoplasmic reticulum where co-translational insertion takes place. Signal peptides are highly hydrophobic but have some positively charged amino acids. Normally, the signal sequence is removed from the growing peptide chain by specific peptidases (signal peptidases) located on the cisternal face of the endoplasmic reticulum.
Category: Domain
- v) Direct protein sequencing
Definition: Protein, whose amino acid sequence has been partially (more than one residue) or completely determined experimentally by Edman degradation or by mass spectrometry.
Category: Technical term
- vi) Neurotoxin
Definition: Proteins, often exquisitely toxic, that inhibit neuronal function. Neurotoxins act typically against sodium channels or block or enhance synaptic transmission. Most venoms contain neurotoxic substances.
Category: Molecular function
- vii) Ion channel impairing toxin
Definition: Protein which interferes with the function of ion channels, which are hydrophilic channels across the lipid bilayer through which specific inorganic ions can diffuse down their electrochemical gradients.
Category: Molecular function
GO: ion channel inhibitor activity [GO:0008200]
- viii) Knottin
Definition: Small disulfide-rich protein characterized by a special 'disulfide through disulfide knot'. This knot is obtained when one disulfide bridge crosses the macrocycle formed by two other disulfides and the interconnecting backbone (disulfide III-VI goes through disulfides I-IV and II-V). The knottin structure is found in some plant protease inhibitors, cyclotides, toxins

from cone snails, spiders, insects, horseshoe crabs and scorpions, gurmamin-like peptides, agouti-related proteins, and some antimicrobial peptides.

Category: Domain

ix) Hydrolase

Definition: enzyme which catalyzes hydrolysis reaction, i.e. the addition of the hydrogen and hydroxyl ions of water to a molecule with its consequent splitting into two or more simpler molecules.

Category: Molecular function

GO: hydrolase activity [GO:0016787]

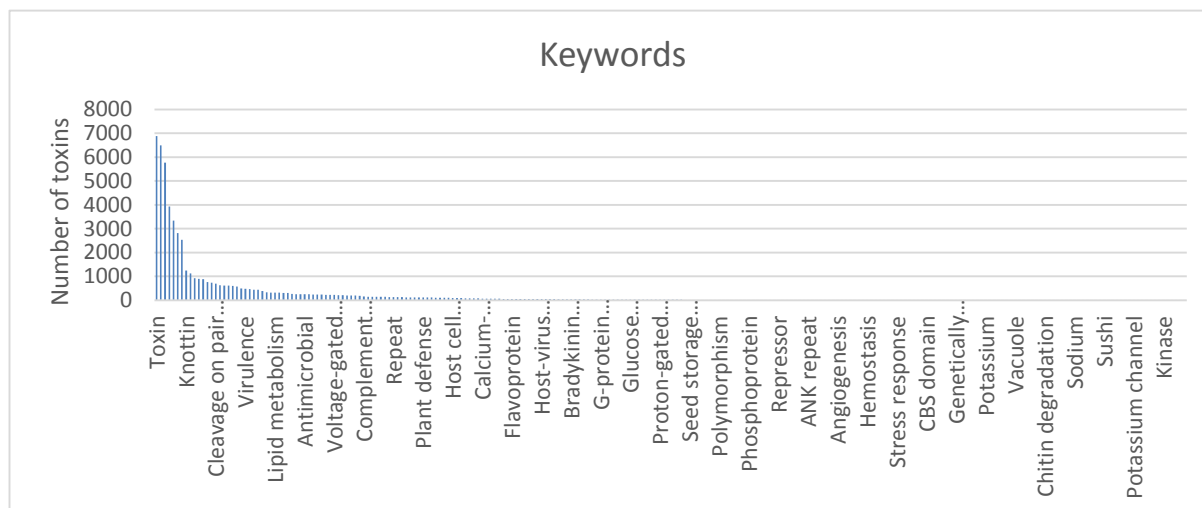


Figure 15: Frequency of keywords

Full information on Keywords and their count can be found in the dedicated Excel files (MC_sequences.xlsx file).

3.1.1.6. Three-dimensional structures

Out of 6,964 proteins identified in this search, 765 have associated one or more experimentally-solved 3D structure/s. As described in the Materials and methods section, 3D structures were carefully checked and only one structure was selected for each protein. Information about experimental determination of the structure such as method, resolution and percentage of sequence coverage were collected from the PDB source and are reported in a separate file (see Folder #1). Table 22 exemplifies the PDB table.

Table 22: Extract of PDB table (see MC_pdb.xlsx)

UniProt:	PDB	Method	Resolution [Å]	Experimental determination of the structure
P01452	1cdt	X-ray	2.5	100%
P0DM15	2n7f	NMR		100%
P11494	2bds	NMR		100%
P10565	5foy	X-ray	2.25	100%
P10565	5foz	X-ray	2.4	100%
P10565	5g37	X-ray	2.5	100%
P07231	1ad7	NMR		17%
...

UniProt: the UniProtKB identifier of the protein/peptide; PDB: the RCSB PDB identifier of the selected structure; Method: 3D structure determination method; Resolution: experimental resolution of the selected 3D structure; Experimental determination of the structure: percentage of primary structure covered by the 3D experimentally determined model.

Figure 16 briefly reports the number of experimentally-solved structures for each protein (labelled with its UniProtKB code), while Table 21 reports the number of available 3D structures in the PDB according to the structure determination methodology used. 'Model' belongs to an obsolete RCSB pdb policy under which also some theoretically-built models were available on the PDB source. This type of entry is now deprecated. As shown in Table 23, there are some toxins with associated two or more experimentally-solved structures.

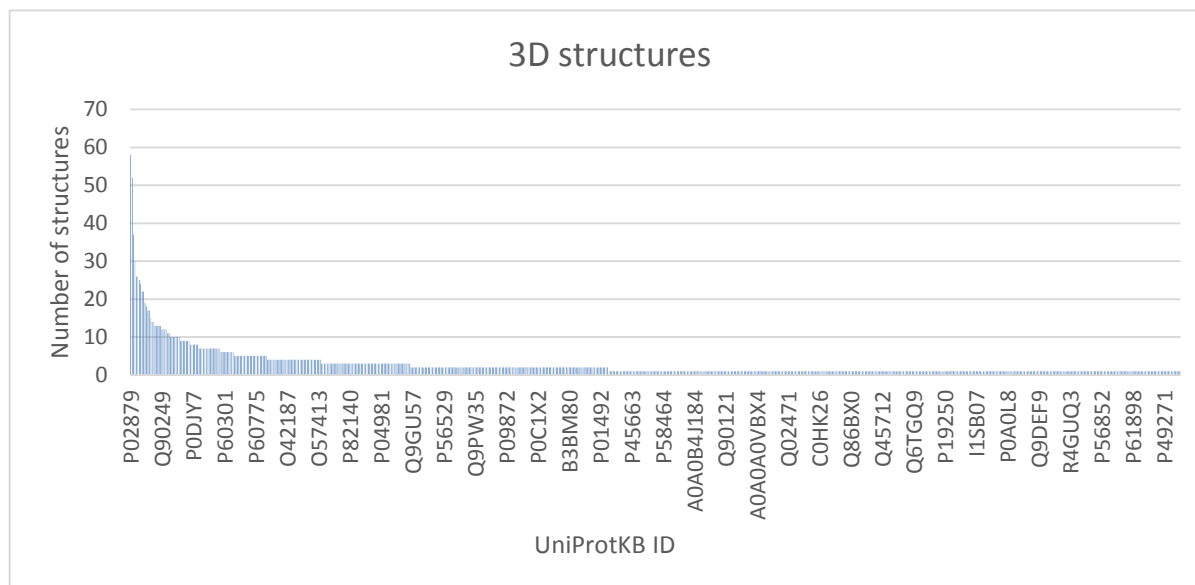


Figure 16: Statistics of experimentally-solved structures per protein

Table 23: Total number of experimentally-solved structures for protein toxins by method

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

Moreover, a total of 5,298 models were downloaded from the SM repository and stored into the Collection. There are some toxins with two or more associated models in the SM repository. As stated in the Materials and Methods section, in these cases we selected the model with the best quality check parameters for our *in silico* analysis. Information about the template, identity and percentage of sequence coverage are shown in summarizing Table 24 which also includes quality parameters.

Table 24: Extract of SM table (See MC_swiss_models.xlsx)

UniProt	Template	Identity	Oligo	Coverage	Qmean	Qmean norm	Gmqe	Origin
P62377	1cvo.1.A	100%	monomer	100%	-6.9	0.3	0.99	SM
P25681	1b41.1.B	93.4%	monomer	100%	0.05	0.8	0.98	SM
P0CG02	2mj4.1.A	70%	monomer	100%	-0.7	0.7	0.93	SM
Q98962	1cb9.1.A	86.6%	monomer	74%	-1.1	0.7	0.69	SM
Q9W716	1kxi.1.B	98.3%	monomer	75%	-0.2	0.8	0.75	SM
Q53B57	1txb.1.A	85.0%	monomer	74%	-5.5	0.4	0.7	SM
P62377	1cvo.1.A	100%	monomer	100%	-6.9	0.3	0.99	SM
P25681	1b41.1.B	93.4%	monomer	100%	0.05	0.8	0.99	SM
...

UniProt: the UniProtKB identifier of the protein/peptide; Template: the PDB code of the 3D protein structure used for homology modelling; Identity: percentage of identity between template and query; Oligo: ; Coverage: percentage normalized of the modelled structure coverage; Qmean(Benkert et al., 2011): parameter involved in the quality estimation (please, refer to

Methods); Qmean norm: parameter involved in the quality estimation; Gmqe: parameter involved in the quality estimation (please, refer to Methods); Origin: tool used for the homology modelling (SM = Swiss-Model; CPH = CPH Model).

As shown in Figure 17, some templates were used to model different proteins (e.g. PDB ID: 2MDN) that belong to the same family.

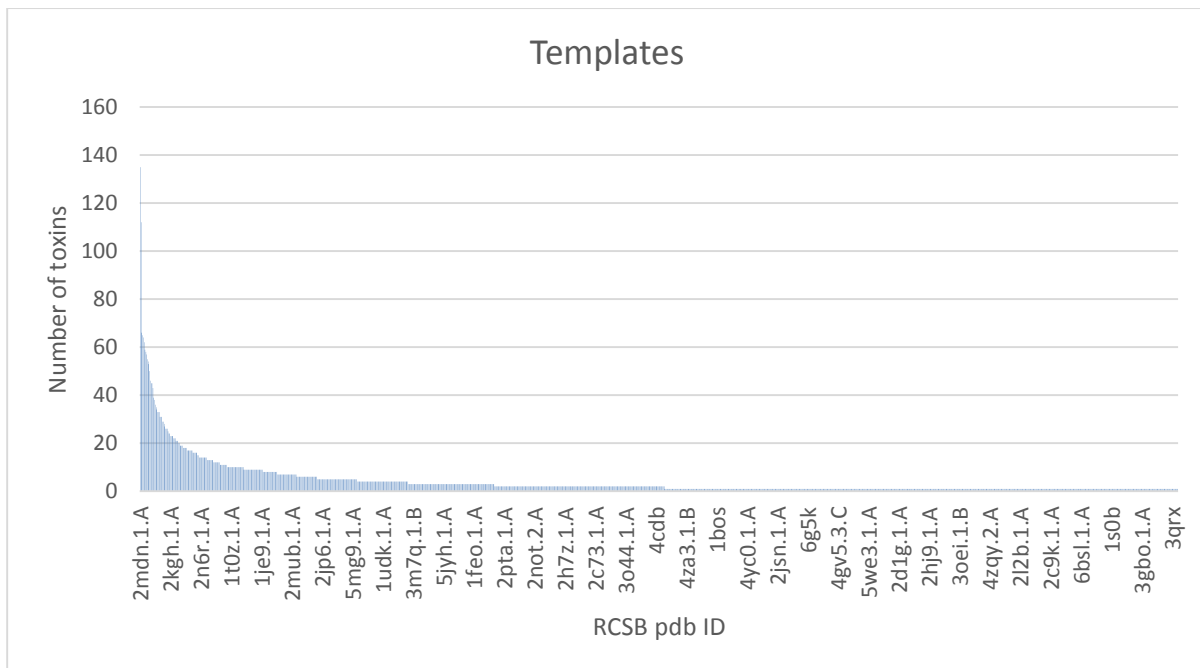


Figure 17: Statistics of crystallographic structure templates per protein

3.1.1.7. Examples of entries in the Main Collection

For each protein, biologically-relevant information was downloaded from UniProtKB. Two examples are reported below to distinguish between a well-described protein (MEL_APIME) and a poorly described one (DIS_CERVI). The examples are directly taken from the Main Collection (Annex A). The experimental or modelling-related information is listed in the respective Excel file ("MC_pdb.xlsx", see Table 15).

P01501 (MEL_APIME)

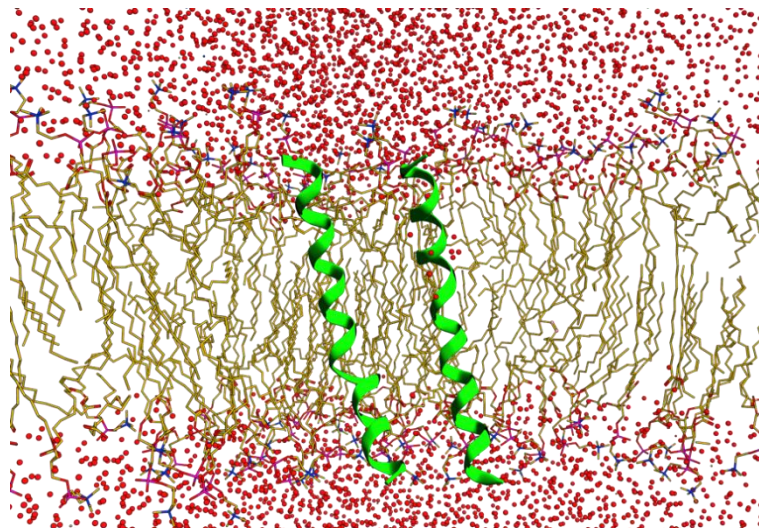


Figure 18: Melittin (green) forms pores by inserting itself into the membrane lipid bilayer (yellow). (PDB ID: 2MLT; membrane, water and ion were added with MOE only for a graphical representation)

Function

Melittin (Figure 18): main toxin of bee venom with strong hemolytic and antimicrobial activity. It enhances the effects of bee venom phospholipase A2. This amphipathic toxin binds to negatively charged membrane surfaces and forms pores by inserting into lipid bilayers. This in turn induces the leakage of ions and molecules from the cells and increases its permeability, ultimately leading to cell lysis. It acts as a voltage-gated pore with higher selectivity for anions over cations. The ion conductance has been shown to be voltage-dependent. Self-association of melittin in membranes is promoted by high ionic strength, but not by the presence of negatively charged lipids. In vivo, intradermal injection into healthy human volunteers produces sharp pain sensation and an inflammatory response. It produces pain by activating primary nociceptor cells directly and indirectly due to its ability to activate plasma membrane phospholipase A2 and its pore-forming activity.

Melittin-S: 1.4-fold less hemolytic and adopts a less organized secondary structure than melittin.

Melittin-2: Has strong hemolytic activity.

Miscellaneous

N-formyl-melittin major has 80% of the activity of melittin.

Melittin: The secretion of this protein into venom follows a seasonal pattern. This variation is synchronized with phospholipase A2 variation, i.e. their production increases in the same months.

Melittin-S: The secretion of this protein into venom follows a seasonal pattern, the maximum secretion occurring during the (southern) winter months.

Exists in two forms, due to cis-trans isomerization at 56-Leu-Pro-57. The trans conformation is the major form. The trans conformation is required for an alpha-helix. The derivative Melp5 is amidated.

Pharmaceutical

Melittin is an attractive candidate for cancer chemotherapy causing more damage to the tumor cell membranes since its membrane potential is higher and cells are less likely to develop resistance to a membrane pore formation. Despite this potential applicability of melittin, its rapid degradation in the blood and its non-specific cellular lytic activity - including hemolysis - poses significant challenges. However, melittin and/or its conjugates can work in conjunction with hormone receptors, gene therapy or as nanoparticles for targeted therapies of some cancer types.

Similarity

Belongs to the melittin family.

Subcellular location

Alpha-helical peptides form toroidal pores in the prey.

Subunit

Monomer (in solution and for integration into membranes), homotetramer (in solution and potentially as a toroidal pore in membranes), and potentially homomultimer (as a toroidal pore in membranes).

Tissue specificity

Expressed by the venom gland.

Toxic dose

LC(50) is 2.7 ug/ml against killifish.

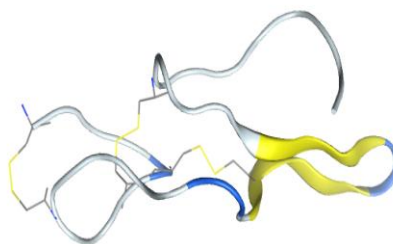
Q3BK17 (DIS_CERVI)

Figure 19: Structure of Q3BK17 obtained via homology modelling. This protein shows a Knottin architecture.

Function

Specifically interacts with the alpha-1/beta-1 integrin (ITGA1/ITGB1). Exhibits highly inhibitory effects on cell adhesion and cell migration to collagens I and IV. Also shows in vivo anti-angiogenic activity (By similarity) (Figure 19).

Similarity

Belongs to the disintegrin family. Short disintegrin subfamily.

Subunit

Monomer.

Tissue specificity

Expressed by the venom gland.

3.1.1.8. Families, domains and signatures

As described in Section 2.4, for each source, families/domains classification was carefully checked (Table 25); redundancy among available sources was investigated leading us to conclude that Pfam can be considered the main reference source for protein family classification, while InterPro can be considered the most complete database for both domains and signatures clustering. The domain/signature classification in all the other databases mirrors that of Interpro, which is also the most up-to-date database. All the information extracted from both Pfam and InterPro is reported in Annex E and F, respectively, while Figures 20 and 21 show the number of proteins for each family or domain, respectively. Signatures are considered in the domain classification.

Table 25: Number of clusters by database (See MC_databases.xlsx)

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

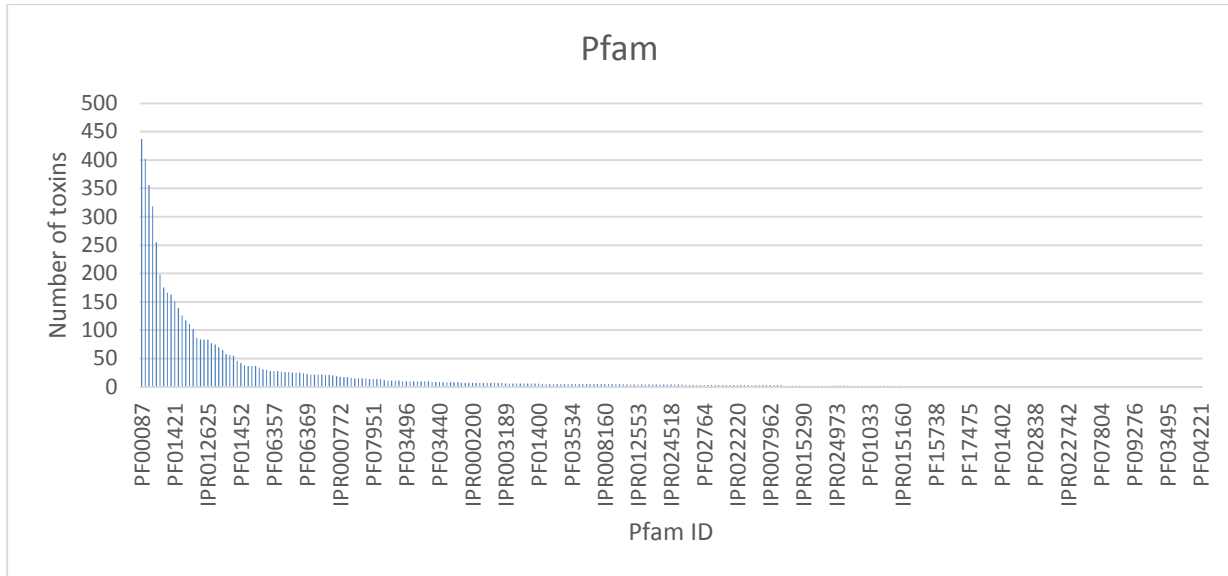


Figure 20: Number of toxins for each Pfam entry

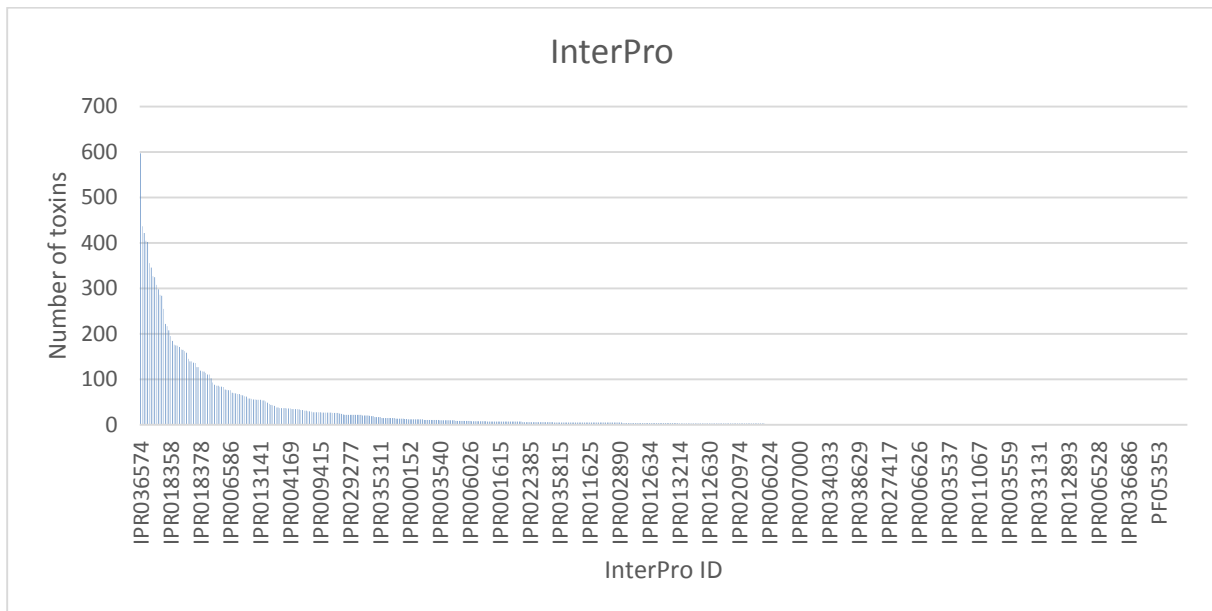


Figure 21: Number of toxins for each InterPro entry

Some examples of Pfam results are reported below. All the family descriptions are extracted from the Pfam web site. Images were produced after primary structure-based alignment with Clustal Omega and 3D structures superposition with MOE. Proteins are coloured according to RMSD within the structures of the family. For each family, a reference structure was selected to highlight architecture (domain composition), conserved within the family. With regards to the toxins identified in the previous exercise, these toxins in the Main Collection can be classified into 290 families, in which 595 different domains can be found. Most of the identified protein families are multidomain, i.e. contain more than a single domain in the same architecture. The complete list of Pfam families and their description, can be found in Annex E.

Alpha-2-macroglobulin family (A2M)

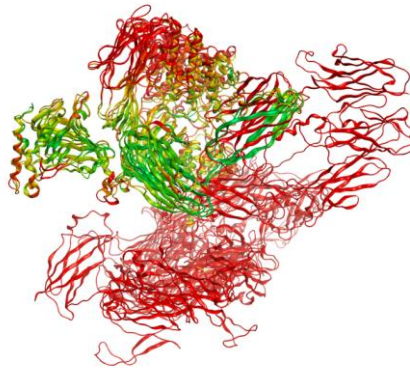


Figure 22: Superposition of A2M family members. Secondary structures are coloured according to RMSD (green represents low RMSD regions while red represents high RMSD regions)

This family includes the C-terminal region of the alpha-2-macroglobulin family (Figure 22 and Table 26).⁴

Reference structure: Q0ZZJ6

Table 26: Q0ZZJ6 architecture

Source	Domain	Start	End
Pfam	MG1	22	122
Pfam	MG2	127	216
Pfam	MG3	218	303
Pfam	MG4	345	435
Pfam	A2M_BRD	445	587
Pfam	ANATO	677	712
Pfam	A2M	754	850
Pfam	TED_complement	963	1263
Pfam	A2M_recep	1377	1472
Pfam	NTR	1513	1623

A2M-BRD



Figure 23: Superposition of A2M_BRD family members. Secondary structures are coloured according to RMSD (green represents low RMSD regions while red represents high RMSD regions)

⁴ Pfam definition

This family includes the C-terminal region of the alpha-2-macroglobulin family (as A2M) (Figure 23 and Table 27).

Reference structure: Q0ZZJ6

Table 27: Q0ZZJ6 architecture

Source	Domain	Start	End
Pfam	MG1	22	122
Pfam	MG2	127	216
Pfam	MG3	218	303
Pfam	MG4	345	435
Pfam	A2M_BRD	445	587
Pfam	ANATO	677	712
Pfam	A2M	754	850
Pfam	TED_complement	963	1263
Pfam	A2M_recep	1377	1472
Pfam	NTR	1513	1623

A2M_recep

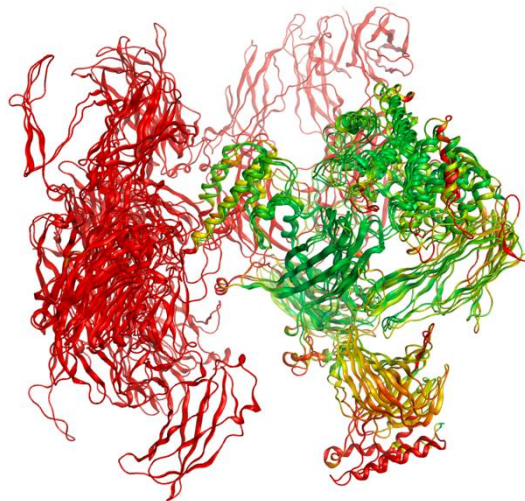


Figure 24: Superposition of A2M_recep family members. Secondary structures are coloured according to RMSD (green represents low RMSD regions while red represents high RMSD regions)

This family includes the C-terminal region of the alpha-2-macroglobulin family (as A2M) (Figure 24 and Table 28).⁴

Reference structure: Q0ZZJ6

Table 28: Q0ZZJ6 architecture

Source	Domain	Start	End
Pfam	MG1	22	122
Pfam	MG2	127	216
Pfam	MG3	218	303
Pfam	MG4	345	435
Pfam	A2M_BRD	445	587
Pfam	ANATO	677	712
Pfam	A2M	754	850
Pfam	TED_complement	963	1263
Pfam	A2M_recep	1377	1472
Pfam	NTR	1513	1623

ATPase family associated with various cellular activities (AAA)

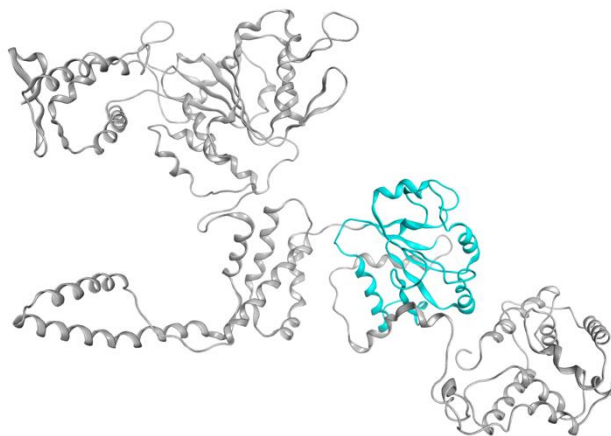


Figure 25: AAA family member. Secondary structures are coloured according to Pfam-defined domains (light blues represents the AAA domain while grey represents other domains, since it is the only member of the AAA family classified as toxin and reviewed.)

Proteins of the AAA family often perform chaperone-like functions that assist in the assembly, operation, or disassembly of protein complexes (Table 29).⁴

Reference structure: q54316 (Figure 25)

Table 29: q54316 architecture

Source	Domain	Start	End
Pfam	Clp_N	18	69
Pfam	Clp_N	94	146
Pfam	AAA	213	347
Pfam	AAA_lid_9	352	456
Pfam	AAA_2	545	717
Pfam	ClpB_D2-small	723	803

Some examples of InterPro results are reported in the following lines. All the domain/signature descriptions are extracted from InterPro web site. For the complete list of domains and their description see Annex F.

IPR000020 - Anaphylatoxin/fibulin - Domain. 2606 proteins.

This entry represents C3a, C4a and C5a anaphylatoxins, which are protein fragments generated enzymatically in serum during activation of complement molecules C3, C4, and C5. They induce smooth muscle contraction. These fragments are homologous to a three-fold repeat in fibulins.

Complement components C3, C4 and C5 are large glycoproteins that have important functions in the immune response and host defence. They have a wide variety of biological activities and are proteolytically activated by cleavage at a specific site, forming a- and b-fragments. a-fragments form distinct structural domains of approximately 76 amino acids, coded for by a single exon within the complement protein gene. The C3a, C4a and C5a components are referred to as anaphylatoxins: they cause smooth muscle contraction, histamine release from mast cells, and enhanced vascular permeability. They also mediate chemotaxis, inflammation, and generation of cytotoxic oxygen radicals. The proteins are highly hydrophilic, with a mainly alpha-helical structure held together by 3 disulphide bridges.

Fibulins are secreted glycoproteins that become incorporated into a fibrillar extracellular matrix when expressed by cultured cells or added exogenously to cell monolayers. The five known members of the family share an elongated structure and many calcium-binding sites, owing to the presence of tandem arrays of epidermal growth factor-like domains. They have overlapping binding sites for several basement-membrane proteins, tropoelastin, fibrillin, fibronectin and proteoglycans, and they participate

in diverse supramolecular structures. The amino-terminal domain I of fibulin consists of three anaphylatoxin-like (AT) modules, each approximately 40 residues long and containing four or six cysteines. The structure of an AT module was determined for the complement-derived anaphylatoxin C3a, and was found to be a compact alpha-helical fold that is stabilised by three disulphide bridges in the pattern Cys1-4, Cys2-5 and Cys3-6 (where Cys is cysteine). The bulk of the remaining portion of the fibulin molecule is a series of nine EGF-like repeats.

References:

C5a fragment of bovine complement. Purification, bioassays, amino-acid sequence and other structural studies. Gennaro R, Simonic T, Negri A, Mottola C, Secchi C, Ronchi S, Romeo D. Eur. J. Biochem. **155** 1 (1986) 77-86. PUBMED 3081348

Sequence of the gene for murine complement component C4. Ogata RT, Rosa PA, Zepf NE. J. Biol. Chem. **264** 28 (1989) 16565-72. PUBMED 2777798

Fibulin is an extracellular matrix and plasma glycoprotein with repeated domain structure. Argraves WS, Tran H, Burgess WH, Dickerson K. J. Cell Biol. **111** 6 Pt 2 (1990) 3155-64 . PUBMED 2269669

Fibulins: a versatile family of extracellular matrix proteins. Timpl R, Sasaki T, Kostka G, Chu ML. Nat. Rev. Mol. Cell Biol. **4** 6 (2003) 479-89. PUBMED 12778127

Structure and expression of fibulin-2, a novel extracellular matrix protein with multiple EGF-like repeats and consensus motifs for calcium binding. Pan TC, Sasaki T, Zhang RZ, Fassler R, Timpl R, Chu ML. J. Cell Biol. **123** 5 (1993) 1269-77. PUBMED 8245130

Primary structure of cobra complement component C3. Fritzinger DC, Petrella EC, Connelly MB, Bredehorst R, Vogel CW. J. Immunol. **149** 11 (1992) 3554-62. PUBMED 1431125

IPR000101 - Gamma-glutamyltranspeptidase - Family. 28201 proteins.

Gamma-glutamyltranspeptidase ([intenz:2.3.2.2]) (GGT) catalyzes the transfer of the gamma-glutamyl moiety of glutathione to an acceptor that may be an amino acid, a peptide or water (forming glutamate). GGT plays a key role in the gamma-glutamyl cycle, a pathway for the synthesis and degradation of glutathione and drug and xenobiotic detoxification. In prokaryotes and eukaryotes, it is an enzyme that consists of two polypeptide chains, a heavy and a light subunit, processed from a single chain precursor by an autocatalytic cleavage. The active site of GGT is known to be located in the light subunit. The sequences of mammalian and bacterial GGT show a number of regions of high similarity. Pseudomonas cephalosporin acylases ([intenz:3.5.1]) that convert 7-beta-(4-carboxybutanamido)-cephalosporanic acid (GL-7ACA) into 7-aminocephalosporanic acid (7ACA) and glutaric acid are evolutionary related to GGT and also show some GGT activity. Like GGT, these GL-7ACA acylases, are also composed of two subunits. This entry also includes the highly similar Scoloptoxin SSD14 from Scolopendra dehaani. SSD14 (which is also cleaved into alpha and beta subunits) has been shown to induce human platelet aggregation.

As an autocatalytic peptidase GGT belongs to MEROPS peptidase family T3 (gamma-glutamyltransferase family, clan PB(T)). The active site residue for members of this family and family T1 is C-terminal to the autolytic cleavage site. The type example is gamma-glutamyltransferase 1 from Escherichia coli.

References:

Nucleotide sequence and expression in Escherichia coli of the cephalosporin acylase gene of a Pseudomonas strain. Ishiye M, Niwa M. Biochim. Biophys. Acta **1132** 3 (1992) 233-9. PUBMED 1358202

DNA sequence of the Escherichia coli K-12 gamma-glutamyltranspeptidase gene, ggt. Suzuki H, Kumagai H, Echigo T, Tochikura T. J. Bacteriol. **171** 9 (1989) 5169-72 . PUBMED 2570061

gamma-Glutamyl transpeptidase from kidney. Tate SS, Meister A. Meth. Enzymol. **113** (1985) 400-19 . PUBMED 2868390

Gamma-glutamyltransferase: nucleotide sequence of the human pancreatic cDNA. Evidence for a ubiquitous gamma-glutamyltransferase polypeptide in human tissues. Courtay C, Oster T, Michelet F, Visvikis A, Diederich M, Wellman M, Siest G. Biochem. Pharmacol. **43** 12 (1992) 2527-33 . PUBMED 1378736

Venomic and transcriptomic analysis of centipede Scolopendra subspinipes dehaani. Liu ZC, Zhang R, Zhao F, Chen ZM, Liu HW, Wang YJ, Jiang P, Zhang Y, Wu Y, Ding JP, Lee WH, Zhang Y. J. Proteome Res. **11** 12 (2012) 6197-212 . PUBMED 23148443

IPR000126 - Serine proteases, V8 family, serine active site - Active_site. 1126 proteins.

A number of prokaryotic proteases have been shown to be evolutionary related; their catalytic activity is provided by a charge relay system similar to that of the trypsin family of serine proteases but which probably evolved by independent convergent evolution. The sequence around the residues involved in the catalytic triad (aspartic acid, serine and histidine) are completely different from that of the analogous residues in the trypsin serine proteases and can be used as signatures specific to that category of proteases. The proteases which are known to belong to this family are listed below.

This entry represents the serine active site of serine proteases from the V8 family.

Staphylococcus aureus V8 proteinase, which preferentially cleaves peptide bonds on the carboxyl-terminal side of aspartate and glutamate and which is widely used in protein sequencing studies.

Bacillus licheniformis glutamate specific endopeptidase (GSE), which like V8 cleaves on the carboxyl-terminal side of acidic residues, but with a strong preference for glutamate.

Bacillus subtilis extracellular "metalloprotease" (gene *mpr*).

Staphylococcus aureus exfoliative (or epidermolytic) toxins A (gene *eta*) and B (gene *etb*). These toxins cause impetiginous diseases commonly referred to as staphylococcal scalded skin syndrome (SSSS) and have been shown to possess proteolytic activity.

References:

Isolation and amino acid sequence of a glutamic acid specific endopeptidase from *Bacillus licheniformis*. Svendsen I, Breddam K. Eur. J. Biochem. **204** **1** (1992) 165-71 . PUBMED 1346764

The epidermolytic toxins are serine proteases. Dancer SJ, Garratt R, Saldanha J, Jhoti H, Evans R. FEBS Lett. **268** **1** (1990) 129-32 . PUBMED 2384148

Gene encoding a novel extracellular metalloprotease in *Bacillus subtilis*. Sloma A, Rudolph CF, Rufo GA Jr, Sullivan BJ, Theriault KA, Ally D, Pero J. J. Bacteriol. **172** **2** (1990) 1024-9 . PUBMED 2105291

Chemistry of collagen cross-linking: biochemical changes in collagen during the partial mineralization of turkey leg tendon. Knott L, Tarlton JF, Bailey AJ. Biochem. J. **322** (Pt 2) (1997) 535-42 . PUBMED 9065774

3.1.2. TAS Collection

As detailed in Section 2.2.1, proteins belonging to a toxin-antitoxin system (TAS) are identified in UniProtKB by the keyword "Toxin-antitoxin system" This chapter summarizes findings on TAS proteins, as covered by the entries in the TAS Collection of additional Excel files (see Table 15).

3.1.2.1. Number of proteins identified

The data on the 627 proteins identified as belonging to a toxin-antitoxin system (Keyword "Toxin-antitoxin system Reviewed:Yes") were downloaded from the UniProtKB database (March 2020). Among them, 624 have precursors into their sequence. All the selected proteins are the toxic component of the toxin-antitoxin system. Full details are included in Annexes C, D and in the Excel files (see Table 15); an example is presented in Table 30 below.

Table 30: Example of primary structure Table (See TC_sequences.xlsx)

Uniprot	Fragment	Precursor	Structure	Length	Experimental determination of the structure	Best pdb
Q46863		true	Swiss Model	535	0%	
P62555			Swiss Model	101	0%	
Q46995			Swiss Model	72	0%	
P62553			PDB	72	50%	3tcj
P64525				124	0%	
...

Fragment: is used to classify if the protein is only a fragment of the primary sequence

Precursor: (logic) does the toxin have a precursor?

Structure: source of the selected structure download

Length: number of amino acids of the primary structure

Experimental determination of the structure: experimentally-derived part of the protein (in percentage with respect to the length of primary structure)

Best pdb: the protein data bank code of the best resolved structure.

3.1.2.2. Sequence length

As shown in Figure 26, ~95% of the identified proteins have a sequence length of less than 200 amino acids, while ~50% of those proteins are shorter than 100 amino acids. Only 5 proteins are shorter than 20 amino acids and were not modelled (Figure 27).

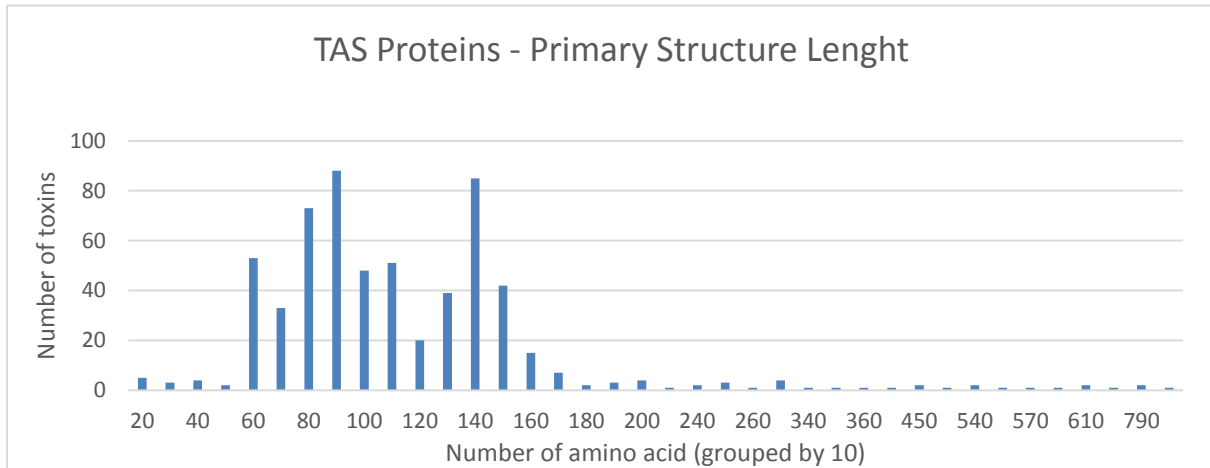


Figure 26: Histogram of the primary structure length for TAS proteins

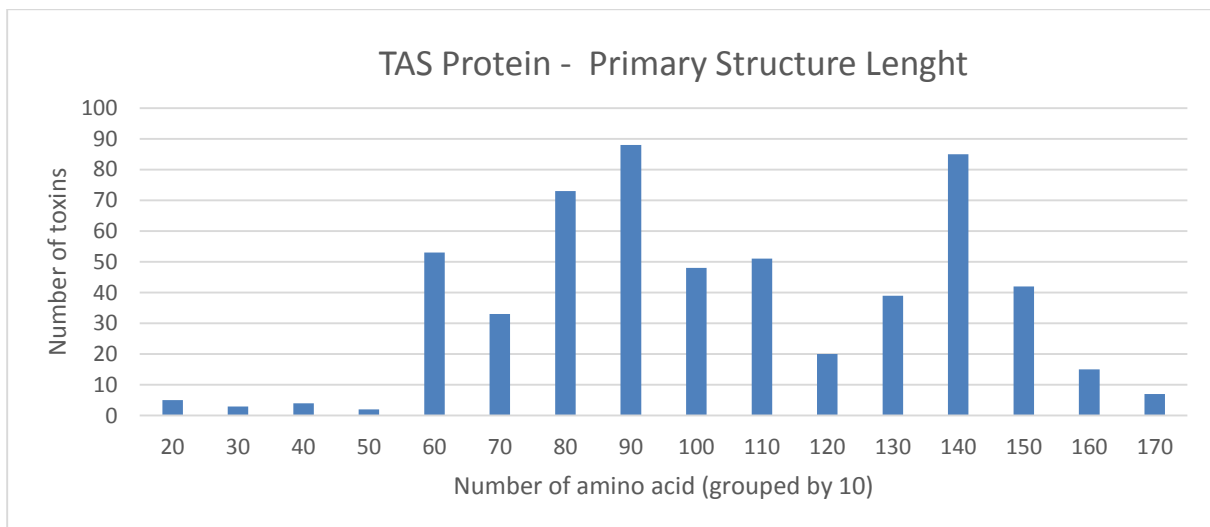


Figure 27: Histogram of the primary structure length for TAS proteins (up to 170 amino acids in length)

3.1.2.3. Organisms

Reviewing the information on the organisms that express the selected toxins, we found that the most representative (i.e. those expressing the highest number of toxins) is *Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv) (145 proteins), followed by *Mycobacterium tuberculosis* (strain CDC 1551 / Oshkosh) (95 proteins), *Escherichia coli* (67 proteins), *Methanocaldococcus jannaschii* (20 proteins) and *Archaeoglobus fulgidus* (19 proteins).

Figure 28 shows the histogram summarising the information about the number of proteins in the TAS Collection on the basis of the organism producing them. In the below section, a brief description is reported for the five most representative organisms Table 31 exemplifies data summarised in the dedicated go.xlsx file (see Table 15).

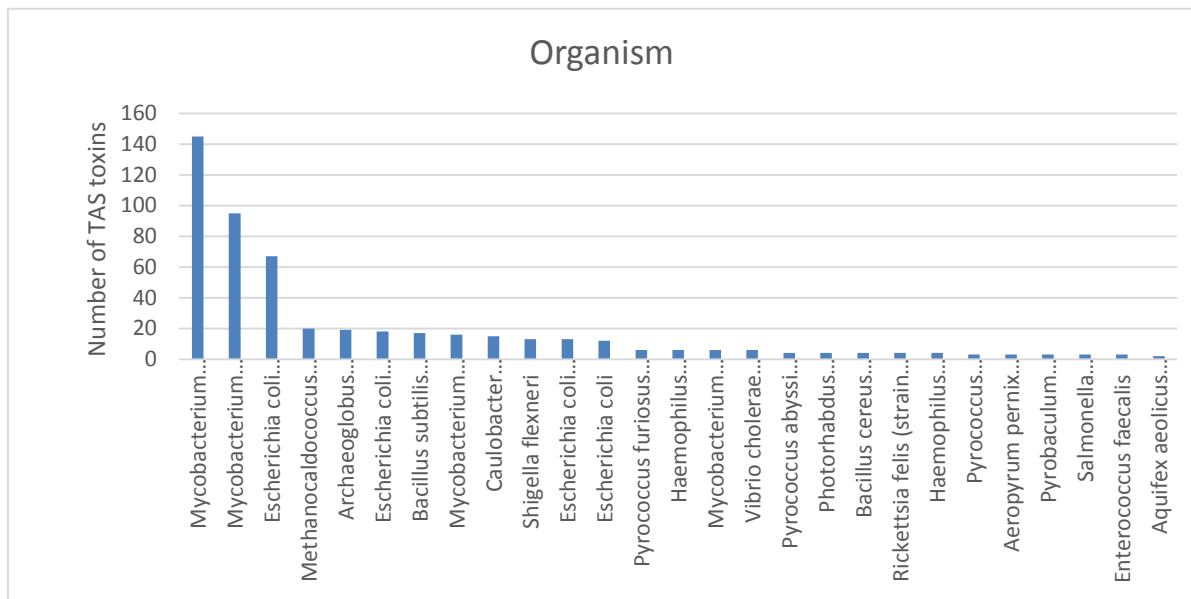
- i) *Mycobacterium tuberculosis* is an acid-fast, obligate aerobic, non-motile, rod-shaped bacterium and it is the causative agent of tuberculosis. The non-replicating persistent

- form is refractory to most drugs, and may be induced by hypoxia (oxygen depletion) and/or nitric oxide exposure;
- ii) *Escherichia coli* is a Gram-negative straight rod, which either uses peritrichous flagella for mobility or is nonmotile. Pathogenic *E.coli* strains are responsible for infection of the enteric, urinary, pulmonary and nervous systems;
 - iii) *Methanocaldococcus jannaschii* is a thermophilic *methanogenic archaean*, meaning that it grows by making methane as a metabolic by-product, in the class *Methanococci*;
 - iv) *Archaeoglobus fulgidus* is the first sulphur-metabolizing organism to have its genome sequence determined. When subjected to stress, *Archaeoglobus fulgidus* can produce biofilms that have been applied for industrial purposes e.g. detoxifying contamination by metals; since it can survive at extremely high temperatures, there are hopes that heat stable enzymes may be derived from it.

Table 31: Extract of organisms table (TC_organisms.xlsx)

UniProt	Organism
P64524	Escherichia coli (strain K12)
P62555	Escherichia coli O157:H7
O34853	Bacillus subtilis (strain 168)
P39394	Escherichia coli (strain K12)
Q4QNL8	Haemophilus influenzae (strain 86-028NP)
P9WF90	Mycobacterium tuberculosis (strain CDC 1551 / Oshkosh)
...	...

UniProt: the UniProtKB identifier of the protein/peptide; Organism: organism of the organism that express the toxin.

**Figure 28:** Histogram of organism source of TAS proteins

3.1.2.4. Gene ontology (GO) terms

The GO terms (see Section 2.1.2) cover three domains: molecular function biological processes, and cellular components. GOs for the TAS proteins (TAS collection) were analysed according to these categories (see Figures 29-31 report the frequencies of the individual GOs in the TAS Collection). More than one GO can be associated with each toxin. Table 32 exemplifies data summarised in dedicated go.xlsx files (see Table 15).

Table 32: Extract of GO terms table (see TC_go.xlsx)

UniProt	ID	GO
P64524	GO:0008092	cytoskeletal protein binding
P62555	GO:0008657	DNA topoisomerase (ATP-hydrolyzing) inhibitor activity
Q46995	GO:0003677	DNA binding
Q8FDB5	GO:0003700	DNA-binding transcription factor activity
Q8FDB5	GO:0097351	toxin-antitoxin pair type II binding
...

UniProt: the UniProtKB identifier of the protein/peptide; ID: the Gene Ontology unique identifier; GO: the name of the GO.

In the following section, the most recurrent GOs for the identified proteins are reported for each category:

Molecular Function

- i) *magnesium ion binding*: interacting selectively and non-covalently with magnesium (Mg) ions;
- ii) *ribonuclease activity*: catalysis of the hydrolysis of phosphodiester bonds in chains of RNA;
- iii) *DNA binding*: any molecular function by which a gene product interacts selectively and non-covalently with DNA (deoxyribonucleic acid);
- iv) *endonuclease activity*: catalysis of the hydrolysis of ester linkages within nucleic acids by creating internal breaks;
- v) *RNA binding*: interacting selectively and non-covalently with an RNA molecule or a portion thereof.

Biological Process

- i) *regulation of transcription, DNA-templated*: any process that modulates the frequency, rate or extent of cellular DNA-templated transcription;
- ii) *negative regulation of growth*: any process that stops, prevents or reduces the rate or extent of growth, the increase in size or mass of all or part of an organism;
- iii) *positive regulation of growth*: any process that activates or increases the phospholipid metabolic process: the chemical reactions and pathways involving phospholipids, any lipid containing phosphoric acid as a mono- or diester.

Cellular Component

- i) *plasma membrane*: the membrane surrounding a cell that separates the cell from its external environment. It consists of a phospholipid bilayer and associated proteins;
- ii) *integral component of membrane*: the component of a membrane consisting of the gene products and protein complexes having at least some part of their peptide sequence embedded in the hydrophobic region of the membrane;
- iii) *extracellular region*: the space external to the outermost structure of a cell. For cells without external protective or external encapsulating structures this refers to space outside of the plasma membrane. This term covers the host cell environment outside an intracellular parasite.

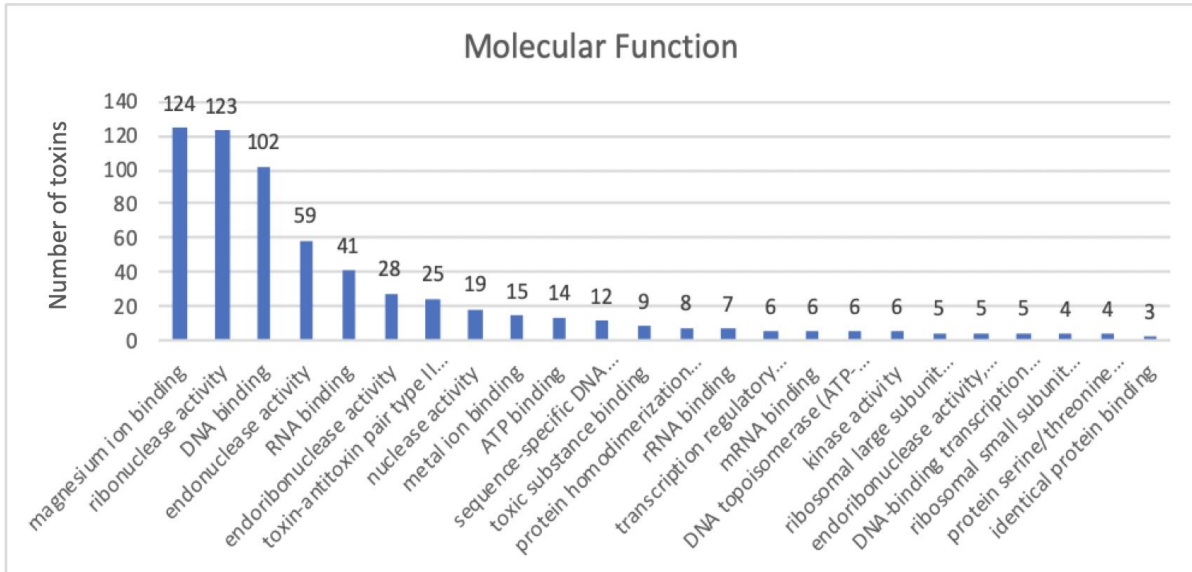


Figure 29: Frequency of the Molecular Function GO terms

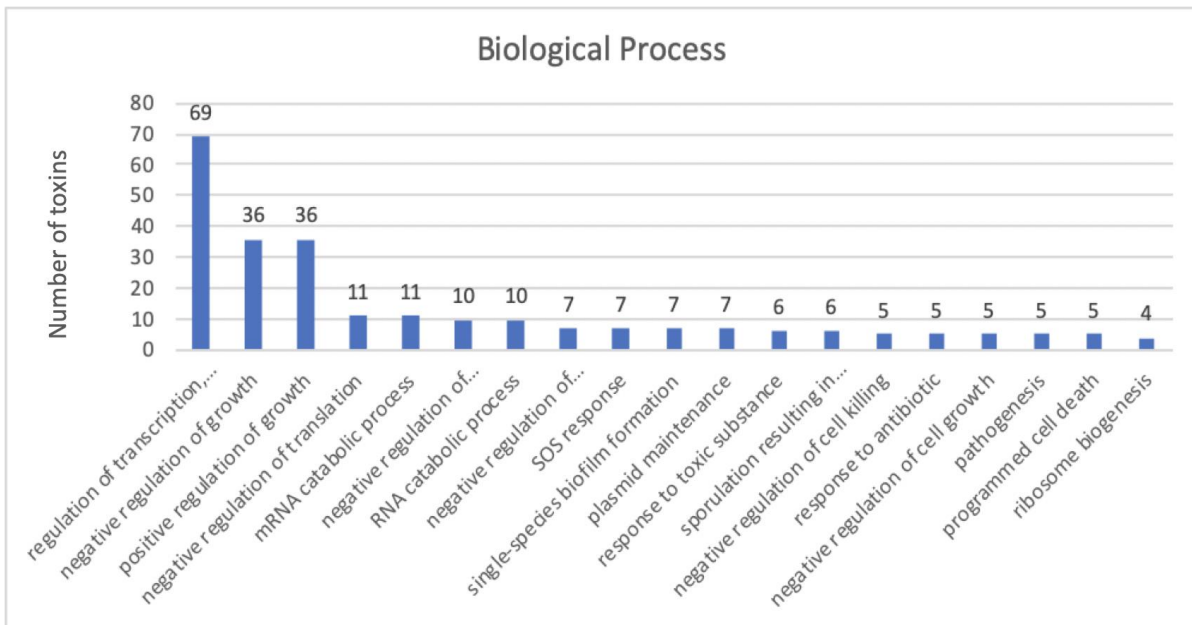


Figure 30: Frequency of the Biological Process GO terms

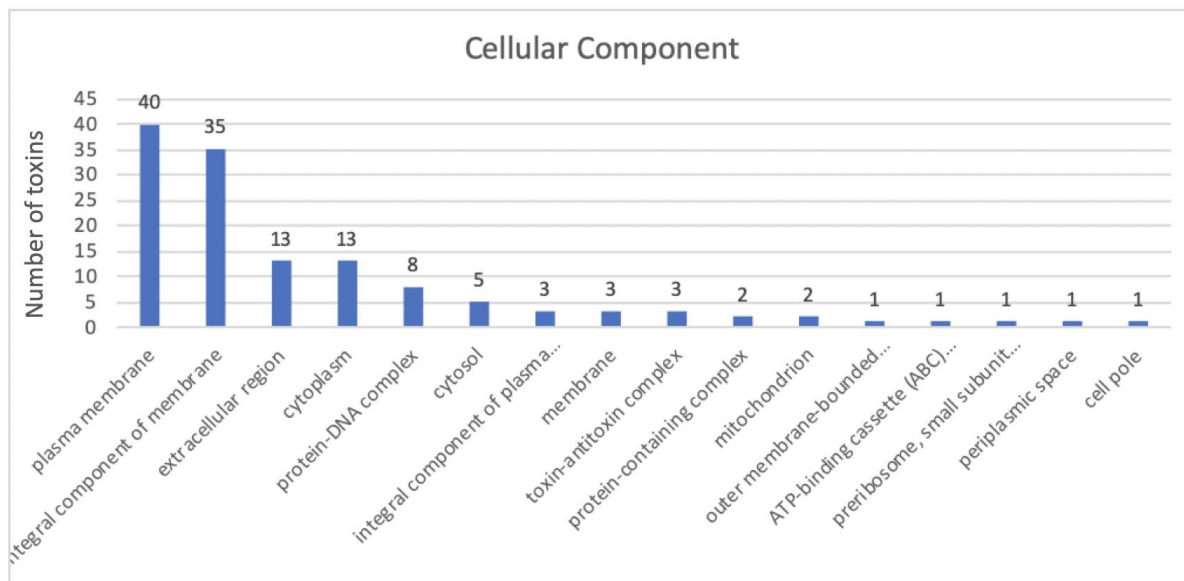


Figure 31: Frequency of the Cellular Component GO terms

3.1.2.5. Keywords

For each protein from the TAS collection, keywords were extracted from UniProtKB (unique ID and category) (Table 33). UniProtKB keywords constitute a controlled vocabulary with a hierarchical structure, different from GO. Keywords are native of UniprotKB and summarize the content of a UniProtKB entry, facilitating the search for proteins of interest. Keywords were assigned by UniProtKB curators and are classifiable in 8 categories:

- Toxin-antitoxin system
- Complete proteome
- Reference proteome
- Hydrolase
- Nuclease
- Metal-binding
- Magnesium
- 3D-structure

Table 33: Extract of keywords table (TC_keywords.xlsx)

Uniprot	ID	Keyword	Category
P64524	KW-0133	Cell shape	Biological process
P64524	KW-0181	Complete proteome	Technical term
P64524	KW-0963	Cytoplasm	Cellular component
P64524	KW-1185	Reference proteome	Technical term
P64524	KW-1277	Toxin-antitoxin system	Biological process
P62555	KW-0181	Complete proteome	Technical term
...

UniProt: the UniProtKB identifier of the protein/peptide; ID: the keyword unique identifier; Keyword: the keyword; Category: the category in which the keyword is categorized.

An entry often contains several keywords. Within a category, the keywords are stored in alphabetical order. Some keywords are also related to GO. Figure 32 reports the count of all the keywords, not divided into categories. The following section describes in more detail the main keywords based on their frequency (>100 counts):

- i) Toxin-antitoxin system
Definition: Bacterial, archaeal or viral protein belonging to a toxin-antitoxin (TA) system. Toxin protein expression is poisonous to the cell and is counteracted by its co-expressed, short-lived "antidote" antitoxin. Antitoxins can be protein (type II, IV, V, VI) or RNA (type I, III) and act directly or indirectly on the toxin protein or transcript. Genes for the two components are closely linked; in many type II TA systems the toxin-antitoxin complex regulates transcription of its operon. Toxins belong to recognizable protein families, antitoxins usually do not. Orphan toxin genes are occasionally found in complete genomes.
Category: Biological process
- ii) Complete proteome
Definition: Protein which is part of a proteome. A proteome is the set of protein sequences that can be derived by translation of all protein coding genes of a completely sequenced genome, including alternative products such as splice variants for those species in which these may occur. Proteomes may include protein sequences from both the reviewed (UniProtKB/Swiss-Prot) and unreviewed (UniProtKB/TrEMBL) sections of the UniProt Knowledgebase. Note that some proportion of the predicted protein sequences of a given proteome may require further review or correction. The precise proportion depends on the relative distributions of protein sequences between the two sections of UniProtKB and the quality of the underlying genome sequence and gene predictions.
Category: Technical term
- iii) Reference proteome
Definition: Protein which is part of a reference proteome. Reference proteomes are a subset of proteomes that have been selected either manually or algorithmically according to a number of criteria to provide a broad coverage of the tree of life and a representative cross-section of the taxonomic diversity found within UniProtKB, as well as the proteomes of well-studied model organisms and other species of interest for biomedical research.
Category: Technical term
- iv) Hydrolase
Definition: Enzyme which catalyzes hydrolysis reaction, i.e. the addition of the hydrogen and hydroxyl ions of water to a molecule with its consequent splitting into two or more simpler molecules.
Category: Molecular function
GO: hydrolase activity [GO:0016787]
- v) Nuclease
Definition: Enzyme that degrades nucleic acids into shorter oligonucleotides or single nucleotide subunits by hydrolyzing sugar-phosphate bonds in the nucleic acid backbone.
Category: Molecular function
GO: nuclease activity
- vi) Metal-binding
Definition: Protein which binds metals.
Category: Ligand
GO: metal ion binding
- vii) Magnesium
Definition: Protein which binds at least one magnesium atom, or protein whose function is magnesium-dependent. Magnesium is a metallic element, chemical symbol Mg.
Category: Ligand
- viii) 3D-structure
Definition: Protein, or part of a protein, whose three-dimensional structure has been resolved experimentally (for example by X-ray crystallography or NMR spectroscopy) and whose coordinates are available in the PDB database. Can also be used for theoretical models.

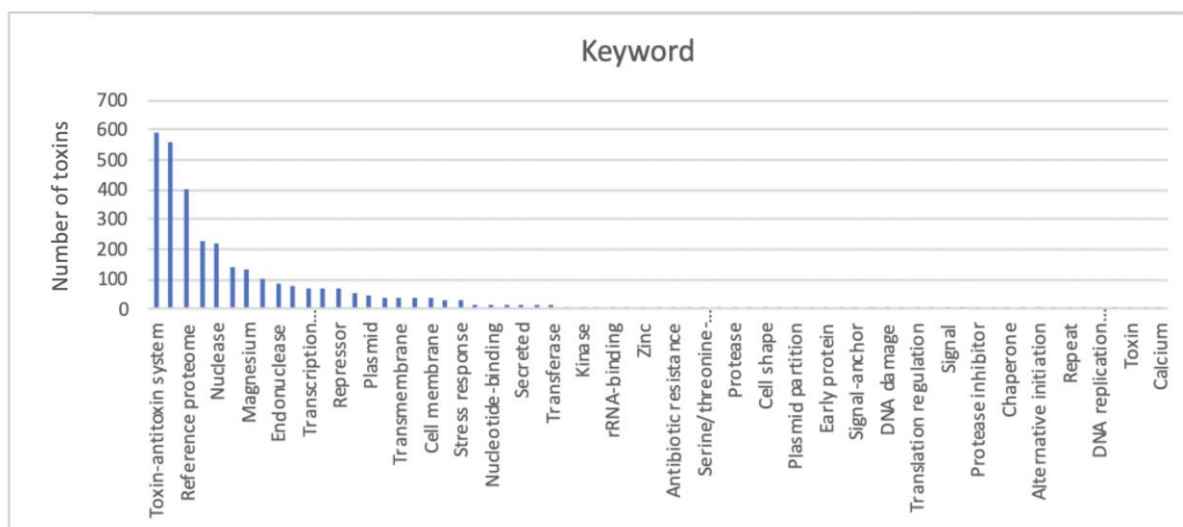


Figure 32: Frequency of keywords

Full information on Keywords and their count can be found in the dedicated Excel files (TC_sequences.xlsx file)

3.1.2.6. Three-dimensional structures

Out of the 627 identified TAS proteins, 114 have associated one or more experimentally-derived 3D structures. As described in the Materials and methods section, 3D structures were carefully checked and only one structure was selected for each protein. Information about the experimental determination of the structure such as method, resolution and percentage of sequence coverage were collected from the PDB source (see Table 15). Table 34 exemplifies the PDB table.

Table 34: Extract of PDB table (see TC_pdb.xlsx)

UniProt	PDB	Method	Resolution [Å]	Experimental determination of the structure
P62553	3tcj	X-ray	1.93	50%
O34853	3o6q	X-ray	2.5	63.3%
E6Z0R3	3shg	X-ray	1.5	41.2%
P9WFA5	6a7v	X-ray	1.67	99.3%
Q8U3V0	1y82	X-ray	2.3	99.3%
A0A140NAP5	4ml0	X-ray	2.1	100%
O53778	5x3t	X-ray	2.65	100%
...

UniProt: the UniProtKB identifier of the protein/peptide; PDB: the RCSB PDB identifier of the selected structure; Method: 3D structure solving method; Resolution: resolution of the selected 3D structure; Experimental determination of the structure: percentage of protein sequence covered in the experimentally-solved structure.

Figure 33 reports the number of experimentally-determined structures for each protein (labelled with its UniProtKB code), while Table 35 reports the number of available 3D structures in the PDB according to the determination methodology used. 'Model' belongs to an obsolete RCSB pdb policy under which also some protein theoretically-built models were available on the PDB source. This type of entry is now deprecated.

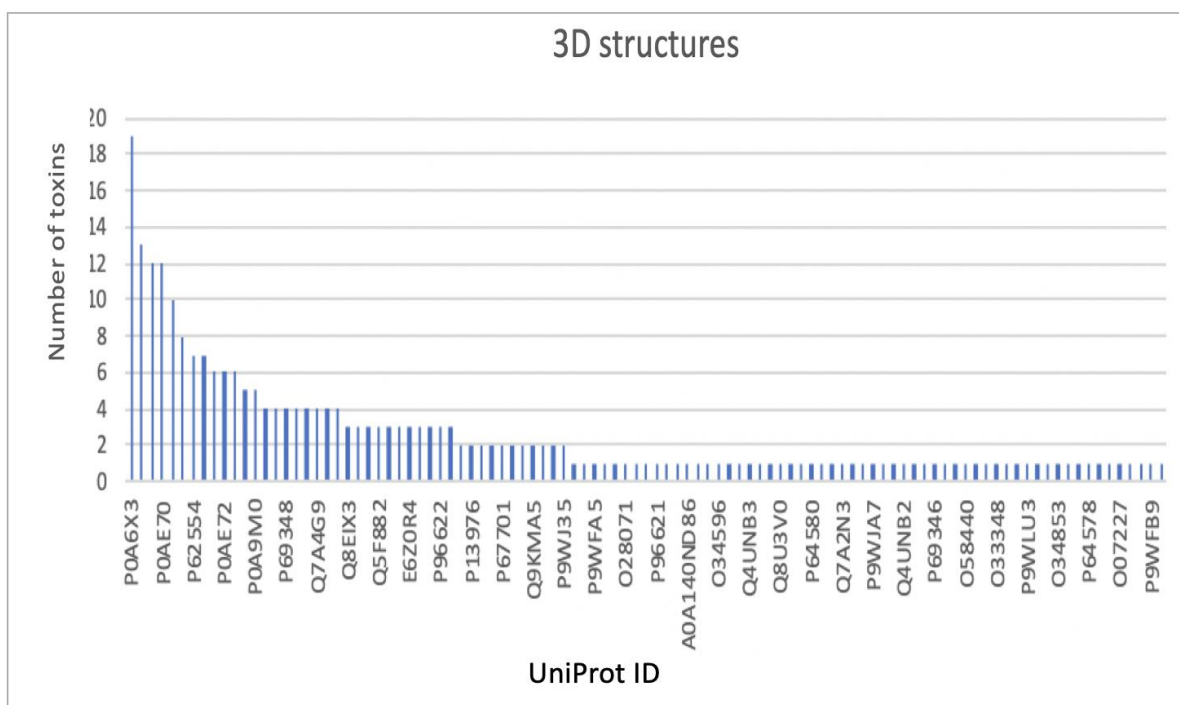


Figure 33: Statistic of experimentally-determined structures per protein.

Table 35: Total number of experimentally-solved structures by method

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

Moreover, a total of 356 models were downloaded from the SM repository and stored into the TAS Collection. Also in this case there are some TAS proteins with 2 or more associated models in the SM repository. As stated in the Materials and Methods section, for our *in silico* analysis, we selected the one with best quality check parameters.

Information about the template, identity and percentage of sequence coverage are shown in a summarizing table that also includes quality parameters (Table 36).

Table 36: Extract of SM table (see TC_swiss-models.xlsx)

UniProt	Template	Identity	Oligo	Coverage	Qmean	Qmean norm	Gmqe	Origin
P9WF86	5wzf.1.a	17.3	Homo-2-mer	0.917	-3.898	0.616	0.502	SM
P39394	4qo6.1.a	30.9	Monomer	0.513	-1.649	0.661	0.293	SM
Q4QNL8	6nkl.1.c	44.2	Monomer	0.948	-1.837	0.677	0.554	SM
P9WF90	5sv2.1.a	99.2	Homo-2-mer	0.978	-0.554	0.751	0.981	SM
Q9Z4V7	3d55.1.c	63.4	Heteromer	0.781	-0.503	0.760	0.78	SM
POCW38	5g5s.1.a	100	Monomer	0.982	-0.172	0.781	0.985	SM
...

UniProt: the UniProtKB identifier of the protein/peptide; Template: the PDB code of the 3D protein structure used for homology modelling; Identity: percentage of identity between template and query sequence; Oligo: ; Coverage: percentage normalized of the modelled structure coverage; Qmean (Benkert et al., 2011): parameter involved in the quality estimation (see Methods); Qmean norm: parameter involved in the quality estimation; Gmqe: parameter involved in the quality estimation (see Methods); Origin: tool used for the homology modelling (SM = Swiss-Model; CPH = CPH Model).

As shown in Figure 34, some templates were used to model different proteins (e.g. PDB ID: 2FE1) that belong to the same family.

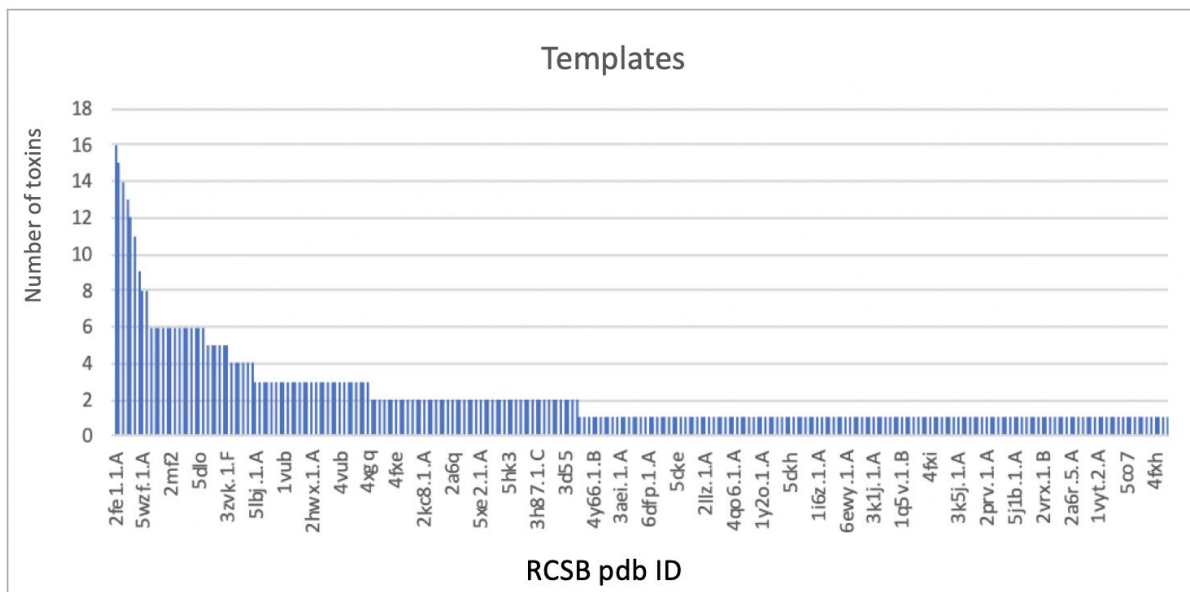


Figure 34: Statistics of experimentally-solved structures per protein

3.1.2.7. Families, domains and signatures

For each source (Table 37), families/domains classification was carefully checked and Pfam was selected as the reference source for protein families, while InterPro was chosen as the reference source for both domains and signatures. All the other domain databases can be reported to the classification in InterPro, which is also the most up-to-date database. All the information extracted from both Pfam and InterPro is reported in Annex E and F, respectively, while Figures 35 and 36 show the number of proteins for each family or domain, respectively. Signatures are considered in the domain classification.

Table 37: Number of families/domains by database (see TC_databases.xlsx)

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

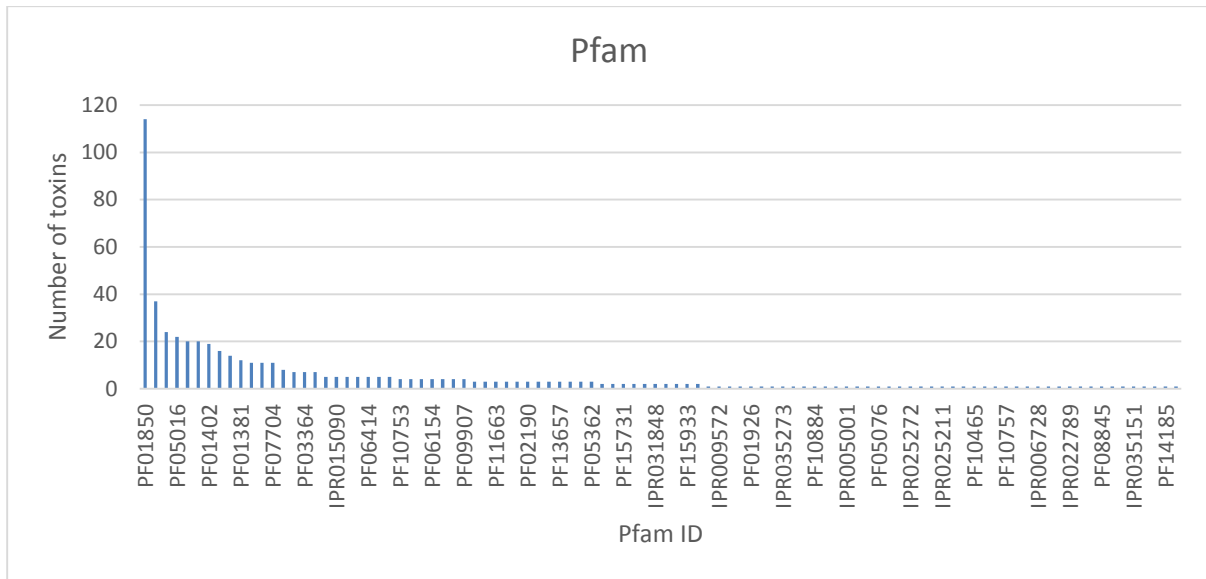


Figure 35: Number of toxins for each Pfam entry

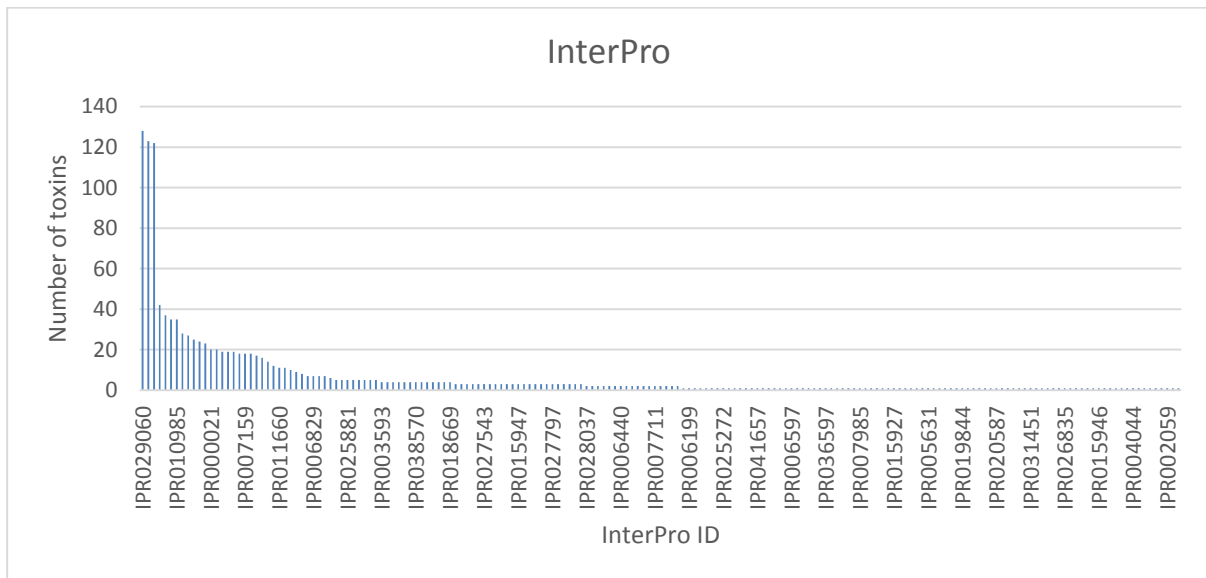
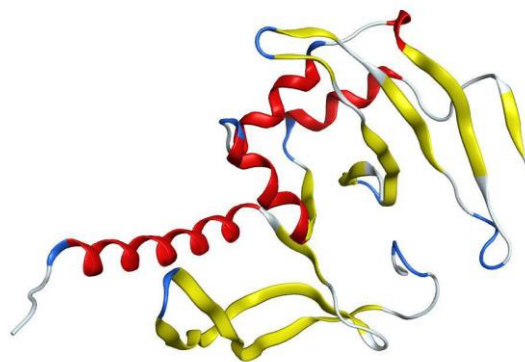


Figure 36: Number of toxins for each InterPro entry

For each protein, information was downloaded from UniProtKB and linked to the literature. Below, two examples are reported to distinguish between a well-annotated protein (Q46865) and a poorly-annotated one (P0A7G6). All the information extracted is reported in Annex C.

Q46865 (MQSR_ECOLI)**Figure 37:** Crystallographic structure of Q46865 (PDB ID: 3HI2).**Function**

Toxic component of a type II toxin-antitoxin (TA) system (Figure 37). Plays a significant role in the control of biofilm formation and induction of persister cells in the presence of antibiotics. An mRNA interferase which has been reported to be translation independent. It has also been reported to be translation dependent. Cleavage has been reported to occur on either side of G in the sequence GCU. Also reported to cleave after C in GC(A/U) sequences. There are only 14 genes in E.coli W3110 (and probably also MG1655) that do not have a GCU sequence and thus are resistant to the mRNA interferase activity; among these is the gene for toxin GhoT. Overexpression of MqsR causes cessation of cell growth and inhibits cell proliferation via inhibition of translation as well as increasing persister cell formation; these effects are overcome by concomitant or subsequent expression of antitoxin MqsA. Cross-talk can occur between different TA systems. Ectopic expression of this toxin induces transcription of the relBEF TA system operon with specific cleavage of the relBEF mRNA produced. Regulates the expression of GhoT/GhoS, a type V TA system. Persistence depends on toxin GhoT activity, which MqsR controls at the post-transcriptional level by selectively degrading the antitoxin ghoS segment of the ghoST mRNA. Persister cells exhibit antibiotic tolerance without genetic change. mRNA interferases play a role in bacterial persistence to antibiotics; overexpression of this protein induces persisters resistant to ciprofloxacin and ampicillin. Overexpression leads to a dramatic increase in tolerance to the antibiotic ofloxacin. This TA system mediates cell growth during bile acid deoxycholate stress by degrading mRNA for probable deoxycholate-binding protein YgiS; bile acid detergents such as deoxycholate are important for host defense against bacterial growth in the gall bladder and duodenum.

Initially reported to act as a cotranscription factor with MqsA. Following further experiments, the MqsR-MqsA complex does not bind DNA and all reported data are actually due to a small fraction of free MqsA alone binding DNA. Addition of MqsR to a preformed MqsA-promoter DNA complex causes dissociation of the MqsA-DNA complex, probably causing derepression of MqsA-repressed transcripts. Does not bind DNA in the presence or absence of MqsA.

Temperature dependence

The MqsR-MqsA complex is exceptionally thermostable with a T_m of 83.4 degrees Celsius versus 48.1 degrees Celsius for MqsR and 61.1 degrees Celsius for MqsA.

Disruption phenotype

No loss of ability to form persister cells in MG1655, represses persister cell formation in BW25113. Deletion decreases biofilm formation in LB medium. Deletion has also been shown to increase biofilm formation. Deletion at 48h, in flow cells, leads to a reduction in biomass, substratum coverage and changes the biofilm architecture from a 54-micron thick film with microcolonies to one with nearly no biomass and only a few colonies remaining. Deletion abolishes motility. A double mqsR-mqsA deletion leads to increased rpoS mRNA levels, resulting in increased cyclic-di-GMP levels, increasing stress resistance, increased biofilm formation. The double mutant has increased metabolism and respiration in the presence of the bile acid deoxycholate and consequently grows less well. Decreases cell survival

in the presence of 20% deoxycholate. mRNA interferases play a role in bacterial persistence to antibiotics; as 10 mRNA interferases are successively deleted reduced levels of persisters are generated.

Expression - Induction

Induced by amino acid starvation, glucose starvation and when translation is blocked. Also induced by nalidixic acid, azlocillin and H₂O₂. It has been suggested that MqsA represses its own operon. Induction is decreased in the absence of the Lon protease suggesting, by homology to other toxin-antitoxin systems, that Lon may degrade the MqsA antitoxin. Transcription is activated by MqsA. A member of the mqsRA operon. Most highly induced gene in persister cells. Degrades its own transcript. This operon induced by ectopic expression of toxins RelE, HicA and YafQ but not by MazF or HicA.

Subunit structure

Might be a dimer. Also reported to be a monomer. Crystallizes as a heterotetramer with MqsA, MqsR-MqsA2-MqsR. Purifies as a possible heterohexamer of 2 MqsR dimers and 1 MqsA dimer. When the 2 dissociate the MqsR mRNA interferase becomes active.

P0A7G6 (RECA_ECOLI)

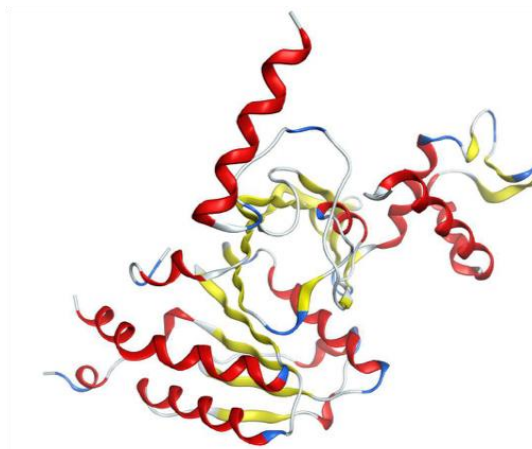


Figure 38: Crystallographic structure of P0A7G6 (PDB ID: 1U98).

Function

Required for homologous recombination and the bypass of mutagenic DNA lesions by the SOS response. Catalyzes ATP-driven homologous pairing and strand exchange of DNA molecules necessary for DNA recombinational repair (Figure 38). Catalyzes the hydrolysis of ATP in the presence of single-stranded DNA, the ATP-dependent uptake of single-stranded DNA by duplex DNA, and the ATP-dependent hybridization of homologous single-stranded DNAs. The SOS response controls an apoptotic-like death (ALD) induced (in the absence of the mazE-mazF toxin-antitoxin module) in response to DNA damaging agents that is mediated by RecA and LexA.

Activity regulation

The rate of DNA-strand exchange is stimulated by RadA.

Disruption phenotype

Triple mazE-mazF-recA mutant cells no longer undergo an apoptotic-like death upon DNA damage characterized by membrane depolarization.

Induction

Induced by DNA damage, repressed by LexA. Induced in response to low temperature. Sensitive to temperature through changes in the linking number of the DNA. Induced by cold shock (42 to 15 degrees Celsius) (at protein level). 5.1-fold induced by hydroxyurea treatment (at protein level). mRNA levels are repressed in a mazE-mazF-mediated manner.

Below some examples of Pfam results are reported. All the family/domain descriptions are extracted from Pfam web site. For the complete list of families and their description, please, see Annex F.

CbtA_toxin of type IV toxin-antitoxin system

CbtA is a family of bacterial and archaeal toxins of type IV toxin-antitoxin system. Toxins from such systems in free-living bacteria inhibit cell growth by targeting essential functions of cellular metabolism. In this case the toxin inhibits cell-division leading to changes in morphology and finally lysis, by interacting with two essential cytoskeletal proteins, *FtsZ* and *MreB*. For *FtsZ* it inhibits its GTPase activity and GTP-dependent polymerisation, and for *MreB* it inhibits its ATP-dependent polymerisation. These actions of *CbtA* appear to occur simultaneously. The cognate antitoxin family is represented by PF06154 (Table 38).

Reference structure: P77692

Table 38: P77692 architecture

Source	Domain	Start	End
Pfam	<i>CbtA_toxin</i>	1	112

On the other hand, some examples of InterPro results are reported in the following lines. All the domain/signature descriptions are extracted from InterPro web site. For the complete list of domains and their description, please, see Annex G.

P64524 - Cytoskeleton-binding toxin CbtA - *Escherichia coli* (strain K12)

disruption phenotype

No visible phenotype. Deletion of 3 type IV toxin genes (*cbtA*, *ykfI*, *ypjF*) leads to a slight reduction in resistance to oxidative stress, has no effect on cell growth.

domain

The N-terminal 15 residues are required for interaction with both *FtsZ* and *MreB*, while the C-terminal 63 residues are required for interaction with *MreB*.

function

Toxic component of a type IV toxin-antitoxin (TA) system. Acts as a dual toxin inhibitor that blocks cell division and cell elongation in genetically separable interactions with *FtsZ* and *MreB*. Interacts with cytoskeletal proteins *FtsZ* and *MreB*; inhibits *FtsZ* GTP-dependent polymerization and GTPase activity as well as *MreB* ATP-dependent polymerization. Binds to both the N- and C-terminus of *FtsZ*, likely blocking its polymerization and localization, leading to blockage of cell division. Overexpression results in inhibition of growth in liquid cultures and decrease in colony formation; these effects are overcome by concomitant expression of antitoxin *CbeA* (*YeeU*). In other experiments cells swell, by 6 hours are lemon-shaped and by 24 hours those that have not lysed are spherical with diminished polar regions. Toxic effects are neutralized by cognate antitoxin *CbeA*, although there is no direct interaction between the 2 proteins. Toxic effects are also neutralized by overexpression of noncognate antitoxins *YafW* and *YpjF*.

induction

Expressed in mid-log phase at lower levels than toxin *relE*.

miscellaneous

Encoded in prophage CP4-44.

similarity

Belongs to the *CbtA/YkfI/YpjF* toxin family.

subunit

Interacts with *FtsZ*. Interacts with *MreB*.

P62555 - Toxin CcdB - *Escherichia coli* O157:H7

function

Toxic component of a type II toxin-antitoxin (TA) system, functioning in plasmid maintenance. Responsible for the post-segregational killing (PSK) of plasmid-free cells, also referred to as a plasmid addiction system. Half-life of over 2 hours. Interferes with the activity of DNA gyrase, inducing it to form a covalent *GyrA*-DNA complex that cannot be resolved, thus promoting breakage of plasmid and

chromosomal DNA. Toxicity is inhibited by labile antitoxin CcdA, which blocks the activity of CcdB; CcdA also removes bound CcdB protein from the CcdB-GyrA complex by forming a CcdA-CcdB complex, a process termed rejuvenation. Functions as a transcriptional corepressor for the ccdAB operon, repression also requires CcdA (By similarity).

similarity

Belongs to the CcdB toxin family.

subunit

Homodimer. Forms a complex with GyrA, probably a tetramer GyrA(2)CcdB(2), in which GyrA is inactive. Forms a complex with toxin CcdB; the CcdA-CcdB(2) trimer is sufficient for rejuvenation, whereas maximal operon repression occurs with CcdA(2)CcdB(2) (By similarity).

Q46995 - Antitoxin CcdA - *Escherichia coli*

function

Antitoxin component of a type II toxin-antitoxin (TA) system which inhibits the post-segregational killing (PSK) of plasmid-free cells, also referred to as a plasmid addiction system. Binds to and blocks the activity of CcdB; will also remove bound CcdB protein from the CcdB-GyrA complex by forming a CcdA-CcdB complex, a process termed rejuvenation. Functions as a transcriptional corepressor for the ccdAB operon, repression also requires CcdB (By similarity).

similarity

Belongs to the CcdA antitoxin family.

subunit

Homodimer in solution and when bound to DNA. Forms a complex with toxin CcdB; the CcdA-CcdB(2) trimer is sufficient for rejuvenation, whereas maximal operon repression occurs with CcdA(2)CcdB(2) (By similarity).

3.2. Proteins with putative toxic activity as aggregates

Some human pathologies are associated with the presence in the affected tissues of insoluble protein formations, whose components are sometimes referred to as 'toxic proteins'. In the majority of cases such formations are highly structured aggregates, known as amyloid deposits; in other cases they feature unstructured aggregates. In accordance with this evidence, amyloidosis and aggregation were explored, giving rise to the below results.

Amyloidosis

The tendency of individual proteins to aggregate into amyloid fibrils varies. Some proteins are non-amyloidogenic when present at their physiological concentration but become amyloidogenic at higher concentration. This is for instance the case of serum amyloid A, an acute phase protein whose circulating levels sustainedly increase in the course of chronic inflammation; of beta-2-microglobulin, a 99 aa polypeptide chain that, in the kidneys, is filtered by the glomerulus to be reabsorbed down the tubules, and whose circulating levels increase in patients on long-term hemodialysis; of complete immunoglobulins and/or portions thereof, which are produced in clonally homogenous form and at high concentration by plasma cells in the course of multiple myeloma (hence the term of M-proteins). Some proteins may form amyloid deposits at physiological concentrations when proteostasis become defective, such as in older age. This is for instance the case of wild-type transthyretin in senile amyloidosis and possibly of all proteins involved in neurodegenerative conditions of the elderly. Mutations in precursor genes may increase the degree of amyloidogenicity of the proteins described above and of other proteins; however, there are some reported cases of mutations that decrease the amyloidogenic potential. The most relevant example for both outcomes is transthyretin, for which more than 50 amyloidogenic, and a handful of anti-amyloidogenic mutants have been described. The quaternary homotetrameric structure of transthyretin involves the association of two monomers then the association of two dimers. The dimer-dimer interfaces define two binding sites for the protein ligands; occupancy of one or both sites by natural or artificial ligands stabilizes the assembly. Conversely, apo-transthyretin is prone to dissociation and the dissociated subunits are in turn prone to partial unfolding/misfolding, which is instrumental to aggregation into amyloid fibrils. Among the artificial ligands interfering with the above processes, tafamidis (2-(3,5-dichlorophenyl)-1,3-benzoxazole-6-

carboxylic acid) is being marketed since 2011 to delay amyloid deposition and neurological sequels in carriers of transthyretin amyloidogenetic mutations.

In a few cases amyloid fibers are made up not by individual components but by an assortment of proteins. The likely reason behind this presentation is an increased rate of cellular death, entailing the impossibility for the tissues to clear all the cellular debris. For instance, in some cutaneous forms of amyloidosis, the main components of the amyloid fibers are keratin and immunoglobulins together with serum amyloid P (SAP). Actually, the latter protein binds to all types of amyloid fibrils in a calcium-dependent, reversible manner, and stabilizes them.

The above background guided our search for proteins that are associated with amyloidosis – and that for this reason may be defined as ‘toxic’ to tissues/organisms. Our search in UniProt database was made with the string: « amyloidosis AND reviewed:yes », and retrieved 71 entries; of these, 44 are human proteins and 6 more are primate proteins (1 of *Macaca fascicularis* / Cynomolgus monkey, 4 of *Macaca mulatta* / Rhesus macaque, 1 of *Saimiri sciureus* / squirrel monkey), 4 are from hamster, 4 are murine and 2 bovine proteins. We downloaded the list with the main information for each of them: Entry; Entry name; Status; Protein names; Gene names; Organism; Length.

We reviewed the list by inspecting the content of each entry. We rated as positive hit each protein that is reported to form amyloid fibrils by itself; we divided such entries as positive in the wild-type sequence and positive in one or more mutated natural variants. While in some cases also mutagenesis has been carried out on the gene, in our classification we did not take note of mutagenesis data. We rated as negative hit each protein that is not reported to form amyloid fibrils by itself, but that is instead able to start events whose final outcome is amyloidogenesis (25 entries).

In 7 cases the term ‘amyloid’ could not be retrieved inside any field of the protein entry. Two such cases refer to parts of immunoglobulins, one to a fragment of apolipoprotein E (in a monkey) so the properties are likely to reflect what is known for the human protein of that name as a whole. In 3 cases we could find information about the direct or indirect involvement of the protein in amyloidosis referring to general scientific literature. In one case we could not find any such link.

Aggregation of human proteins

The term ‘aggregat(ion)’ is much broader than ‘amyloidosis’ and as such it is likely to yield many false positive hits. The word, in fact, is used to describe not only some pathologic events involving proteins but also a few physiologic processes involving various biological items.

With reference to physiological processes, the most systematic and specific use of ‘aggregation’ refers to platelet function: in the final step of hemostasis, activated platelets with exposed GPIIb/IIIa bind fibrinogen and do aggregate. It was then straightforward to exclude from the output all UniProt entries connected with hemostasis by including in the search string the words ‘NOT platelet’. Conversely, we could not find any other general term that could be used to omit whole groups of proteins from the search. The more than 300 entries retrieved were thus inspected one by one.

As anticipated, most of the proteins in the output turned out to be false positive hits (289/337).

In many cases ‘aggregation’ was actually found in the entry, but it referred to such biological items as organelles, e.g. mitochondria, or cells, e.g. in homotypic recognition through adhesion molecules. In some cases, along the lines intended by our search, it did refer to proteins, still ‘aggregation’ was used to describe the way their biological function is actually carried out through interaction/binding/complex formation with similar or with different macromolecules, either at baseline or as a reaction to specific stress conditions. In a number of cases the protein was described as being able to prevent the aggregation of another protein, e.g. by acting as a chaperone; in a couple of cases, to cause the aggregation of another protein.

In many cases ‘aggregation’ was found as being the hallmark of a disease and mutations/natural variants of the protein were reported to either predispose to the disease or to be its cause; however, the protein dealt with in the test entry was not reported as specifically prone to form pathological aggregates.

Only in a minority of cases, pathological aggregates of the protein itself were reported as pathological feature of a disease, and this was almost invariably connected with specific variants of the protein rather than with the wild-type form; even in several such cases, however, aggregation was associated with some but not all of the variants defined as cause of the relevant disease, so that the presence of protein aggregates *per se* seems unlikely to be the reason of the clinical findings (rather than e.g. the reduced availability/activity of the free protein). In a very large number of cases (72) 'aggregation' could not be found in any of the fields of the UniProt entry. Some of the relevant proteins were part of a complex, some were reported to bind other (macro)molecules so that it could be assumed that in these cases 'aggregation' could have been taken as equivalent to interaction/binding/complex formation/polymerization and the like. For some other proteins, however, no obvious reason for such a discrepancy was evident.

Aggregation of animal proteins

After collecting the properties of all the hits, in almost the totality of cases the connection with actual protein aggregation appears as loose as seen above for the human molecules (subunits in oligomeric proteins, protein components in supramolecular assemblies; chaperones preventing protein aggregation; proteins involved in organelle or cell aggregation; etcetera). As it may be expected, contrary to the human entries, no reference is made to pathologies or to natural mutants; in few cases it is made to mutagenesis or to an experimentally null phenotype (Table 39).

Table 39: Number of proteins forming aggregates per animal species

Common name	Genus and species	Number of proteins
<i>laboratory animals:</i>		
mouse	<i>Mus musculus</i>	151
rat	<i>Rattus norvegicus</i>	91
African clawed frog	<i>Xenopus levis</i>	19
zebrafish	<i>Danio rerio</i>	10
guinea pig	<i>Cavia porcellus</i>	7
<i>farm animals:</i>		
bovine	<i>Bos taurus</i>	56
chicken	<i>Gallus gallus</i>	19
pig	<i>Sus scrofa</i>	17
rabbit	<i>Oryctolagus cuniculus</i>	11
horse	<i>Equus caballus</i>	7
sheep	<i>Ovis aries</i>	7
goat	<i>Capra hircus</i>	1
<i>pets:</i>		
dog	<i>Canis familiaris</i>	10
hamster	<i>Mesocricetus auratus</i>	8
cat	<i>Felis catus</i>	5
<i>primates:</i>		
Cynomolgus monkey	<i>Macaca fascicularis</i>	14
Chimpanzee	<i>Pan troglodytes</i>	4
Japanese macaque	<i>Macaca fuscata fuscata</i>	2
Rhesus macaque	<i>Macaca mulatta</i>	2

As an example, PrP was captured 41 times.

In conclusion, summing up the overall evidence from this part of the work the words 'toxic' and 'toxicity' appear to be used sometimes in the scientific literature as synonyms of pathological/pathology or even of pathology-associated/pathology-associated phenomena. A similarly extended use up to improperly stretching and almost subverting the meaning occurs for aggregation when applied to proteins: as

written few lines above, the word is sometimes used as synonym of interaction, or binding, or complex formation, or polymerization.

In addition to these mainly semantic observations, a biological issue seems relevant as well: while insoluble proteins, in the form of either amyloid fibers or amorphous aggregates, are observed in a number of diseases, the link between the presence of such insoluble structure and the pathogenesis of the disorder is not always clear. Efforts to prevent the formation of amyloid was effective in preventing neurological sequels in carriers of amyloidogenic transthyretin variants but interference on beta-amyloid processing had little impact on the development of Alzheimer disease. This raises the hypothesis that, at least in some cases, amyloid formation may be just a symptom/a marker of an ongoing condition rather than its cause.

4. Quality assurance of 3D structures

4.1. Protein Data Bank

Out of the more than 6000 proteins identified with the “toxin activity” search, 738 have associated one or more experimentally-solved structure/s. On the other hand, 102 proteins have associated one or more experimentally-solved structures that were identified with the “toxin-antitoxin system” search. As described in the Materials and methods section of present report, 3D structures were carefully checked and only one experimentally-solved structure was selected for each protein. The selection criteria were based both on experimental determination of the structure and on resolution, weighting these two terms 80% and 20%, respectively.

The preparation of the structures via *in silico* methods (e.g. energy minimization, loop modelling and protonation) ensures that all the experimentally-solved related errors were corrected and that the selected structures are in an optimized geometry. In order to show the distribution of some parameters, we use a boxplot⁵ graph obtained by reporting the 5 summary numbers [minimum, 1st quartile (Q1), median, 3rd quartile (Q3), maximum] on a vertical (or horizontal) axis. As shown in Figure 39, the medians of the primary structure coverage are 89% and 100% for both Main Collection and TAS Collection, respectively. All selected 3D structures from the RCSB Protein Data Bank have a very good resolution (lower than 2.5 Å) (Figure 39), except one structure in the TAS Collection (higher than 8 Å), since in structures with resolution 3 Å or lower only the backbone of the protein chain is resolved with reasonable accuracy and the positions of the amino acid side chains are inferred based on predefined geometries. For the Main Collection, the interquartile range spans from 65% to 100%, a minimum is found around 15%, with some outliers. On the other hand, for the TAS Collection, the interquartile range is close to 100%, with many outliers.

Overall, the selected experimentally-solved structures were considered of a good quality, considering both structural determination and resolution. All the outliers in structure determination (points in the boxplot) are related to proteins with a very long primary structure and the experimentally-solved protein parts are actually involved in the toxicological mechanism. In fact, for these proteins, the structural determination was focused on domains related to protein toxic effects, excluding structurally unstable/disordered regions. For the Main Collection, the median is 2 Å and the maximum is found around 3.5 Å, without outliers. On the other hand, for the TAS Collection, the median is approx. 1.5 Å, with an outlier. The single outlier shown in the resolution boxplot (Figure 40) represents a protein for which only the alpha-carbons of the protein were resolved (no side chains could be modelled). For this protein, a 3D model was generated using the MOE Homology Modelling tool.

⁵ The box of the plot has Q1 and Q3 respectively as lower and upper ends. The median divides the box into two parts. The moustache is obtained by joining Q1 to the minimum and Q3 to the maximum. The distance between Q3 and Q1 is called interquartile range. By comparing the lengths of the two whiskers (which represent the distances between Q1 and the minimum and between Q3 and the maximum) and the heights of the two rectangles that make up the box (which represent the distances between Q1 and median and between median and Q3) information on the symmetry of the distribution is obtained: this is all the more symmetrical as the lengths of the moustache are similar to each other and the heights of the two rectangles are similar to each other. The moustache also highlights the presence of any outliers (exceptional observations).

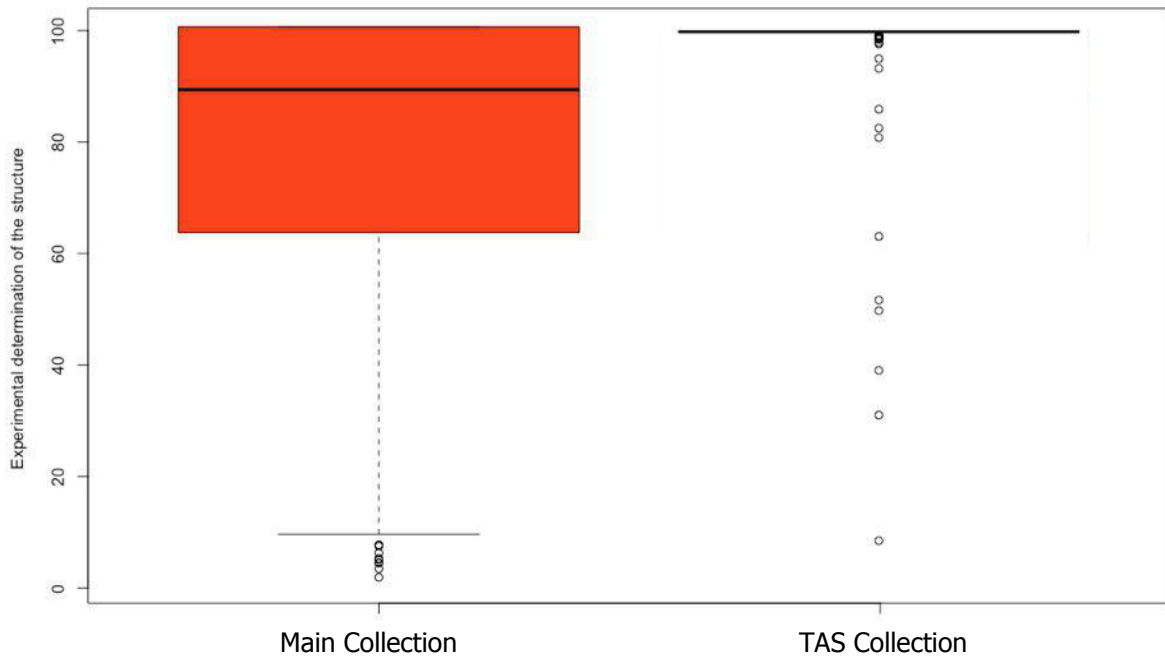


Figure 39: Boxplot of the sequence coverage of the experimentally-solved structure for both Main and TAS Collections. The lower and the upper parts of the box represent the first and the third quartile, respectively, while the black line is the median. Points are outliers.

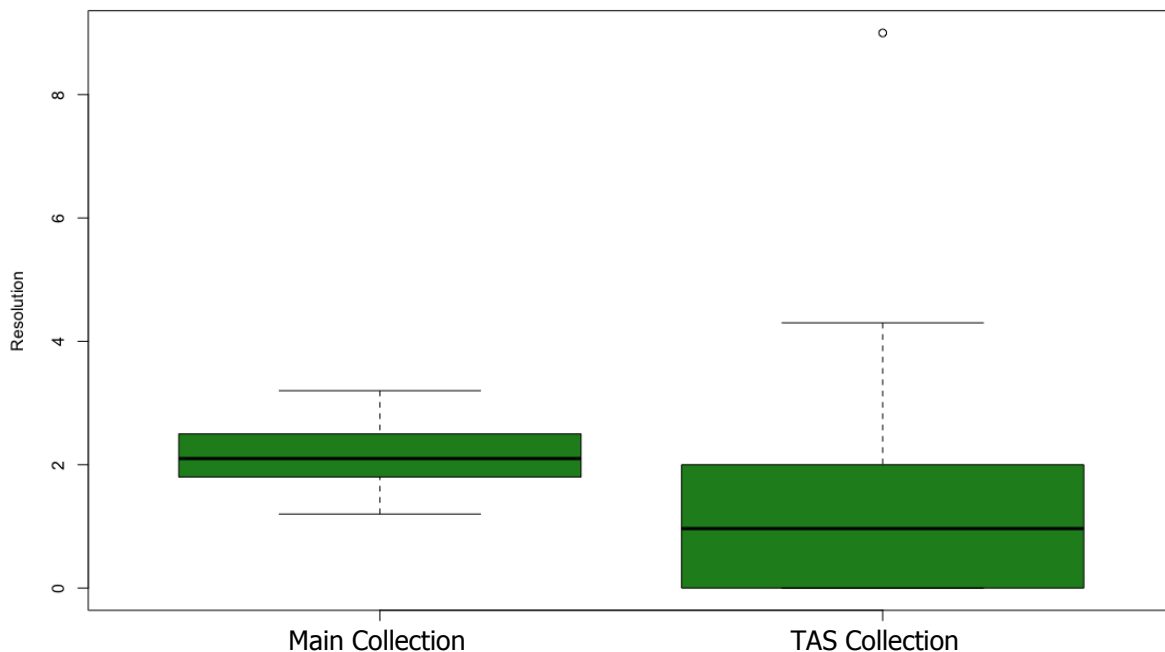


Figure 40: Boxplot of resolution of crystallographic structures for both Main and TAS Collections. The lower and the upper parts of the box represent the first and the third quartile, respectively, while the black line is the median. The point represents an outlier.

4.1.1. Models downloaded from SM repository

Swiss-Model, the protein structure homology-modelling server used to either download existing 3D model or to newly produce additional 3D models of protein toxins, has two different parameters for the model quality estimations:

- i) **GMQE** (*Global Model Quality Estimation*) combines properties from the target–template alignment and the template search method. The resulting GMQE score is expressed as a number between 0 and 1, reflecting the expected accuracy of a model built with that alignment and template and the coverage of the target. Higher numbers indicate higher reliability.
- ii) **Qmean** is a composite estimator based on different geometrical properties and provides both global (i.e. for the entire structure) and local (i.e. per residue) absolute quality estimates on the basis of one single model. The QMEAN Z-score provides an estimate of the "degree of nativeness" of the structural features observed in the model on a global scale. It indicates whether the QMEAN score of the model is comparable to what one would expect from experimental structures of similar size. QMEAN Z-scores around zero indicate good agreement between the model structure and experimental structures of similar size. Scores of -4.0 or below are an indication of models with low quality.

Both the GMQE and Qmean values were considered in the quality check of the obtained 3D models. As shown in Figures 41-44, the 3D structures downloaded from Swiss-Model repository have an acceptable GMQE value, while some of them have a low Qmean value (lower than -4.0).

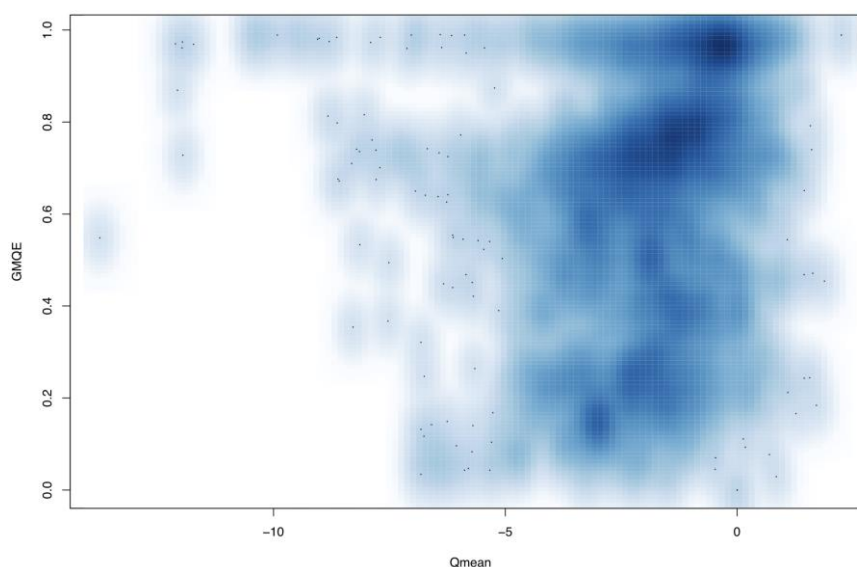


Figure 41: Scatterplot of Qmean and GMQE quality parameters for the Main Collection.

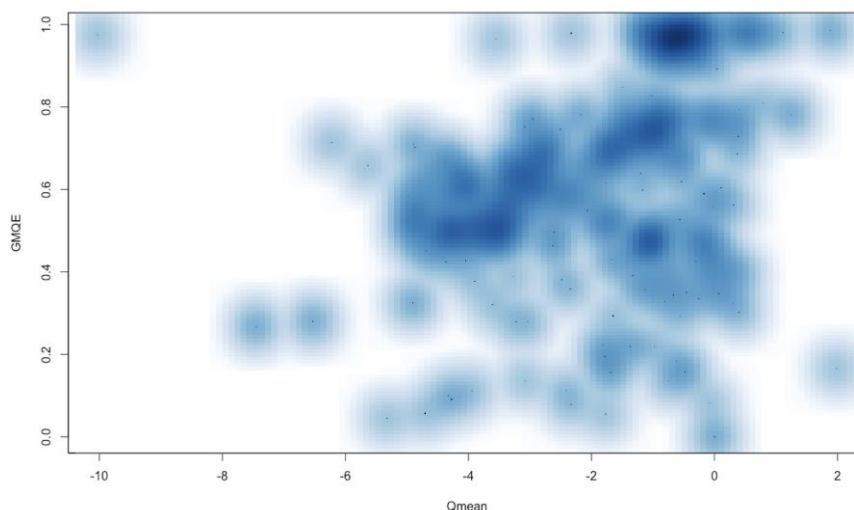


Figure 42: Scatterplot of Qmean and GMQE quality parameters for the TAS Collection.

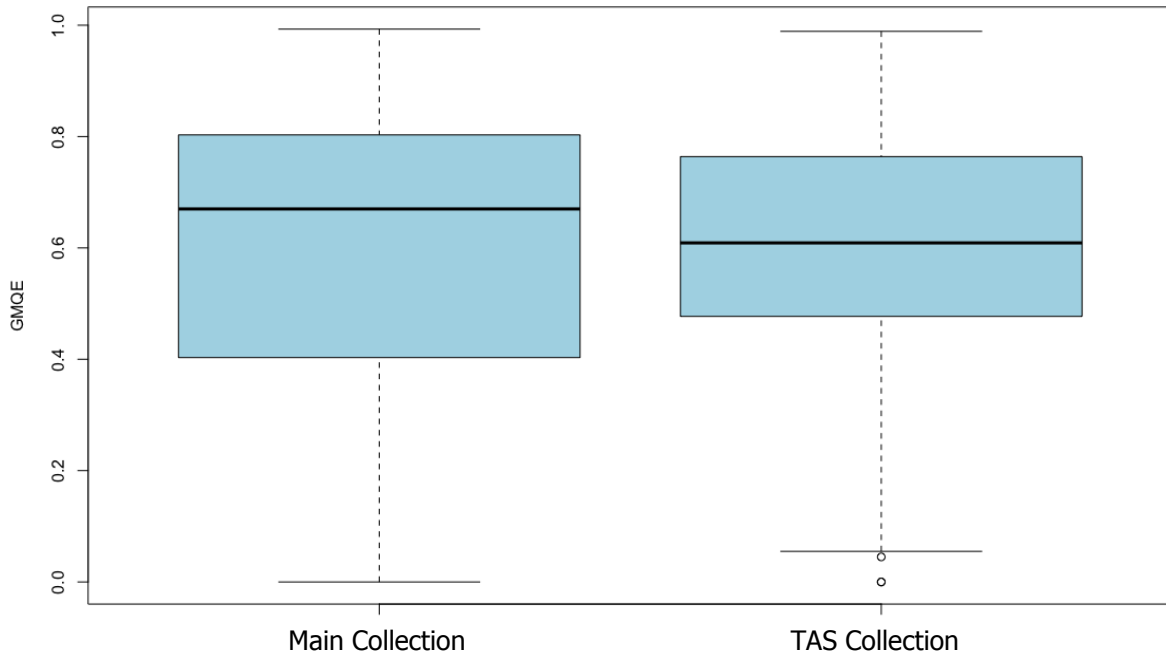


Figure 43: Boxplot of model GMQE parameter for both Main and TAS databases. The lower and the upper parts of the box represent the first and the third quartile, respectively, while the black line is the median. The point represents an outlier.

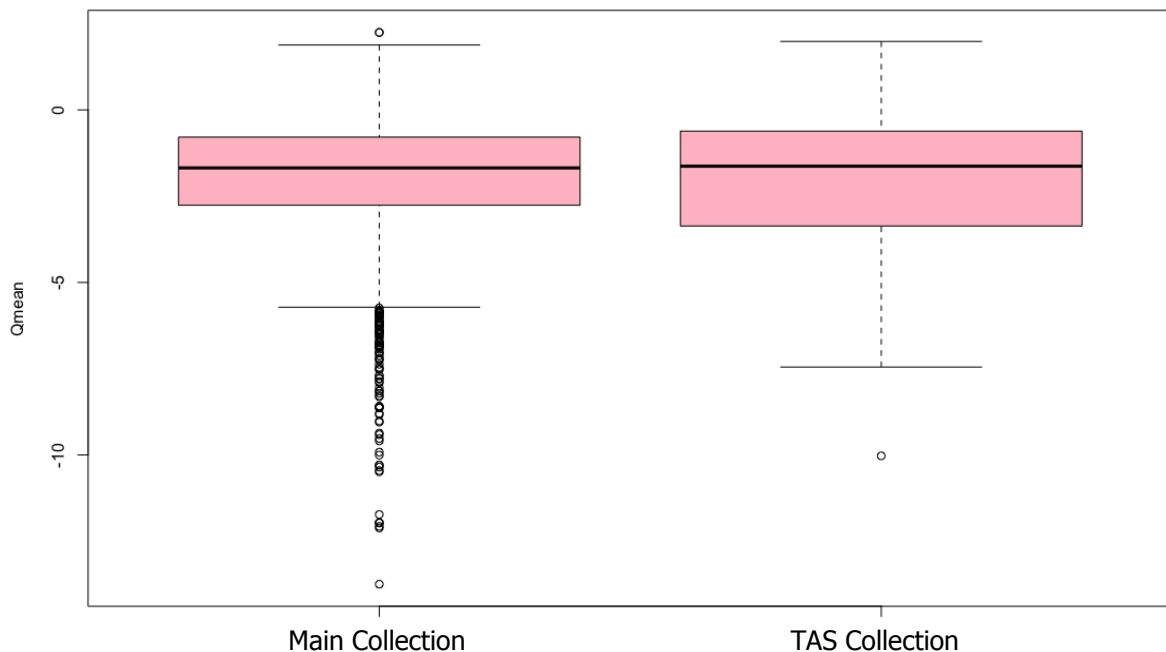


Figure 44: Boxplot of model Qmean parameter for both Main and TAS databases. The lower and the upper parts of the box represent the first and the third quartile, respectively, while the black line is the median. The point represents an outlier.

5. Evaluation of toxin prediction tools

In parallel with the *in silico* pipeline developed to reach the milestones required by EFSA, we also tested the most important freely available toxin prediction tools (Table 41). These tools, developed by academic researchers, can be used to predict if a query protein could be a toxin and were selected starting from

a web search, also evaluating the available scientific literature. Obsolete or not documented tools were discarded.

To test these tools, we selected some subsets of toxins from the Main or TAS Collection (as true positives) and some proteins known to be non-toxins (as true negatives). The true negatives were chosen among well-known proteins, not involved in toxicity and that are being studied by the UMIL group (Table 42). The tools were evaluated by computing the sensitivity, specificity and accuracy of the predictions. For some tools that predict toxicity in a specific field or for specific organisms, we also added a negative set, containing toxins that do not belong to the prediction field specific for the tool of interest (see below for further details). For each tool, we evaluated the Cooper's statistic considering the parameter values in range (0.5-0.75] as fair; (0.75-0.95] as good; (0.95-0.99] as very good and 1 as excellent.

Table 41: Selected tools

<i>Source</i>	<i>Link</i>
NTXpred	http://crdd.osdd.net/raghava/ntxpred/
BTXpred	http://crdd.osdd.net/raghava/btxpred/
KNOTTIN	http://www.dsimb.inserm.fr/KNOTTIN/
CLANTOX	http://www.clantox.cs.huji.ac.il
CONOSERVER	http://www.conoserver.org
ToxinPred	https://webs.iitd.edu.in/raghava/toxinpred/index.html

Table 42: Selected True Negative subset of well-known non-toxic proteins

<i>UniProtKB ID</i>	<i>Protein Name</i>	<i>Organism</i>
P60063	Arginine/agmatine antiporter	Escherichia coli O157:H7
Q01650	Large neutral amino acids transporter small subunit 1	Human
P16301	Phosphatidylcholine-sterol acyltransferase	Mouse
	Apolipoprotein A-I	Mouse
Q00623		
P10276	Retinoic acid receptor alpha	Human
	ATP-dependent zinc metalloprotease FTSH 7, chloroplastic	Arabidopsis thaliana
Q9SD67		
Q09QM4	Uracil nucleotide/cysteinyl leukotriene receptor	Rat
P03372	Estrogen receptor	Human
P35869	Aryl hydrocarbon receptor	Human
P04629	High affinity nerve growth factor receptor	Human
Q15233	Non-POU domain-containing octamer-binding protein	Human
O43174	Cytochrome P450 26A1	Human
Q9UKV0	Histone deacetylase 9	Human
P08684	Cytochrome P450 3A4	Human
P70041	Cerberus	Xenopus laevis
Q8MI17	Retinal dehydrogenase 1	Rabbit
P28700	Retinoic acid receptor RXR-alpha	Mouse
Q00175	Progesterone receptor	Mouse
Q8C2B3	Histone deacetylase 7	Mouse
P40394	All-trans-retinol dehydrogenase [NAD(+)] ADH7	Human

5.1. NTXpred

Link: <http://crdd.osdd.net/raghava/ntxpred/>

Neurotoxins are a broad class of exogenous chemical compounds that can adversely affect function in both developing and mature nervous tissues. Neurotoxins act on ion concentrations across the cell membrane, or synaptic communication between neurons. Neurotoxins are key players in science and medicine and are used in receptor studies, drug discovery and insecticides research.

NTXpred allows users to discriminate between neurotoxins and non-toxins and to classify neurotoxin proteins on the basis of their function using a feed-forwarded neural network (FNN) and partial recurrent neural network (RNN) with a single hidden layer.

The server can accept 2 types of primary structure format:

1. amino acids in single letter code;
2. a standard primary structure format (PIR, FASTA, EMBL...).

The user can select 4 different types of classification:

1. neurotoxin or not neurotoxin;
2. based on source;
3. based on function;
4. sub-classification of ion channel inhibitors.

The server allows the prediction on the basis of 3 different approaches:

1. amino acid composition (and length);
2. dipeptide (and length);
3. PSI-Blast.

To test the performance of NTXpred, a true positive subset (TP) and two true negative subsets (TN1 and TN2) of proteins were used. The proteins in the TP subset were randomly extracted from the Main Collection using an R script, selecting only proteins with the "Neurotoxin" keyword; the TN1 subset was created extracting proteins from the Main Collection, selecting only proteins with no associated neurotoxic effects. For TN2, protein listed in Table 42 were used. These proteins are not associated with any toxic effect. Tables 43-45 report the NTXpred results for the three used subsets.

Table 43: NTXpred TP results

<i>UniProt ID</i>	<i>NTXpred prediction (amino acid)</i>	<i>NTXpred prediction (dipeptide)</i>	<i>NTXpred prediction (PSI-BLAST)</i>
F1CGT6	Neurotoxin	Neurotoxin	Neurotoxin
P0CI04	Neurotoxin	Neurotoxin	Neurotoxin
P0C194	Neurotoxin	Neurotoxin	Neurotoxin
P0DP10	Neurotoxin	Neurotoxin	Neurotoxin
A6YR42	Neurotoxin	Neurotoxin	Neurotoxin
Q9BPC0	Neurotoxin	Neurotoxin	Neurotoxin
P85273	Neurotoxin	Neurotoxin	Neurotoxin
P45664	Neurotoxin	Neurotoxin	Neurotoxin
D2Y218	Neurotoxin	Neurotoxin	Neurotoxin
P84864	Neurotoxin	Neurotoxin	Neurotoxin
Q6T178	Neurotoxin	Neurotoxin	Neurotoxin
P58370	Neurotoxin	Neurotoxin	Neurotoxin
P0DL24	Neurotoxin	Neurotoxin	Neurotoxin
P86989	Neurotoxin	Neurotoxin	Neurotoxin
A7X3V0	Neurotoxin	Neurotoxin	Neurotoxin
Q98993	Non-toxin	Non-toxin	Non-toxin
P0CB08	Neurotoxin	Neurotoxin	Neurotoxin
Q9BPC7	Neurotoxin	Neurotoxin	Neurotoxin
P84093	Neurotoxin	Neurotoxin	Neurotoxin
P0C8E0	Neurotoxin	Neurotoxin	Neurotoxin

Table 44: NTXpred TN1 results

<i>UniProt ID</i>	<i>NTXpred prediction (amino acid)</i>	<i>NTXpred prediction (dipeptide)</i>	<i>NTXpred prediction (PSI-BLAST)</i>
Q0TV31	Non-toxin	Non-toxin	Non-toxin
P83047	Non-toxin	Non-toxin	Non-toxin

D2Y225	Non-toxin	Non-toxin	Non-toxin
P13423	Non-toxin	Non-toxin	Non-toxin
P96622	Non-toxin	Non-toxin	Non-toxin
Q07730	Non-toxin	Non-toxin	Non-toxin
P09167	Non-toxin	Non-toxin	Non-toxin
P0CE82	Non-toxin	Non-toxin	Non-toxin
COJAZ3	Non-toxin	Non-toxin	Non-toxin
E3PQQ8	Non-toxin	Non-toxin	Non-toxin
Q6STF1	Non-toxin	Non-toxin	Non-toxin
O35126	Non-toxin	Non-toxin	Non-toxin
Q0T963	Non-toxin	Non-toxin	Non-toxin
P0DP16	Non-toxin	Non-toxin	Non-toxin
P81428	Non-toxin	Non-toxin	Non-toxin
Q9PTU8	Non-toxin	Non-toxin	Non-toxin
P02879	Non-toxin	Non-toxin	Non-toxin
P69840	Non-toxin	Non-toxin	Non-toxin
P0C7B0	Non-toxin	Non-toxin	Non-toxin
G8XQX1	Non-toxin	Non-toxin	Non-toxin

Table 45: NTXpred TN2 results

<i>UniProt ID</i>	<i>NTXpred prediction (amino acid)</i>	<i>NTXpred prediction (dipeptide)</i>	<i>NTXpred prediction (PSI-BLAST)</i>
P60063	Non-toxin	Non-toxin	Non-toxin
Q01650	Non-toxin	Non-toxin	Non-toxin
P16301	Non-toxin	Non-toxin	Non-toxin
Q00623	Non-toxin	Non-toxin	Non-toxin
P10276	Non-toxin	Non-toxin	Non-toxin
Q9SD67	Non-toxin	Non-toxin	Non-toxin
Q09QM4	Non-toxin	Non-toxin	Non-toxin
P03372	Non-toxin	Non-toxin	Non-toxin
P35869	Non-toxin	Non-toxin	Non-toxin
P04629	Non-toxin	Non-toxin	Non-toxin
Q15233	Non-toxin	Non-toxin	Non-toxin
O43174	Non-toxin	Non-toxin	Non-toxin
Q9UKV0	Non-toxin	Non-toxin	Non-toxin
P08684	Non-toxin	Non-toxin	Non-toxin
P70041	Non-toxin	Non-toxin	Non-toxin
Q8MI17	Non-toxin	Non-toxin	Non-toxin
P28700	Non-toxin	Non-toxin	Non-toxin
Q00175	Non-toxin	Non-toxin	Non-toxin
Q8C2B3	Non-toxin	Non-toxin	Non-toxin
P40394	Non-toxin	Non-toxin	Non-toxin

The web tool output is very simple, and easy to interpret. For each query, it indicates if the test protein is predicted to be a neurotoxin or not (Figures 45-46). For those proteins deemed as possible neurotoxins, NTXpred also provides a prediction of their cellular target and molecular function. NTXpred has very good sensitivity and excellent specificity and accuracy (Table 44). Based on our analysis, NTXpred appears to be a reliable tool in predicting neurotoxins.

Predicted protein
NEUROTOXIN
Predicted Function (target and action)
Block ion Channels

Figure 45: Extract of a positive results from the NTXpred webpage, using a sequence from TP dataset.

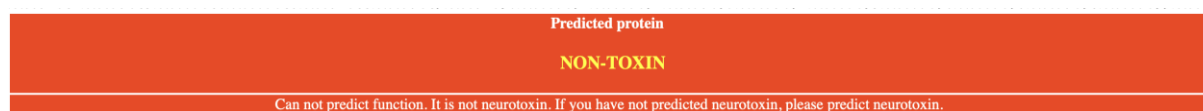


Figure 46: Extract of a negative results from the NTXpred webpage, using a sequence from TN2 dataset.

Table 46: NTXpred Cooper' statistics (values between 0 and 1)

<i>Dataset</i>	<i>Method</i>	<i>Sensitivity</i>	<i>Specificity</i>	<i>Accuracy</i>
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

5.2. BTXpred

Link: <http://crdd.osdd.net/raghava/btxpred/>

The BTXpred server accepts an amino acid sequence as input with the aim of predicting whether the queried protein is a bacterial toxin. If the protein is deemed as a bacterial toxin the tool also predicts its potential toxic function. The server also allows users to classify bacterial toxins into:

- endotoxins (constituent of the cell wall released into host tissues when bacteria die)
- exotoxins (soluble substance secreted into host tissues).

Exotoxins are then classified into 7 different function groups depending upon their molecular targets:

1. adenylate cyclase activation;
2. guanylate cyclase activation;
3. food poisoning;
4. neurotoxins;
5. macrophage cytotoxin;
6. vacuolating cytotoxin;
7. thiol-activated cytotoxin.

The BTXpred server allows the submission of a sequence in any of the standard formats.

The server allows 3 types of prediction:

1. bacterial toxin or non-toxin;
2. types of toxin – endotoxin or exotoxin;
3. function of exotoxins.

Moreover, the server allows the prediction on the basis of three different approaches:

1. SVM (Support Vector Machine, for Toxin and types of toxin);
2. PSI-Blast (for Toxin, Types of Toxin, and functions of exotoxins);
3. HMM (Hidden Markov Model, only for function of exotoxins).

To test the performance of BTXpred, a true positive subset (TP) and two true negative subsets (TN1 and TN2) of proteins were used. The proteins in the TP subset were randomly extracted from the TAS Collection using an R script; the TN1 subset was created extracting random proteins from the Main Collection. For TN2, proteins listed in Table 42 were used. These proteins have not associated with any toxic effect. Tables 47-49 report the BTXpred results for the three used subsets.

Table 47: BTXpred TP results

UniProt ID	BTXpred prediction (amino acid composition – SVM)	BTXpred prediction (dipeptide composition – SVM)	Organism (Control)
Q47156	Non-toxin	Non-toxin	<i>Escherichia coli</i> (strain K12)
P76364	Bacterial toxin	Non-toxin	<i>Escherichia coli</i> (strain K12)
Q49Z22	Bacterial toxin	Bacterial toxin	<i>Staphylococcus saprophyticus subsp. saprophyticus</i>
P9WIH6	Bacterial toxin	Non-toxin	<i>Mycobacterium tuberculosis</i> (strain CDC 1551 / Oshkosh)
P11519	Bacterial toxin	Non-toxin	<i>Escherichia coli</i> (strain K12)
O06777	Bacterial toxin	Non-toxin	<i>Mycobacterium tuberculosis</i> (strain ATCC 25618 / H37Rv)
A1JUB1	Non-toxin	Bacterial toxin	<i>Yersinia enterocolitica</i> serotype O:8 / biotype 1B
Q8U2Q8	Bacterial toxin	Non-toxin	<i>Pyrococcus furiosus</i>
P9WII4	Non-toxin	Bacterial toxin	<i>Mycobacterium tuberculosis</i> (strain CDC 1551 / Oshkosh)
L0TGFO	Non-toxin	Non-toxin	<i>Mycobacterium tuberculosis</i> (strain ATCC 25618 / H37Rv)
P9WFA4	Bacterial toxin	Non-toxin	<i>Mycobacterium tuberculosis</i> (strain CDC 1551 / Oshkosh)
Q06259	Bacterial toxin	Bacterial toxin	<i>Escherichia phage P1</i>
Q52042	Bacterial toxin	Bacterial toxin	<i>Escherichia coli</i>
P22995	Bacterial toxin	Non-toxin	<i>Escherichia coli</i>
Q54944	Non-toxin	Non-toxin	<i>Streptococcus pyogenes</i>
P9WFB9	Non-toxin	Non-toxin	<i>Mycobacterium tuberculosis</i> (strain ATCC 25618 / H37Rv)
P64595	Non-toxin	Non-toxin	<i>Escherichia coli</i> O157:H7
Q46953	Non-toxin	Non-toxin	<i>Escherichia coli</i> (strain K12)
Q38278	Bacterial toxin	Bacterial toxin	<i>Lactococcus phage c2</i>
P96621	bacterial toxin	Non-toxin	<i>Bacillus subtilis</i> (strain 168)

Table 48: BTXpred TN1 results

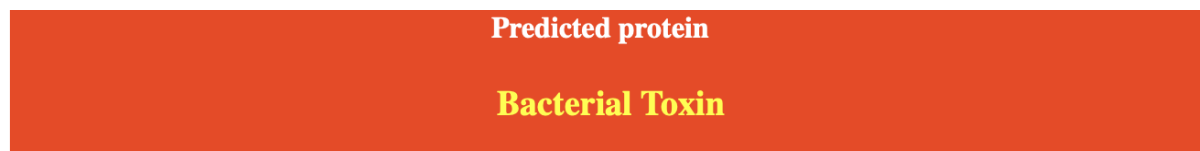
UniProt ID	BTXpred prediction (Amino acid composition - SVM)	BTXpred prediction (dipeptide composition - SVM)	Organism (Control)
U5L3M7	Bacterial toxin	Bacterial toxin	<i>Amanita exitialis</i>
F5C3U4	Non-toxin	Non-toxin	<i>Conus marmoreus</i>
C0JB52	Bacterial toxin	Bacterial toxin	<i>Sicarius patagonicus</i>
A6YR37	Bacterial toxin	Bacterial toxin	<i>Californiconus californicus</i>
C0JB68	Bacterial toxin	Non-toxin	<i>Sicarius patagonicus</i>
P01519	Bacterial toxin	Bacterial toxin	<i>Conus geographus</i>
E0SDG8	Bacterial toxin	Bacterial toxin	<i>Dickeya dadantii</i> (strain 3937)
P86470	Bacterial toxin	Non-toxin	<i>Bunodosoma caissarum</i>
P86544	Bacterial toxin	Bacterial toxin	<i>Naja naja</i>
P0CH14	Bacterial toxin	Non-toxin	<i>Conus marmoreus</i>
A6YR23	Bacterial toxin	Bacterial toxin	<i>Californiconus californicus</i>
P28878	Bacterial toxin	Bacterial toxin	<i>Conus striatus</i>
C0HKZ8	Bacterial toxin	Bacterial toxin	<i>Walterinnesia aegyptia</i>
Q7T2I1	Bacterial toxin	Non-toxin	<i>Laticauda laticaudata</i>
D2Y3T8	Bacterial toxin	Bacterial toxin	<i>Californiconus californicus</i>
C0HL53	Bacterial toxin	Bacterial toxin	<i>Heteractis crispa</i>
P0DMY5	Bacterial toxin	Non-toxin	<i>Anemonia viridis</i>
P0C6S2	Bacterial toxin	Bacterial toxin	<i>Conus cancellatus</i>
C0JB27	Bacterial toxin	Bacterial toxin	<i>Loxosceles laeta</i>

COJB36	Bacterial toxin	Bacterial toxin	<i>Sicarius peruensis</i>
---------------	-----------------	-----------------	---------------------------

Table 49: BTXpred TN2 results

UniProt ID	BTXpred prediction (Amino acid composition - SVM)	BTXpred prediction (dipeptide composition - SVM)
P60063	Non-toxin	Non-toxin
Q01650	Non-toxin	Non-toxin
P16301	Non-toxin	Non-toxin
Q00623	Non-toxin	Non-toxin
P10276	Non-toxin	Non-toxin
Q9SD67	Non-toxin	Non-toxin
Q09QM4	Non-toxin	Non-toxin
P03372	Non-toxin	Non-toxin
P35869	Non-toxin	Non-toxin
P04629	Non-toxin	Non-toxin
Q15233	Non-toxin	Non-toxin
O43174	Non-toxin	Non-toxin
Q9UKV0	Non-toxin	Non-toxin
P08684	Non-toxin	Non-toxin
P70041	Non-toxin	Non-toxin
Q8MI17	Non-toxin	Non-toxin
P28700	Non-toxin	Non-toxin
Q00175	Non-toxin	Non-toxin
Q8C2B3	Non-toxin	Non-toxin
P40394	Non-toxin	Non-toxin

The web tool produces the same output as NTXpred (Figures 47, 48). For each query, it warns if the test protein is a bacterial toxin or not. Differently from NTXpred, BTXpred does not provide predictions for cellular target(s) or molecular function. Sensitivity, specificity and accuracy show that BTXpred is able to reliably predict non-toxic proteins (Table 50) but has a reduced ability in discriminating between animal and bacterial toxins. Its sensitivity is fair, and accuracy using the TN1 dataset is unacceptably low (<0.5), using both amino acid- and dipeptide-based methods. In fact, the accuracy for TP+TN1 databases is lower than the probability to predict if a toxin is bacterial toxin using an unbiased coin.

**Figure 47:** Extract of a positive results from the BTXpred webpage, using a sequence from TP dataset.**Figure 48:** Extract of a negative results from the BTXpred webpage, using a sequence from TN2 dataset.**Table 50:** BTXpred Cooper' statistics (values between 0 and 1)

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

5.3. Knottin

Link: <http://www.dsimb.inserm.fr/KNOTTIN/>

Knottins (sometimes referred as “Inhibitor Cysteine Knots”) are small disulphide-rich proteins containing at least 3 disulphide bridges and characterized by a *disulphide through disulphide “knot”*, formed when one disulphide bridge crosses the macrocycle formed by the two other disulphides and the interconnecting backbone. The structural family of knottins has the disulphide between cysteines III and VI going through disulphides I-IV and II-V (<http://www.dsimb.inserm.fr/KNOTTIN/>) (Figure 49).

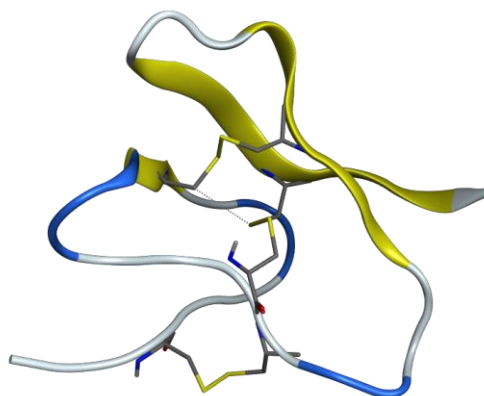


Figure 49: Crystal structure of asteropsin A from marine sponge *Asteropus* sp. (PDB ID: 3Q8J); the three disulphides bridges are highlighted.

The connection between cysteines is very important: some small disulphide-rich proteins may resemble knottins, but are not considered as knottins because they do not contain the “knot”; some other types of proteins (for example the growth factor cysteine knots) may contain a knot but the connectivity is different, so they belong to a distinct structural family (Postic et al., 2018).

The Knottin database is a relational database that stores information on known structures and sequences extracted from the Protein Data Bank and the UniProtKB. The section “Tools” allows users to select if a test protein is predicted to be a knottin based on its amino acid sequence (KNOTER1D) or its 3D structure (KNOTER3D). KNOTER3D was not assessed for consistency with other database tests. The text below refers only to KNOTER1D. KNOTER1D works as follow:

- i) the UniProtKB is filtered for cysteine-rich proteins;
- ii) each query sequence (seq1) is compared to all known knottin sequences (knot_seq2) using a similarity score based on:
 - the BLAST P-value for the seq1/knot_seq2 comparison;
 - the number of conserved cysteines of the knot when aligning seq1 onto the multiple alignment of the family of knot_seq2;
 - the compatibility of intercysteine loop lengths with known knottin loop lengths;
 - the taxonomic proximity between seq1 and knot_seq2;
 - the similarity between the seq1 function and knottin functions;
 - the similarity between seq1 keywords and knottin keywords, and/or dissimilarity between seq1 keywords and non-knottin keywords.

The query sequence is analysed using BLAST to compare the sequence against the Knottin database. The results of the analysis of the server provide the user with the indication of whether the query sequence is predicted to be a knottin, a putative knottin or not a knottin, on the basis of similarity score (Figure 50-51).

2 lines in sequence INPUT:
>mytest
DWECLPLHSSCDNDVCCKNHCHCPYSNVSKLEKWLPEWAKIPDALKRCSQRNDKDGKINTCDKYKN

The results below are predictions and should be used carefully.

Knoter1D said: **Query mytest is a knottin**

[Click here for complete Knoter1D results](#)


Results	Collier de Perles
<p>Cysteine I : Cysteine II : Cysteine III: Cysteine IV : Cysteine V : Cysteine VI :</p> <p><i>Cysteine numbering</i></p> <p>Cys IV standardized number = 61 deduced from the closest knottin family: <i>Spider</i> Nomenclature: (-1)-1.-1(-1)-1</p>	

Figure 50: Extract of some positive results from the Knottin webpage, using a sequence from TP dataset.

5 lines in sequence INPUT:
>mytest
IALILVCWSVLSQAAQTDVEGRADKRRPIWIMGHMVNAIAQIDFVNLGANSIETDVSFDDNANPEYTYHGVPCDCGRSCLK
WENFNDFLKLRSATTPGNAKYQAKLILVFDLKTGSLYDNQANEAGKKLAKNLLKHYWNNNGNGGRAYIVLSIPDLNHYP
LIKGFKDQLTHDGHPELMDKVGHDGFGSNGDAIGDVGNGAYKKAGISGHVWQSDGITNCLLRGLDRVKQAIANRDSNGNGFINK
VYYWTVDKRATTRDALDAGVDGVMNTNYPDVITDVLNESAYKNKFRVASYEDNPWETFKK

The results below are predictions and should be used carefully.

Knoter1D said: **Query mytest is not a knottin**

[Click here for complete Knoter1D results](#)

The sequence is probably not a knottin: No action taken.

Figure 51: Extract of some negative results from the Knottin webpage, using a sequence from TN dataset.

To test the performance of KNOTTIN, a true positive subset (TP) and two true negative subsets (TN1 and TN2) of proteins were used. The proteins in the TP subset were randomly extracted from the Main Collection using an R script, selecting only proteins with the “Knottin” keyword; the TN1 subset was created extracting proteins from the Main Collection, selecting only proteins that are not knottins. For TN2, proteins listed in Table 42 were used. These proteins are not associated with any toxic effect. Tables 51-53 report the KNOTTIN results for the three used subsets.

Table 51: KNOTTIN TP results

<i>UniProt ID</i>	<i>KNOTTIN prediction</i>
P83257	Putative knottin
POC8M0	Knottin
POC2L8	Putative knottin

P28881	putative knottin
P85264	Not a knottin
P0DL48	Not a knottin
D2Y1Y6	knottin
W4VSI3	Putative knottin
D2Y2P8	Knottin
B1P1F0	Knottin
P61789	Putative knottin
B3FIS3	Knottin
D2Y2I9	Not a knottin
P83400	Knottin
Q5Y4W2	Knottin
D2Y1Z6	Knottin
P37045	Knottin
D2Y267	Knottin
P0CH13	Putative knottin
Q9BP95	Putative knottin

Table 52: KNOTTIN TN1 results

<i>UniProt ID</i>	<i>KNOTTIN prediction</i>
P08434	is not a knottin
P0DJP8	is not a knottin
A8S6A4	is not a knottin
P59938	is not a knottin
P89059	is not a knottin
Q9DEZ9	is not a knottin
P0C201	knottin
B2KKV7	is not a knottin
P0DPU4	is not a knottin
C0HJI9	is not a knottin
D2Y2M5	knottin
Q9YGC7	is not a knottin
O06014	is not a knottin
E5AJX2	is not a knottin
Q3SAF4	is not a knottin
Q3HXY4	is not a knottin
D2Y2D4	is not a knottin
P0C828	is not a knottin
A0S865	is not a knottin
B3FIS5	knottin

Table 53: KNOTTIN TN2 results

<i>UniProt ID</i>	<i>KNOTTIN prediction</i>
P60063	is not a knottin
Q01650	is not a knottin
P16301	is not a knottin
Q00623	is not a knottin
P10276	is not a knottin
Q9SD67	is not a knottin
Q09QM4	is not a knottin
P03372	is not a knottin
P35869	is not a knottin
P04629	is not a knottin
Q15233	is not a knottin
O43174	is not a knottin

Q9UKV0	is not a knottin
P08684	is not a knottin
P70041	is not a knottin
Q8MI17	is not a knottin
P28700	is not a knottin
Q00175	is not a knottin
Q8C2B3	is not a knottin
P40394	is not a knottin

We tested only the 1D prediction, in order to directly compare KNOTTIN's ability to predict toxicity using the primary structure. Sensitivity, specificity and accuracy calculations show that KNOTTIN is a very good predictor in discriminating between nontoxic proteins and knottins when taking the more conservative approach (putative knottin = knottin), while it is less reliable if the less conservative approach (putative knottin = non-knottin) is taken (Table 54). Overall, accuracy is good. The user interface is very user-friendly and also the results are simple to interpret.

Table 54: KNOTTIN Cooper' statistics (values between 0 and 1)

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

5.4. CLANTOX

Link: <http://www.clantox.cs.huji.ac.il/>

ClanTox is a classifier of animal toxins. It is based on the extraction of sequence-driven functions from the primary protein sequence followed by the application of a classification system. This tool predicts whether a given sequence is similar to a toxin.

Queries in ClanTox are first transformed into a vector on the basis of various features extracted from the sequence and then analysed by ten sub-classifiers using specific algorithms. As a result, each classifier will produce a slightly different prediction:

- i) class P3 ("Toxin-like")
- ii) class P2 ("Probably toxin-like")
- iii) class P1 ("Possibly toxin-like")
- iv) class N ("Probably not toxin-like")

The labels P1, P2 and P3 are considered positive (i.e. either toxin or toxin-like) while the label N is considered negative (non-toxins). The mean score and standard deviation correspond to the average and standard deviation of all ten predictions.

To test the performance of Clantox, a true positive subset (TP) and two true negative subsets (TN1 and TN2) of proteins were used. The proteins in the TP subset were randomly extracted from the Main Collection using an R script, verifying that all the selected entries are animal proteins; the TN1 subset was created extracting proteins from the TAS Collection. For TN2, proteins listed in Table 42 were used. These proteins are not associated with any toxic effect. Table 55-57 reports the Clantox results for the three used subsets.

Table 55: Clantox TP results

UniProt ID	Clantox prediction	Organism
P0C1U0	Toxin-like	<i>Conus consors</i>
P50984	Probably toxin-like	<i>Conus pennaceus</i>
C0JAV6	Probably not toxin-like	<i>Loxosceles apachea</i>

P0DMA1	Toxin-like	<i>Conus pictus</i>
D0PX84	Possibly toxin-like	<i>Conus imperialis</i>
Q8UW25	Probably not toxin-like	<i>Hydrophis hardwickii</i>
P0DMY1	Toxin-like	<i>Anemonia viridis</i>
COJB45	Probably not toxin-like	<i>Loxosceles spinulosa</i>
P84823	Probably not toxin-like	<i>Ascapus truei</i>
A0A2I6EDL6	Possibly toxin-like	<i>Conus regius</i>
P0DM28	Probably not toxin-like	<i>Conus spurius</i>
P0C7J9	Probably not toxin-like	<i>Crotalus adamanteus</i>
P0DPM0	Possibly toxin-like	<i>Conus stercusmuscarum</i>
Q9BPH3	Possibly toxin-like	<i>Conus arenatus</i>
P01386	Toxin-like	<i>Ophiophagus hannah</i>
A0A1P8NVT3	Possibly toxin-like	<i>Conus frigidus</i>
Q1W694	Probably not toxin-like	<i>Loxosceles intermedia</i>
P86721	Probably not toxin-like	<i>Bothrops atrox</i>
P0C8V4	Probably toxin-like	<i>Conus ventricosus</i>
Q86DU6	?	<i>Conus striatus</i>

Table 56: Clantox TN1 results

UniProt ID	Clantox prediction	Organism
Q47156	Probably not toxin-like	<i>Escherichia coli</i> (strain K12)
P76364	Probably not toxin-like	<i>Escherichia coli</i> (strain K12)
Q49Z22	Probably not toxin-like	<i>Staphylococcus saprophyticus</i>
P9WIH6	Probably not toxin-like	<i>Mycobacterium tuberculosis</i>
P11519	Probably not toxin-like	<i>Escherichia coli</i> (strain K12)
O06777	Probably not toxin-like	<i>Mycobacterium tuberculosis</i>
A1JUB1	Probably not toxin-like	<i>Yersinia enterocolitica</i> serotype O:8
Q8U2Q8	Probably not toxin-like	<i>Pyrococcus furiosus</i>
P9WII4	Probably not toxin-like	<i>Mycobacterium tuberculosis</i>
LOTGF0	Probably not toxin-like	<i>Mycobacterium tuberculosis</i>
P9WFA4	Probably not toxin-like	<i>Mycobacterium tuberculosis</i>
Q06259	Probably not toxin-like	<i>Escherichia phage P1</i>
Q52042	Probably not toxin-like	<i>Escherichia coli</i>
P22995	Probably not toxin-like	<i>Escherichia coli</i>
Q54944	Probably not toxin-like	<i>Streptococcus pyogenes</i>
P9WFB9	Probably not toxin-like	<i>Mycobacterium tuberculosis</i>
P64595	Probably not toxin-like	<i>Escherichia coli</i> O157:H7
Q46953	Probably not toxin-like	<i>Escherichia coli</i> (strain K12)
Q38278	Probably not toxin-like	<i>Lactococcus phage c2</i>
P96621	?	<i>Bacillus subtilis</i> (strain 168)

Table 57: Clantox TN2 results

UniProt ID	Clantox prediction
P60063	Probably not toxin-like
Q01650	Probably not toxin-like
P16301	Probably not toxin-like
Q00623	Probably not toxin-like
P10276	Probably not toxin-like
Q9SD67	Probably not toxin-like
Q09QM4	Probably not toxin-like
P03372	Probably not toxin-like
P35869	Probably not toxin-like
P04629	Probably not toxin-like
Q15233	Probably not toxin-like
O43174	Probably not toxin-like

Q9UKV0	Probably not toxin-like
P08684	Probably not toxin-like
P70041	Probably not toxin-like
Q8MI17	Probably not toxin-like
P28700	Probably not toxin-like
Q00175	Probably not toxin-like
Q8C2B3	Probably not toxin-like
P40394	?

During the test, the last sequence of the dataset was not read by the web tool (prediction = "?").

The web tool output is very simple and easy to interpret. ClanTox categorizes proteins into 4 different levels of prediction. For each query, it indicates if it is an animal toxin or not (Figures 52-53). To evaluate the results, we followed an "increasing conservatism" approach as follows:

- less conservative approach: only the proteins classified as P3 ("Toxin-like") are considered as toxins;
- conservative approach: only the proteins classified as P3 and P2 ("Probably toxin-like") are considered as toxins;
- more conservative approach: the proteins classified as P3 ("Toxin-like"), P2 ("Probably toxin-like") or P1 ("Possibly toxin-like") are considered as toxins.

ClanTox has acceptable sensitivity and a good accuracy in the more conservative and conservative approaches. In the less conservative approach, it has an unacceptable sensitivity (<0.5). Specificity is excellent in all the three approaches (Table 58).

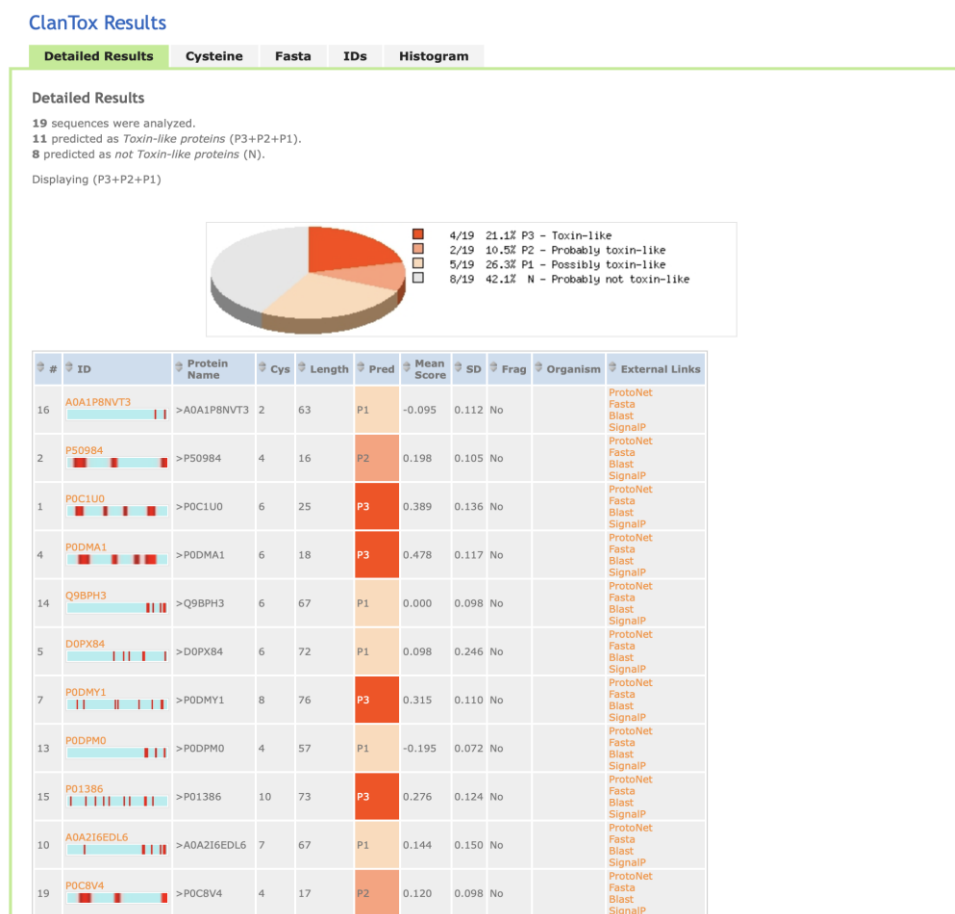


Figure 52: Extract of some positive results from the ClanTox webpage, using the TN dataset.

ClanTox Results



Figure 53: Extract of some positive results from the ClanTox webpage, using the TN dataset.

Table 58: Clantox Cooper' statistics (values between 0 and 1)

<i>Dataset</i>	<i>Method</i>	<i>Sensitivity</i>	<i>Specificity</i>	<i>Accuracy</i>
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

5.5. ConoServer

Link: <http://www.conoserver.org/>

ConoServer is a database containing information on toxins found in the *Conidae* family (cone snails) of predatory sea snails; these toxins are known as conotoxins or conopeptides. Conotoxins are 10 to 30 amino acid long and usually contain one or more disulphide bonds. The evidence for their function is not clear however it appears that many of them modulate the activity of ion channels via mechanisms of action that seem to vary (Kaas et al., 2012). Because of their high specificity and affinity towards human ion channels, receptors and transporters of the nervous system, conopeptides are of significant importance to medical research: their features make them both an interesting tool for physiology studies on neuroreceptors and promising drug leads (Kaas et al., 2012). ConoServer contains information on protein sequences, nucleic acid sequences and structural information of conopeptides. ConoServer's data are first collected from the peer-reviewed literature and from publicly available databases, including UniProtKB/Swiss-Prot, NCBI nucleotide (nt), and the World-Wide Protein Data Bank.

To test the performance of ConoServer, a true positive subset (TP) and two true negative subsets (TN1 and TN2) of proteins were used. The proteins in the TP subset were randomly extracted from the Main Collection using an R script, verifying that all the selected entries are conotoxins; the TN1 subset was created extracting non-conotoxins proteins from the Main Collection. For TN2, proteins listed in Table 42 were used. These proteins are not associated with any toxic effect. Tables 59-61 report the ConoServer results for the three used subsets.

Table 59: ConoServer TP results

<i>UniProt ID</i>	<i>ConoServer prediction</i>
Q5K0D6	Conotoxin
Q9BPI5	Conotoxin
Q9BP84	Conotoxin
Q9XZL5	Conotoxin
Q9BP68	Conotoxin
Q9UAB2	Conotoxin
P0C831	Conotoxin
P0C8V8	Conotoxin
Q5EHP2	Conotoxin
Q9BPB9	Conotoxin
P0DJC7	Conotoxin
P0C1N1	Conotoxin
Q9BP80	Conotoxin
D6C4L7	Conotoxin
Q2I2R3	Conotoxin
P69751	Conotoxin
C8CK74	Conotoxin
Q9BPC1	Conotoxin
Q2I2Q2	Conotoxin
Q9N6A4	Conotoxin

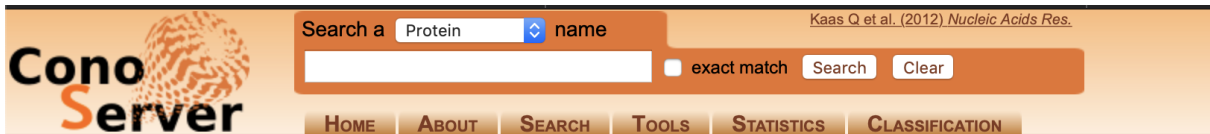
Table 60: ConoServer TN1 results

<i>UniProt ID</i>	<i>ConoServer prediction</i>
P82683	No
C0JB41	No
P81428	No
P13495	No
P0A376	No
Q9PVF0	No
C9X4K7	No
Q45718	No
Q98957	No
P43445	No
A7X3X3	No
F8S0Y4	No
P59867	No
C5H5D1	No
P19321	No
P82981	No
P0A377	No
P00629	No
D2Y292	No
P15228	No

Table 61: ConoServer TN2 results

<i>UniProt ID</i>	<i>ConServer prediction</i>
P60063	No
Q01650	No
P16301	No
Q00623	No
P10276	No
Q9SD67	No
Q09QM4	No

P03372	No
P35869	No
P04629	No
Q15233	No
O43174	No
Q9UKV0	No
P08684	No
P70041	No
Q8MI17	No
P28700	No
Q00175	No
Q8C2B3	No
P40394	No



Search a name exact match

HOME | ABOUT | SEARCH | TOOLS | STATISTICS | CLASSIFICATION

SEQUENCE

Protein List

Your search: subsequence =

'MKLTCVVIVAVLLLACQLITADDSRGTQKHRSLRSTTKVSKATDCIEAGNYCGPTVMKICCGFCSPYSKICMNYPKN'

Found 1 entry.

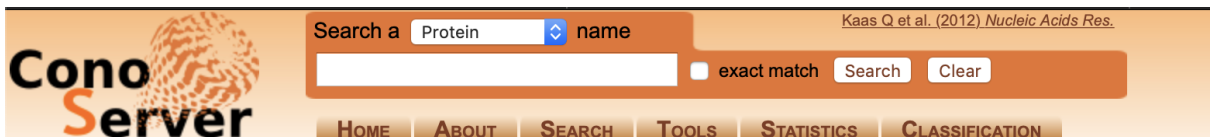
Select / Deselect All

FASTA

Format

ID	Name	Alternative name(s)	Conopeptide class	Gene superfamily	Organism
<input type="checkbox"/> P01061	SO4_precursor		conotoxin	O1 superfamily	Conus striatus

Figure 54: Extract of some positive results from the ConoServer webpage.



Search a name exact match

HOME | ABOUT | SEARCH | TOOLS | STATISTICS | CLASSIFICATION

SEQUENCE

Protein List

Your search: subsequence =

'MNAVMDSRGAWVSCFLILGLVFGATVKAETKFSYERLRLRVTHQTTGDEYFRFITLLRDYVSSGFSNEIPLLRQSTIPVSDAQI'

Found 0 entries.

Select / Deselect All

FASTA

Format

Figure 55: Extract of some negative results from the ConoServer webpage.

ConoServer accepts a single sequence as input. The produced output is very simple, and easy to interpret (Figures 54-55) and it shows outstanding accuracy in discriminating between conotoxins and non-conotoxins (Table 62). The Cooper's statistic is excellent.

Table 62: ConoServer Cooper' statistics (values are between 0 and 1)

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

5.6. ToxinPred

Link: <https://webs.iitd.edu.in/raghava/toxinpred/index.html>

ToxinPred is an *in silico* tool to predict toxicity of peptides/proteins starting from their primary structure (Figures 56, 57), also identifying specific toxic regions in proteins. ToxinPred was trained using a database created by extracting small toxins proteins/peptides from different databases including ATDB, Arachno-Server, Conoserver, DBETH, BTXpred, NTXpred and SwissProt. To build prediction models, ToxinPred uses Support Vector Machine (SVM), based on a freely available machine learning method; its performance was evaluated on independent datasets. The identification of possible motifs in toxins was based on MEME (Multiple Em for Motif Elicitation): this tool was used to assign unknown sequences as toxic or non-toxic by scanning known toxicity-associated motifs, using another application of the MEME suite called Motif Alignment & Search Tool (MAST). The hybrid approach, which combines motif identification with SVM output, offers an extra advantage for a more biologically reliable prediction of toxic peptides.

The true positive subset (TP) used to test the performance of ToxinPred was randomly extracted from Main Collection, using an R script. The true negative subset (TN) was composed by proteins listed in Table 42. The output is very simple, and easy to interpret. ToxinPred accepts a single sequence as input and shows a good accuracy in discriminating between toxins and non-toxins, a good specificity and a fair sensitivity (Tables 63-65).

Table 63: ToxinPred TP results

UniProt ID	ToxinPred prediction
Q7LZ09	Non-Toxin
P0C8B6	Non-Toxin
P85011	Toxin
B3EWW9	Non-Toxin
P0C6S6	Non-Toxin
P0C7K1	Toxin
A0A023UBA8	Non-Toxin
P85224	Non-Toxin
P20416	Toxin
P86255	Toxin
P56855	Toxin
P85277	Toxin
P0DKX1	Non-Toxin
A8W7N4	Non-Toxin
P61508	Toxin
P0C7J9	Toxin
P0C349	Toxin
P0DOZ5	Toxin

Table 64: ToxinPred TN results

<i>UniProt ID</i>	<i>ToxinPred prediction</i>
P37817	Non-Toxin
Q8IVG9	Non-Toxin
E3YBA4	Toxin
P62945	Non-Toxin
P13191	Non-Toxin
P36937	Non-Toxin
C0HJF7	Non-Toxin
P83835	Non-Toxin
P84643	Non-Toxin
Q9NRI7	Non-Toxin
P84384	Non-Toxin
P85170	Non-Toxin
P80056	Non-Toxin
P16312	Non-Toxin
P83951	Non-Toxin
P58020	Non-Toxin
POCH48	Non-Toxin
PODMB9	Non-Toxin

Peptide ID	Peptide Sequence	SVM Score	Prediction	Hydrophobicity	Hydrophobicity	Hydrophilicity	Charge	Mol wt
	MKFYTIKLPKFLGGIVRAMLGSRK	-1.36	Non-Toxin	-0.09	0.19	-0.10	5.00	3018.13

Figure 56: Extract of a positive results from the ToxinPred webpage, using a sequence from TP dataset.

Peptide ID	Peptide Sequence	SVM Score	Prediction	Hydrophobicity	Hydrophobicity	Hydrophilicity	Charge	Mol wt
	GCCSDPRCRYRC	1.24	Toxin	-0.50	-0.93	0.50	2.00	1418.78

Figure 57: Extract of a negative results from the ToxinPred webpage, using a sequence from TN dataset.**Table 65:** ToxinPred Cooper' statistics (values are between 0 and 1)

<i>Dataset</i>	<i>Sensitivity</i>	<i>Specificity</i>	<i>Accuracy</i>
TP+TN	0.6	0.95	0.78

6. Evaluation of other freely available tools

6.1. Meme

The MEME tool identifies motifs in sequences, splitting variable-length patterns into two or more separate motifs. A motif in MEME is defined as *an approximate sequence pattern that occurs repeatedly in a group of related sequences*. The motifs representation is very simple to interpret because MEME is based on position-dependent letter-probability matrices that describe the probability of each possible letter at each position in the pattern. MEME accepts as input a group of sequences in FASTA format and outputs as many motifs as requested by the user, reporting the best width, number of occurrences, and description for each motif. MEME is available both as an online tool and as a downloadable locally installed program. The standard output is an .html file (Figure 58), containing all the results in a human-readable format and an .xml file for further machine processing. MEME is simple to use, and it may also be integrated into an automated *in silico* pipeline

A Meme search was performed using both the Main and TAS Collections (results not provided) in order to identify motifs for each family/domain as defined in the relevant databases used for this work (see Table 4).

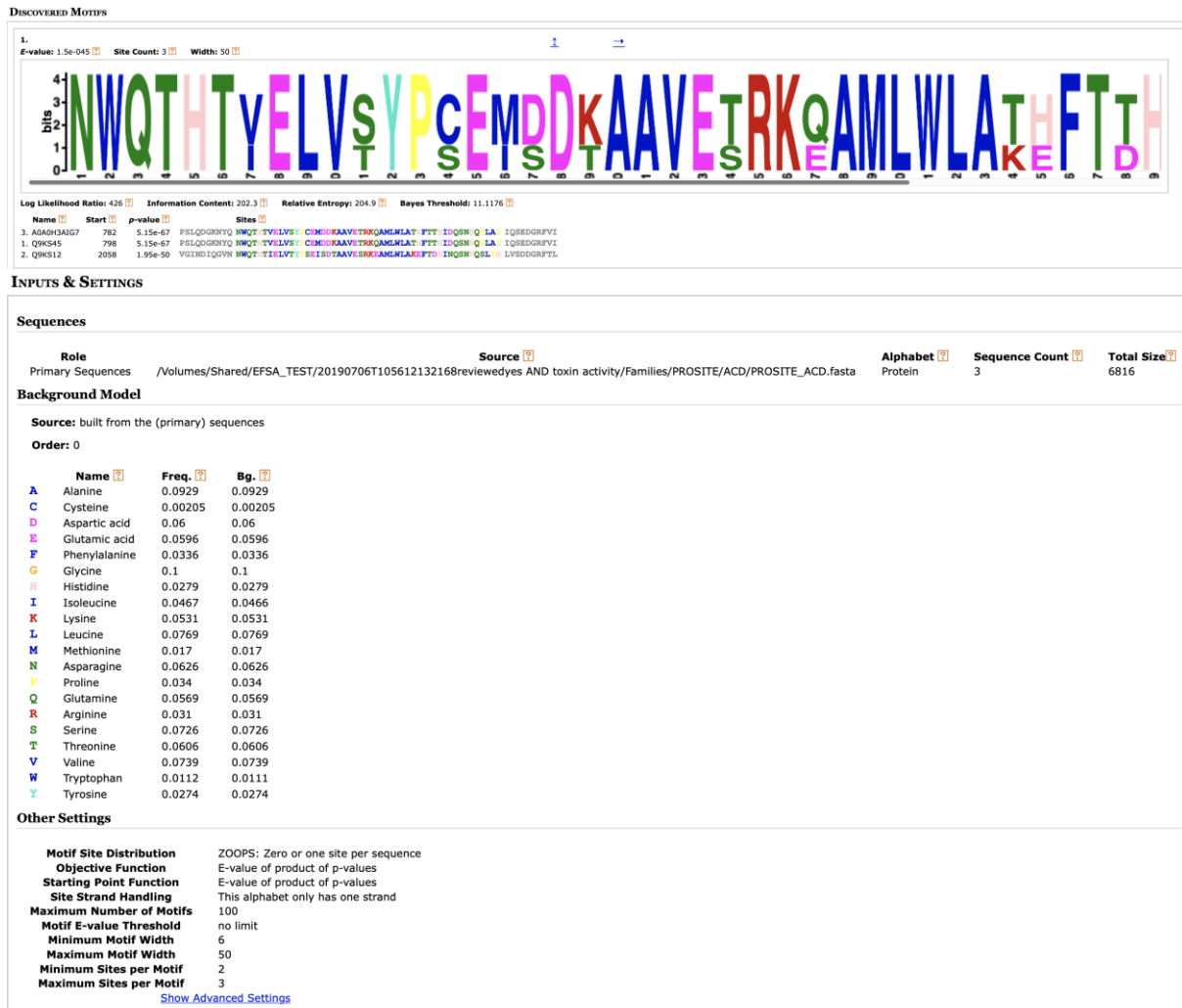


Figure 58: Example of a MEME discovered motif output. ACD domain from PROSITE was taken as example.

7. Discussion, conclusions and future perspectives

This work presents the outcome of a comprehensive search in peer-reviewed literature and in public protein databases designed to identify and retrieve information published on proteins with associated toxic effects in humans and animals (mammals, fish and birds). Our work addressed separately 1) toxic protein, i.e. proteins showing toxic activity either *per se* or as part of a toxin-antitoxin system, and 2) proteins prone to form endogenous aggregates in humans and animals and associated to disease.

For **toxic proteins** we adopted an integrated search strategy. Using the primary search string “toxin activity” on both protein (UniProtKB) and literature (WoS, SCOPUS and Pubmed) databases we found 6,964 proteins. By applying a growing strategy using the secondary search string “toxin-antitoxin system” we found an additional 627 proteins. Accordingly, we created two Collections to group the results from the two searches. For all the identified proteins, relevant toxicity information (e.g. GOs, keywords, source organism) was retrieved from UniProtKB and clustered. Excel tables were created for each Collection in order to summarise both data and related literature and to provide “raw” data to the user. The pathogenesis leading to the adverse effect(s) was also described for each toxin, and information on the underlying molecular mechanism of action(s) was collected and annexed to both Collections. UniProtKB unique IDs were used to manage all the information, classifying each toxic protein through this tag. The integrated strategy applied allowed us to reach the milestone of a comprehensive

literature search of proteins with well-recognized toxic activity. A Python-based automated tool (script) was developed to retrieve information and facilitate the management of the whole pipeline.

3D experimentally-determined and predicted (modelled) structures were also collected respectively from the RCSB Protein Data Bank and from the Swiss-Model repository for proteins with sequence more than 20 amino acids in length. We also computed novel 3D models when neither experimentally-determined structures nor already modelled 3D structures were available. Quality check of both experimentally-determined structures and predicted models was carried out. All the 3D structures were refined via MOE tools to fill gaps or add missing atoms. With this *in silico* pipeline, we obtained, to the best of our knowledge, the first 'complete set' of the 3D structures of proteins with a well-recognised toxic activity.

A comprehensive search was also performed to cluster selected toxins based on protein family classification as well as identified domains and other molecular signatures: eleven databases were queried for these analyses; family classification was retrieved from the Pfam database, while domains and signatures were obtained from InterPro. These two databases were considered as reference sources for the two types of information. On the basis of these classifications, sequence alignments were performed using primary structures of toxins belonging to the same family or having the same architecture (same domains). 3D structure superimpositions were also carried out, highlighting common features. This *in silico* and database-oriented strategy allowed us to reach the milestone of the comprehensive search of families, domains and signatures.

Other freely available tools were also tested, in order to provide to EFSA an evaluation for the ability of these tools in discriminating between toxins and non-toxic proteins. The performance of these tools was assessed by computing Cooper's statistics using different sub-sets of data.

Proteins prone to form endogenous aggregates in humans and animals when in mutated form or expressed at abnormally high levels in a disease-state, or accumulated when inefficiently catabolized, do not fit into the definition of toxin as set by the GO database (protein that "interacts selectively with one or more biological molecules in another organism - the "target" organism). Their nature and behaviour were anyway investigated by manually inspecting their features as reported in the relevant UniProtKB entries. It was possible to verify that all these proteins (retrieved with keywords 'amyloid' and 'aggregat*') are physiological components of human/animal tissues and that aggregation occurs only under specific conditions. While aggregation is positively associated with disease, it is not clear if aggregates may be themselves the cause of the pathological events or the adverse effects may arise from the reduced availability of the native monomeric protein or because of several factors that may interact synergistically.

Overall, this work features a literature search and an up-to-date collection of information on proteins associated with adverse (toxic) effects. All the evidence is gathered from databases that are considered reference sources by the scientific community. Reference databases for protein sequence, structure and activity and key literature databases were used. Thus, the implemented strategy described in this report highlighted the key protein information sources available and resulted in the production of two comprehensive and complementary "collections" of proteins associated with adverse (toxic) effects and related relevant information.

Such collections can be an important reference set to support the risk assessment of novel or poorly characterised proteins. These can be mined for further investigations in several directions for either research or risk assessment purposes. For instance, structural data can be used for other *in silico* analyses, e.g. molecular docking and molecular dynamics simulations. This could provide an insight towards the understanding of the molecular initiating event (MIE) of toxicological pathways.

The data collected in this work may contribute to the development of a comprehensive *in silico* risk assessment strategy and more in general can be the basis for setting new risk assessment strategies for new proteins. Methodologies could be devised and tested for sensitivity, specificity and accuracy, in a similar way to what has already been extensively done for chemicals. For example, homology detection can be performed for inferring (toxic) properties from well characterized biopolymers, since homologous proteins can share similar structural architecture and functions.

These data can also be used in conjunction with already available and well-established methods in protein science. For example, using 'protein BLAST' to analyse a protein sequence for similarity against toxic proteins and restricting the analysis to the identified proteins in the constructed Collections instead of the whole UniProtKB database, could increase the probability of identifying biologically significant similarities against known and well-annotated toxins. Additional strategies can be developed using the gathered data that are based on the use of multiple alignments, specific domains and/or molecular signatures as well as hidden Markov models (hMMs)-oriented approaches. Such strategies could be used to assess whether a protein belongs to a family that is known to contain toxic proteins. Moreover, artificial intelligence approaches could also be applied combining findings from methods such as the above-mentioned in order to try to increase the accuracy of predictive risk assessment for proteins.

The data collected for this work can form the basis for further information gathering in order to identify those protein structural/functional elements relevant in the molecular initiating events leading to toxicity.

References

- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: Tool for the unification of biology. *Nat. Genet.* <https://doi.org/10.1038/75556>
- Attwood, T.K., Beck, M.E., Bleasby, A.J., Parry-Smith, D.J., 1994. PRINTS - A database of protein motif fingerprints. *Nucleic Acids Res.*
- Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L.G., Figueira, L., Garmiri, P., Georghiou, G., Gonzalez, D., Hatton-Ellis, E., Li, W., Liu, W., Lopez, R., Luo, J., Lussi, Y., MacDougall, A., Nightingale, A., Palka, B., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Speretta, E., Turner, E., Tyagi, N., Volynkin, V., Wardell, T., Warner, K., Watkins, X., Zaru, R., Zellner, H., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., ArgoudPuy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M.C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., De Castro, E., Coudert, E., CuChe, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Noupikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A.L., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Garavelli, J.S., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Yeh, L.S., Zhang, J., 2017. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169. <https://doi.org/10.1093/nar/gkw1099>
- Benkert, P., Biasini, M., Schwede, T., 2011. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27, 343–350. <https://doi.org/10.1093/bioinformatics/btq662>
- Benson, M.D., 2012. Pathogenesis of transthyretin amyloidosis, in: *Amyloid*. pp. 14–15. <https://doi.org/10.3109/13506129.2012.668501>
- Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D., Zardecki, C., 2002. The protein data bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 58, 899–907. <https://doi.org/10.1107/S0907444902003451>
- Bienert, S., Waterhouse, A., De Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., Schwede, T., 2017. The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res.* 45, D313–D319. <https://doi.org/10.1093/nar/gkw1132>

- Bratosiewicz-Wasik, J., Wasik, T.J., Liberski, P.P., 2004. Molecular approaches to mechanisms of prion diseases. *Folia Neuropathol.* <https://doi.org/10.1080/00074919812331337390>
- Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J.M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D.S., Green, R.K., Guranović, V., Guzenko, D., Hudson, B.P., Kalro, T., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Periskova, I., Prlić, A., Randle, C., Rose, A., Rose, P., Sala, R., Sekharan, M., Shao, C., Tan, L., Tao, Y.P., Valasatava, Y., Voigt, M., Westbrook, J., Woo, J., Yang, H., Young, J., Zhuravleva, M., Zardecki, C., 2019. RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* 47, D464–D474. <https://doi.org/10.1093/nar/gky1004>
- Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N.L., Lewis, S.E., Mungall, C.J., Basu, S., Chisholm, R.L., Dodson, R.J., Hartline, E., Fey, P., Thomas, P.D., Albou, L.P., Ebert, D., Kesling, M.J., Mi, H., Muruganujan, A., Huang, X., Poudel, S., Mushayahama, T., Hu, J.C., LaBonte, S.A., Siegele, D.A., Antonazzo, G., Attrill, H., Brown, N.H., Fexova, S., Garapati, P., Jones, T.E.M., Marygold, S.J., Millburn, G.H., Rey, A.J., Trovisco, V., Dos Santos, G., Emmert, D.B., Falls, K., Zhou, P., Goodman, J.L., Strelets, V.B., Thurmond, J., Courtot, M., Osumi, D.S., Parkinson, H., Roncaglia, P., Acencio, M.L., Kuiper, M., Lreid, A., Logie, C., Lovering, R.C., Huntley, R.P., Denny, P., Campbell, N.H., Kramarz, B., Acquaah, V., Ahmad, S.H., Chen, H., Rawson, J.H., Chibucos, M.C., Giglio, M., Nadendla, S., Tauber, R., Duesbury, M.J., Del, N.T., Meldal, B.H.M., Perfetto, L., Porras, P., Orchard, S., Shrivastava, A., Xie, Z., Chang, H.Y., Finn, R.D., Mitchell, A.L., Rawlings, N.D., Richardson, L., Sangrador-Vegas, A., Blake, J.A., Christie, K.R., Dolan, M.E., Drabkin, H.J., Hill, D.P., Ni, L., Sitnikov, D., Harris, M.A., Oliver, S.G., Rutherford, K., Wood, V., Hayles, J., Bahler, J., Lock, A., Bolton, E.R., De Pons, J., Dwinell, M., Hayman, G.T., Laulederkind, S.J.F., Shimoyama, M., Tutaj, M., Wang, S.J., D'Eustachio, P., Matthews, L., Balhoff, J.P., Aleksander, S.A., Binkley, G., Dunn, B.L., Cherry, J.M., Engel, S.R., Gondwe, F., Karra, K., MacPherson, K.A., Miyasato, S.R., Nash, R.S., Ng, P.C., Sheppard, T.K., Shrivatsav Vp, A., Simison, M., Skrzypek, M.S., Weng, S., Wong, E.D., Feuermann, M., Gaudet, P., Bakker, E., Berardini, T.Z., Reiser, L., Subramaniam, S., Huala, E., Arighi, C., Auchincloss, A., Axelsen, K., Argoud, G.P., Bateman, A., Bely, B., Blatter, M.C., Boutet, E., Breuza, L., Bridge, A., Britto, R., Bye-A-Jee, H., Casals-Casas, C., Coudert, E., Estreicher, A., Famiglietti, L., Garmiri, P., Georghiou, G., Gos, A., Gruaz-Gumowski, N., Hatton-Ellis, E., Hinz, U., Hulo, C., Ignatchenko, A., Jungo, F., Keller, G., Laiho, K., Lemercier, P., Lieberherr, D., Lussi, Y., Mac-Dougall, A., Magrane, M., Martin, M.J., Masson, P., Natale, D.A., Hyka, N.N., Pedruzzi, I., Pichler, K., Poux, S., Rivoire, C., Rodriguez-Lopez, M., Sawford, T., Speretta, E., Shypitsyna, A., Stutz, A., Sundaram, S., Tognolli, M., Tyagi, N., Warner, K., Zaru, R., Wu, C., Chan, J., Cho, J., Gao, S., Grove, C., Harrison, M.C., Howe, K., Lee, R., Mendel, J., Muller, H.M., Raciti, D., Van Auken, K., Berriman, M., Stein, L., Sternberg, P.W., Howe, D., Toro, S., Westerfield, M., 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338. <https://doi.org/10.1093/nar/gky1055>
- Carnate, M., Ed, M., 2008. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci* 95, 5857–5864. <https://doi.org/10.1073/pnas.95.11.5857>
- Comenzo, R.L., 2006. Systemic immunoglobulin light-chain amyloidosis. *Clin. Lymphoma Myeloma.* <https://doi.org/10.3816/CLM.2006.n.056>
- Consortium, T.U., 2008. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 36, 190–195. <https://doi.org/10.1093/nar/gkm895>
- Costanzo, M., Zurzolo, C., 2013. The cell biology of prion-like spread of protein aggregates: mechanisms and implication in neurodegeneration. *Biochem. J.* 452, 1–17. <https://doi.org/10.1042/BJ20121898>
- Dang, L., Van Damme, E.J.M., 2015. Toxic proteins in plants. *Phytochemistry* 117, 51–64. <https://doi.org/10.1016/j.phytochem.2015.05.020>
- Dawson, N.L., Lewis, T.E., Das, S., Lees, J.G., Lee, D., Ashford, P., Orengo, C.A., Sillitoe, I., 2017.

- CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* 45, D289–D295. <https://doi.org/10.1093/nar/gkw1098>
- de Castro, E., Sigrist, C.J.A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., Hulo, N., 2006. ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* 34. <https://doi.org/10.1093/nar/gkl124>
- Durand, S., Jahn, N., Condon, C., Brantl, S., 2012. Type I toxin-antitoxin systems in *Bacillus subtilis*. *RNA Biol.* <https://doi.org/10.4161/rna.22358>
- Finn, R.D., Coghill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. <https://doi.org/10.1093/nar/gkv1344>
- Forloni, G., Balducci, C., 2018. Alzheimer's Disease, Oligomers, and Inflammation. *J. Alzheimer's Dis.* <https://doi.org/10.3233/JAD-170819>
- Franceschi, N., Paraskevopoulos, K., Waigmann, E., Ramon, M., 2017. Predictive Protein Toxicity and Its Use in Risk Assessment. *Trends Biotechnol.* <https://doi.org/10.1016/j.tibtech.2017.03.010>
- Gooptu, B., Lomas, D.A., 2009. Conformational Pathology of the Serpins: Themes, Variations, and Therapeutic Strategies. *Annu. Rev. Biochem.* 78, 147–176. <https://doi.org/10.1146/annurev.biochem.78.082107.133320>
- Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., Lopez, R., 2010. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* 38. <https://doi.org/10.1093/nar/gkq313>
- Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K., Beck, E., 2013. TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.* 41, 387–395. <https://doi.org/10.1093/nar/gks1234>
- Huang, X., Zheng, Y., Zhang, F., Wei, Z., Wang, Y., Carrell, R.W., Read, R.J., Chen, G.Q., Zhou, A., 2016. Molecular mechanism of α 1-antitrypsin deficiency. *J. Biol. Chem.* 291, 15674–15686. <https://doi.org/10.1074/jbc.M116.727826>
- Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B.A., De Castro, E., Lachaize, C., Langendijk-Genevaux, P.S., Sigrist, C.J.A., 2008. The 20 years of PROSITE. *Nucleic Acids Res.* 36. <https://doi.org/10.1093/nar/gkm977>
- Iadanza, M.G., Jackson, M.P., Hewitt, E.W., Ranson, N.A., Radford, S.E., 2018. A new era for understanding amyloid structures and disease. *Nat. Rev. Mol. Cell Biol.* <https://doi.org/10.1038/s41580-018-0060-8>
- Kaas, Q., Yu, R., Jin, A.H., Dutertre, S., Craik, D.J., 2012. ConoServer: Updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkr886>
- Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., Park, Y.M., Buso, N., Lopez, R., 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* 43, W580–W584. <https://doi.org/10.1093/nar/gkv279>
- Lucas, A., Yaron, J.R., Zhang, L., Ambadapadi, S., 2018. Overview of serpins and their roles in biological systems, in: *Methods in Molecular Biology*. pp. 1–7. https://doi.org/10.1007/978-1-4939-8645-3_1
- Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Geer, L.Y., Bryant, S.H., 2017. CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45, D200–D203. <https://doi.org/10.1093/nar/gkw1129>

- Masood, R., Ullah, K., Ali, H., Ali, I., Betzel, C., Ullah, A., 2018. Spider's venom phospholipases D: A structural review. *Int. J. Biol. Macromol.* <https://doi.org/10.1016/j.ijbiomac.2017.09.081>
- Nativi-Nicolau, J., Maurer, M.S., 2018. Amyloidosis cardiomyopathy: update in the diagnosis and treatment of the most common types. *Curr. Opin. Cardiol.* 1. <https://doi.org/10.1097/HCO.0000000000000547>
- Nielsen, M., Lundegaard, C., Lund, O., Petersen, T.N., 2010. CPHmodels-3.0-remote homology modeling using structure-guided sequence profiles. *Nucleic Acids Res.* 38, 576–581. <https://doi.org/10.1093/nar/gkq535>
- Nikolskaya, A.N., Arighi, C.N., Huang, H., Barker, W.C., Wu, C.H., 2006. PIRSF Family Classification System for Protein Functional and Evolutionary Analysis. *Evol. Bioinforma.* 2, 117693430600200. <https://doi.org/10.1177/117693430600200033>
- Ohtake, S., Kita, Y., Payne, R., Manning, M., Arakawa, T., 2013. Structural Characteristics of Short Peptides in Solution. *Protein Pept. Lett.* <https://doi.org/10.2174/092986652012131112121417>
- Parisi, K., Shafee, T.M.A., Quimbar, P., van der Weerden, N.L., Bleackley, M.R., Anderson, M.A., 2018. The evolution, function and mechanisms of action for plant defensins. *Semin. Cell Dev. Biol.* <https://doi.org/10.1016/j.semdb.2018.02.004>
- Postic, G., Gracy, J., Périn, C., Chiche, L., Gelly, J.C., 2018. KNOTTIN: The database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkx1084>
- Saraiva, M.J.M., 2001. Transthyretin amyloidosis: A tale of weak interactions. *FEBS Lett.* [https://doi.org/10.1016/S0014-5793\(01\)02480-2](https://doi.org/10.1016/S0014-5793(01)02480-2)
- Sigrist, C.J.A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., Bucher, P., 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* 3, 265–274. <https://doi.org/10.1093/bib/3.3.265>
- Sigrist, C.J.A., De Castro, E., Cerutti, L., Cuče, B.A., Hulo, N., Bridge, A., Bougueleret, L., Xenarios, I., 2013. New and continuing developments at PROSITE. *Nucleic Acids Res.* 41. <https://doi.org/10.1093/nar/gks1067>
- Soto, C., Pritzkow, S., 2018. Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. *Nat. Neurosci.* <https://doi.org/10.1038/s41593-018-0235-9>
- Tasoulis, T., Isbister, G.K., 2017. A review and database of snake venom proteomes. *Toxins (Basel).* <https://doi.org/10.3390/toxins9090290>
- Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., Narechania, A., 2003. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129–2141. <https://doi.org/10.1101/gr.772403>
- Vandenborre, G., Smagghe, G., Van Damme, E.J.M., 2011. Plant lectins as defense proteins against phytophagous insects. *Phytochemistry.* <https://doi.org/10.1016/j.phytochem.2011.02.024>
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., De Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T., 2018. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303. <https://doi.org/10.1093/nar/gky427>
- Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., Gough, J., 2009. SUPERFAMILY - Sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* 37, 380–386. <https://doi.org/10.1093/nar/gkn762>
- Yuan, X., Lyu, S., Zhang, H., Hang, X., Shi, W., Liu, L., Wu, Y., 2019. Complete genome sequence of novel isolate SYJ15 of *Bacillus cereus* group, a highly lethal pathogen isolated from Chinese soft shell turtle (*Pelodiscus Sinensis*). *Arch. Microbiol.* <https://doi.org/10.1007/s00203-019-01723-y>