bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Razor: annotation of signal peptides from toxins

Bikash K. Bhandari[1], Paul P. Gardner[1,2], Chun Shen Lim[1,*]

[1]Department of Biochemistry, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand
[2]Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand

*Corresponding author. Email: chunshen.lim@otago.ac.nz

ABSTRACT

Motivation: Signal peptides are responsible for protein transport and secretion and are ubiquitous to all forms of life. The annotation of signal peptides is important for understanding protein translocation and toxin secretion and evolution.

Results: Here we explore the features of these signal sequences from eukaryotic proteins. Strikingly, we find that the signal peptides from secretory toxins have common features across kingdoms, supporting the idea of horizontal gene transfer or convergence of toxin genes across kingdoms. We leverage these features to build Razor, a simple yet powerful tool specialised in identifying signal peptides from toxins using the first 23 N-terminal residues. We demonstrate the usability of Razor by analysing all the sequences reviewed by UniProt. Indeed, Razor is able to identify toxins using their N-terminal sequences only. Strikingly, we also discover that many defensive proteins across kingdoms harbour a toxin-like signal peptide; some of these defensive proteins have emerged through convergent evolution, e.g. defensin and defensin-like protein families, and phospholipase families. In sum, Razor uses an approach independent of homology search to identify novel and known toxin classes across species using N-terminal residues.

Availability and implementation: Razor is available as a web application (https://tisigner.com/razor) and a command-line tool (https://github.com/Gardner-BinfLab/Razor).

INTRODUCTION

Secretory proteins are translocated in the secretory pathway with the assistance of a short peptide extension at the N-terminus. This special targeting peptide is known as the Signal Peptide (SP) (von Heijne, 1990). Secretory pathways and their corresponding SPs have evolved across organisms to carry out different functions (Hegde and Bernstein, 2006; Owji *et al.*, 2018). Despite being ubiquitous across all domains of life, SPs do not share a consensus. Nevertheless, a SP usually consists of three regions: a positively charged domain (N-region), a hydrophobic core (H-region), followed by a polar but electrically neutral domain (C-region) containing a cleavage site (von Heijne, 1985, 1990; Nielsen and Krogh, 1998). Apart from translocating proteins, SPs are also responsible for several other roles, such as in regulatory functions, antigen presentation, and some human diseases (Borrego *et al.*, 1998; Datta *et al.*, 2007; Owji *et al.*, 2018).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

An important group of secretory proteins is toxins, whose precursors almost always contain SPs (Fry *et al.*, 2009). Toxins have evolved in all domains of life primarily as a defense mechanism or for predation (Casewell *et al.*, 2013). Furthermore, several organisms in the animal kingdom have evolved to create venoms, which consist of a complex mixture of different types of toxins, usually with a specialised apparatus to facilitate their delivery. Such adaptations may have evolved through convergence or duplication and neofunctionalisation (Casewell, 2020). However, a recent study found that at least five toxin gene families were horizontally transferred from bacteria and fungi to centipedes (Undheim and Jenner, 2021), suggesting common features exist in these gene families. Besides, the pharmacological actions of toxins on living cells are often employed to develop anti-toxins, novel drugs, and pathogen-resistant transgenic crops (King, 2011; Estrada *et al.*, 2007; Bidondo *et al.*, 2019; Samy *et al.*, 2017; Li *et al.*, 2018). Hence, annotating SPs is essential in the functional and structural studies of proteins in fundamental research, commercial, and pharmaceutical industries. In addition, understanding the presence or absence of SPs in the genes of interest is critical for choosing the appropriate recombinant protein expression and purification systems, as the intracellular accumulation of secretory proteins and toxins may be toxic to the host cells. Indeed, the ability of SPs to translocate proteins has been utilised in recombinant protein expression systems for high quality and quantity results (Futatsumori-Sugai and Tsumoto, 2010; Cho *et al.*, 2019; Karyolaimos *et al.*, 2019; Peng *et al.*, 2019).

Despite the immense use cases of toxins, there are very few tools to predict them, such as ClanTox, ToxinPred, TOXIFY, and ToxClassifier, some being specialised such as SpiderP for spider toxins (Naamati *et al.*, 2010; Gupta *et al.*, 2013; Wong *et al.*, 2013; Gacesa *et al.*, 2016; Cole and Brewer, 2019). Moreover, these methods are based on the properties of the mature peptides (or the propeptides), rather than the SPs. To address these issues, we first examined the features of SPs from eukaryotic proteins and toxins. We then exploited those features to build Razor, a new tool for annotating SPs. We have optimised the command-line version of Razor for high-throughput analysis and used it to predict new SPs by scanning all the sequences reviewed by UniProt (UniProt Consortium, 2019). We were able to predict novel toxins and defensive proteins using only the first 23 N-terminal residues, as evidenced by the protein family annotations.

MATERIALS AND METHODS

Datasets

We retrieved the training dataset for the state-of-the-art SP prediction program SignalP 5.0, which is a curated set of the N-terminal sequences from all domains of life (Almagro Armenteros *et al.*, 2019). To get the full sequences and annotations of eukaryotic proteins, we used UniProt's ID mapping service (UniProt Consortium, 2019) and obtained 17,264 fully annotated sequences, of which 2,609 sequences have been experimentally validated to harbour functional SPs. These sequences were used to build a generic, eukaryotic SP classifier. For feature analysis, we clustered these sequences (60 N-terminal residues) at an identity threshold of 70% using CD-HIT v4.8.1 (Fu *et al.*, 2012). A single representative sequence was retained for each cluster to reduce sequence redundancy (Supplementary Table S1).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To build a classifier specialised for annotating toxin SPs, we manually curated a separate positive set using the dataset from the animal toxin annotation project (Jungo *et al.*, 2012) and a subset from the above training set. Other SPs were assigned as a negative set. We then clustered the sequences as above and analysed the representative sequences (Supplementary Table S1).

The SP classifiers were compared using an independent test set retrieved from UniProt on 16 February 2021. In particular, the eukaryotic SP classifier was evaluated using 241 SPs with experimental evidence and 52,055 non-SPs, whereas the toxin SP classifier was evaluated using a subset of this independent set (toxin SPs=47, non-SPs=52,055). We also scanned the reviewed sequences from UniProt (N=561,776, retrieved on 2 September 2020).

Bit score

The bit scores of the N-terminal residues were computed as:

$$Bit\ score_{residue} = log_2(\frac{Normalised\ count\ of\ residue\ in\ the\ positive\ set}{Normalised\ count\ of\ residue\ in\ the\ background\ set})$$

For eukaryotic proteins, the positive set and the background set were SPs and non-SPs, respectively. For toxins, the positive set and the background set were toxin SPs and non-toxin SPs, respectively.

Protein sequence properties

The standard protein sequence properties, implemented in BioPython, were calculated using the Bio.SeqUtils.ProtParam module v1.73 (Cock *et al.*, 2009). These features include GRand AVerage of hydropathicitY (GRAVY), Flexibility, Helix, Sheet and Turn propensities, Instability Index, Aromaticity, and Isoelectric Point. An additional feature included is the Solubility-Weighted Index (SWI; (Bhandari *et al.*, 2020).

SP classifiers

We built a random forest classifier based on several sequence features (GRAVY, flexibility, helix, and SWI), as well as the counts of residues (R, K, N, D, C, E, V, I, Y, F, W, L, Q, and P) of the first 30 N-terminal residues. The residues were chosen such that they maximised Matthew's correlation coefficient (MCC) in five-fold cross-validations. After the cross-validation step, we generated five random forest models, which are used for scoring the N-terminal of a given sequence. The scores from these classifiers are comparable to the S-score of SignalP 4.0 except that our scores are non-position-specific (Petersen *et al.*, 2011).

For the prediction of the cleavage site, we took a total of 30 residues such that the cleavage site is aligned in between positions 15 and 16 in order to capture the major differences in residue distribution around the cleavage site. We built a 20×30 matrix and populated it with the hydrophobicity scale (Kyte and Doolittle, 1982) as initial weights. We then used multi-objective simulated annealing (Kirkpatrick *et al.*, 1987) at each position such that the new weights maximised the AUC and precision-recall curve based on the training set. The scoring of the cleavage site (C-score) is done using the random forest classifier trained on

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the aligned set encoded using the optimised weight matrix. Small limitation of our approach is that we are unable to detect the correct cleavage site if it is located before the 15th position. Yet, based on training data, this is rarely observed (N=13).

After detecting the cleavage site, the final score for classification (Y-score) is the geometric mean $Y = \sqrt{S \times C}$, where S is the S-score and C is the max of C-scores along the sequence. For the final classifier, we chose a threshold of Y-score that maximised the MCC after five-fold cross-validations (MCC=0.914) on the training set.

We then built models specialised in annotating the toxin SPs based on hydrophobicity, SWI, flexibility, and turn. These features were selected such that they maximised the MCC using five-fold cross-validations on the training set. The N-terminal length of 23 was found to generate the maximum median MCC score for the toxin SP classifier (MCC=0.741, see also Supplementary Table S2). Similar to the SP prediction models, the toxin SP classifiers consist of five models each.

Performance measures

We use MCC as a measure of performance to correctly identify eukaryotic SPs. We also use cleavage site precision $(CS_P = N_{corr}/N_P)$ and recall $(CS_R = N_{corr}/N)$, where N_{corr} is the number of the correctly identified cleavage site, N_P is the number of predicted SPs and N is the number of SPs (Almagro Armenteros *et al.*, 2019; Savojardo *et al.*, 2018).

Tool

We developed Razor for annotating SPs using the eukaryotic and toxin SP classifiers (Fig 1). Razor accepts either a nucleotide sequence or a protein sequence. Sequences with a length of lower than 30 residues are padded with Serine (Ser, S), because it shows equal enrichment across all datasets, in particular after the H-region (Fig 2). Razor is available both as a command-line tool (https://github.com/Gardner-BinfLab/Razor) and a web application (https://tisigner.com/razor). For the web application, predictions from five models are displayed as stars. The final score is the median of scores from five models and is displayed along with the region for SP. A plot of C-scores along the sequence is also displayed along with the annotation for the cleavage site. In addition, we integrated the Razor web application with our protein expression and solubility optimisation tools, TIsigner and SoDoPE, respectively (Bhandari *et al.*, 2020, 2021). Our web tools assist users in annotating SPs and protein domains, and making the decisions from gene cloning to protein expression and purification.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

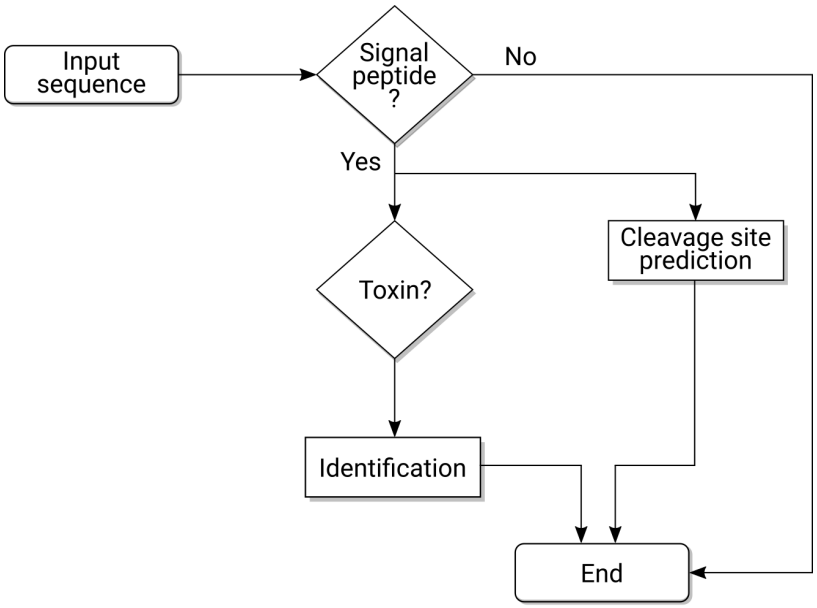

Fig 1. Flow chart of toxin SP classification using Razor.

Statistical analysis

Data analysis was performed using pandas v1.0.3 (McKinney, 2010). Hydrophobicity and SWI were smoothed for the classifier training using the Savitzky-Golay filter implemented in SciPy v1.4.1 (Virtanen *et al.*, 2020). Random forest classifier and MCC computation were done using scikit-learn v0.23.1 (Pedregosa *et al.*, 2011). Plots were generated using Matplotlib v3.1.3 and Seaborn v0.10.0 (Hunter, 2007; Waskom *et al.*, 2020).

Code and data availability

Jupyter notebooks for reproducing our analyses are available at https://github.com/Gardner-BinfLab/Razor_paper_2021. The source code for Razor, our SP annotation server can be found at https://github.com/Gardner-BinfLab/TISIGNER-ReactJS.

RESULTS

Toxin SPs have distinct sequence properties

We investigated the sequence composition of SPs by first aligning the sequences from the N-terminal residue or by centering at the cleavage sites, followed by computing bit scores for each residue (Fig 2). These approaches provide sufficient leverage to enumerate the tripartite domains of SPs (N-, H-, and C-domains). In general, hydrophobic residues are enriched towards the N-termini (H-region), which are characteristic features of SPs (von Heijne, 1990) (Supplementary Fig S1). Strikingly, the SPs of toxins show a strong abundance of isoleucine (I) and lack leucine (L) and alanine (A) residues in contrast to other eukaryotic SPs (Fig 2). This is supported by an amino acid composition analysis of the N-terminal subsequences (Supplementary Fig S2). We also analysed other features of these N-terminal subsequences, including GRAVY, structural flexibility, helix, sheet and turn propensities, instability index, aromaticity, isoelectric point, and SWI. Interestingly, isoelectric point appears as a prominent feature of toxin SPs (Supplementary Fig S3).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The cleavage sites mark the end of SPs and the beginning of the mature region (or the propeptide), which is a unique feature of SPs (Fig 2). By aligning the sequences at the cleavage sites, we observed a clear emergence of (−3,−1) rule preceding the cleavage sites, i.e. a distinctive presence of small and charged residues such as alanine (A) and valine (V) (von Heijne, 1983).

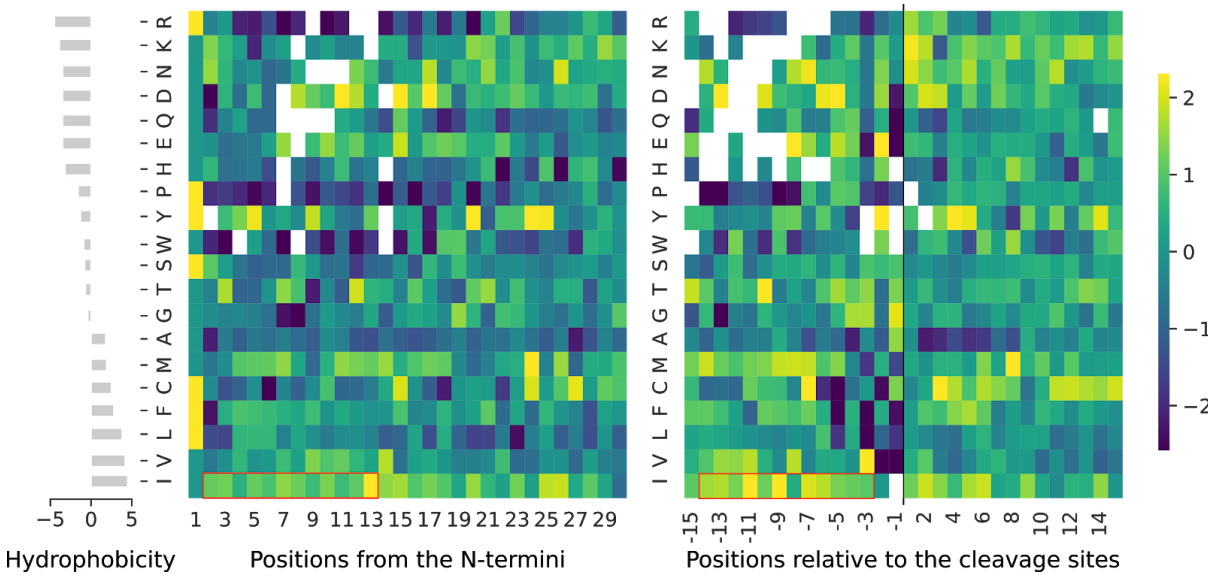

Fig 2. The Signal Peptides (SPs) from toxins are enriched with isoleucine residues in contrast to other eukaryotic SPs. The bar plot shows Kyte and Doolittle's hydrophobicity scale. The heatmaps show the enrichment of residues in bit scores by aligning SPs from the N-termini (left) and at the cleavage sites (right, black vertical line). The unfilled, red rectangles indicate the enrichment of isoleucine residues (I). The white spaces correspond to the absence of residues at certain positions due to limited sample size (261 toxin SPs and 1,738 non-toxin SPs that have been experimentally validated).

Razor accurately predicts toxin SPs

By taking these important features into account, we built SP classifiers to annotate eukaryotic and toxin SPs using random forest (Fig 1). Only SPs with experimental evidence were used for training. We compared these classifiers using an independent test set, where, the MCC, and the cleavage site precision and recall of Razor for eukaryotic SP prediction were 0.405, 0.136, 0.596, respectively (SPs vs non-SPs, see Supplementary Fig S4, Table S3 and S4). More importantly, Razor outperforms state-of-the-art in toxin SP prediction, achieving an MCC score of 0.611, and the cleavage site precision and recall of 0.355 and 0.831, respectively (toxin SPs vs non-SPs, see Fig 3, and Supplementary Table S5 and S6).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

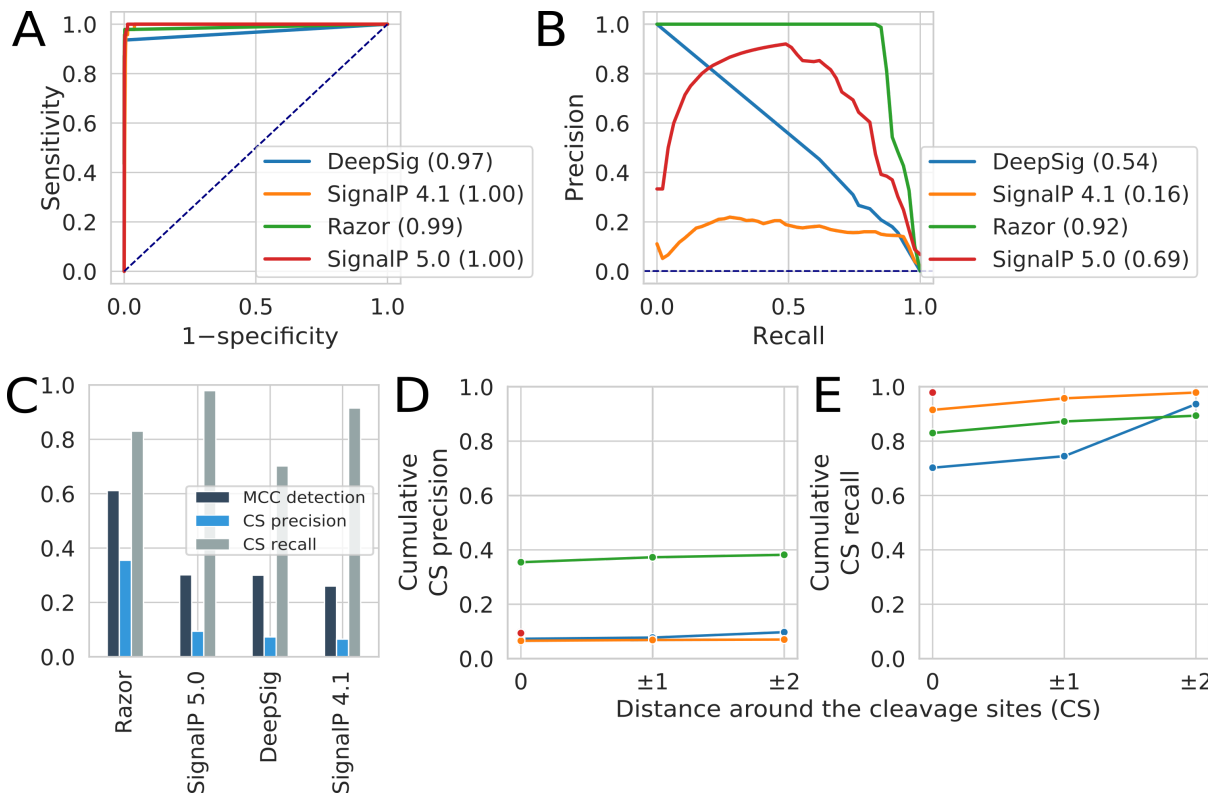

Fig 3. Razor outperforms other tools in predicting toxin SPs. Benchmarks were carried out using an independent test set (47 experimentally validated toxin SPs and 52,055 non-SPs). **(A)** Receiver operating characteristic curves **(B)** and precision recall curves **(C)** of the SP prediction tools. Areas under the curves are shown in parentheses. The dotted lines show the performance of a random classifier. **(C)** Matthew's Correlation Coefficients (MCC) of the SP prediction tools. The cleavage site (CS) precisions **(D)** and recalls **(E)** of windows surrounding the cleavage sites are shown. Data are available in Supplementary Tables S5 and S6.

Defensive proteins harbour a toxin-like SP

The training set for the toxin SP classifier was mainly composed of the SPs from animal toxins, e.g. snake three-finger toxins, scorpion toxins, and phospholipase A_2, and plant toxins, e.g. ribosome-inactivating proteins (Fig 4A). To further assess our new toxin SP classifier, we scanned the reviewed sequences from UniProt (N=561,776). A total of 910 sequences were predicted positive from all SP detection models.

In Fig 4, we excluded potential false positive hits, i.e. computationally annotated transmembrane proteins by UniProt (N=33). The remaining sequences were divided into two groups based on the presence or absence of toxin annotation. From these probable toxin SPs, 759 sequences had annotations for toxins. They included protein families such as scorpion toxin, phospholipase A_2 and ribosome-inactivating protein (Fig 4B). The remaining 110 sequences had no annotations for toxins. These sequences were clustered at an identity threshold of 70%, which gave rise to 100 representative sequences. Interestingly, many of these proteins without toxin annotations have some defensive properties such as antibacterial peptides and cyclotides. Furthermore, other defensive proteins such as

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

beta-defensin and defensin-like (DEFL) are the results of convergent evolution. For example, beta-defensin-like motifs are also found in toxins from lepidosauria (rattlesnakes and bearded dragons) and mammalia (platypus) (Fry *et al.*, 2009, 2010; Whittington *et al.*, 2008). This suggests why their SPs show some remote similarity with toxin SPs.

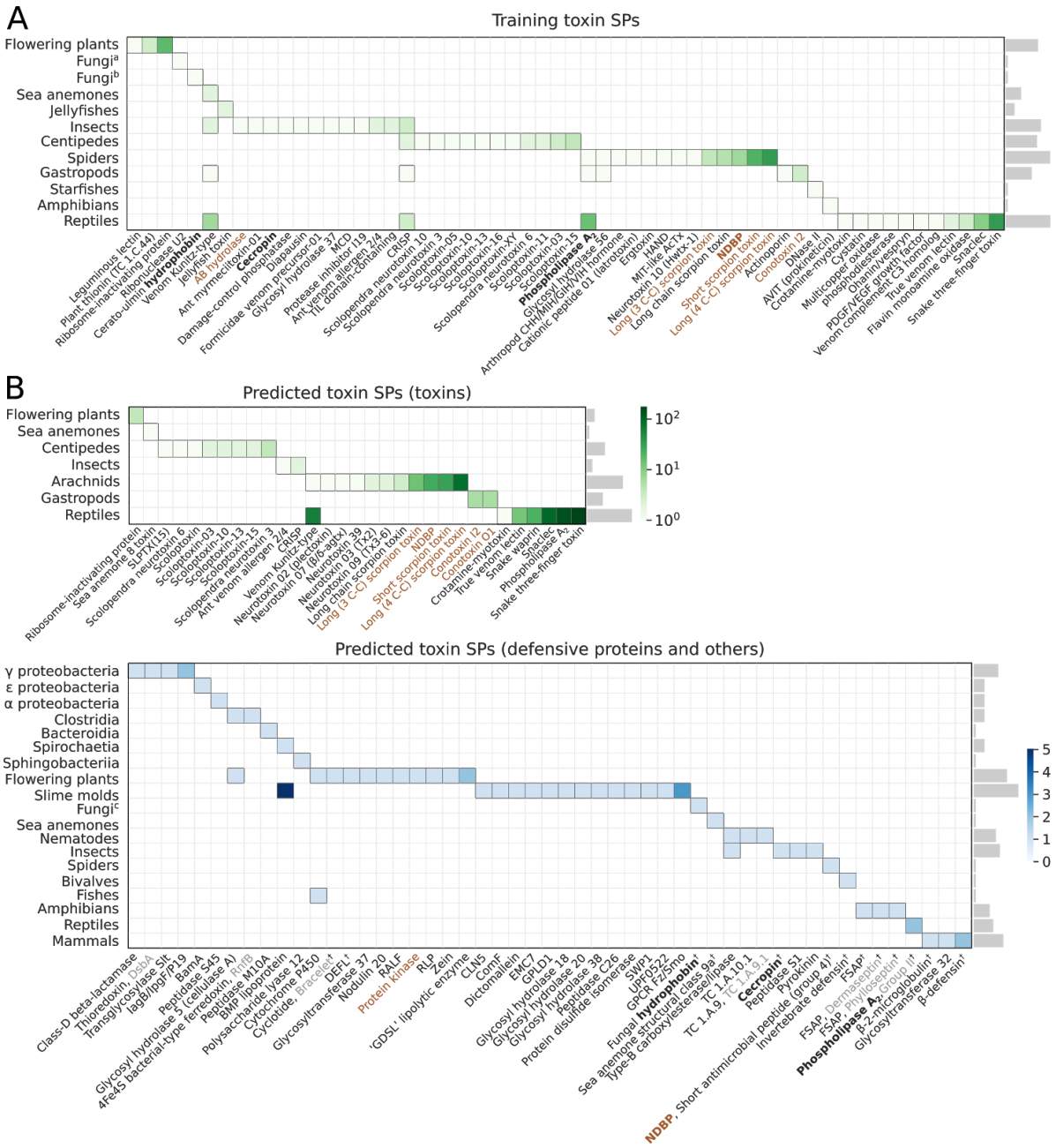

Fig 4. Razor identifies SPs from toxins along with several classes of defensive proteins. The reviewed sequences from UniProt were examined (N=561,776). **(A)** Heatmap shows the abundance of protein families in the training toxin sequences with SPs by taxa. A total of 237 of 261 training toxins had protein family annotations. **(B)** Heatmaps show the abundance of protein families in the sequences predicted to harbour toxin SPs. A total of 753 of 759 toxins predicted to harbour toxin SPs had protein family annotations (top). A total of 110 other types of proteins were predicted to harbour toxin SP, in which 76 of them had protein family annotations (bottom). The scale bars indicate the frequencies of protein

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

families. Those protein families that have defensive properties are marked with † (bottom). Protein families that are in common between the training and predicted toxin SP sequences are bolded (bottom panel). Protein subfamily, family and superfamily are shown in grey, black and brown, respectively. Fungi[a], Eurotiomycetes; Fungi[b], Sordariomycetes; Fungi[c], Agaricomycetes; CLN5, Ceroid-Lipofuscinosis Neuronal protein; ComF, Competence protein F; CRISP, Cysteine RIch Secretory Protein; DEFL, DEFensin Like; EMC7, ER membrane protein complex subunit 7; FSAP, Frog Skin Antimicrobial Peptide; GPLD1, Glycosyl-phosphatidylinositol-specific phospholipase D; HAND, Helical Arthropod-Neuropeptide-Derived; RALF, Rapid ALkalinization Factor; RLP, Receptor Like Protein; SLPTX, Scoloptoxin; UPF, Uncharacterised Protein Family.

DISCUSSION

We have studied the features of SPs from eukaryotic proteins. While SPs share a common hydrophobic nature, we have found several differences between toxin SPs and other eukaryotic SPs in their residue compositions and consequently the sequence properties. We have used these features to develop Razor for annotating eukaryotic SPs, which have specialised functionalities in annotating toxin SPs. Razor outperforms other sophisticated methods in predicting toxin SPs. Using Razor, we were able to predict several classes of probable toxins, which are yet to be annotated (Fig 4). Our predicted results consist of toxins and defensive proteins from diverse species, which gives us an overview of the source of toxins.

Since toxins and defensive proteins occur naturally in organisms to attack and neutralise foreign invaders, many of our predicted results include proteins involved in innate immune response and signalling. Some of the frequently observed biological processes of these proteins were ‘killing of cells of other organism [GO:0031640]’, ‘defense response to fungus [GO:0050832]’, ‘defense response to bacterium [GO:0042742]’ and ‘innate immune response [GO:0045087]’ (Supplementary Fig S5 and S6). Many toxins and defensive proteins are commercially important. For example, plant toxins such as defensin-like protein, animal toxins such as cecropin are used to develop disease-resistant transgenic crops (Stotz *et al.*, 2009; Lacerda *et al.*, 2014; Wu *et al.*, 2016; Boccardo *et al.*, 2019; Ali *et al.*, 2018). Similarly, the cytotoxic activity of phospholipase A_2 on cancer cells makes it a promising candidate for cancer therapy (Xiao *et al.*, 2017; Hiu and Yap, 2020; Lomonte and Rangel, 2012).

Taken together, Razor uses an approach independent of homology search to identify known and novel toxin classes across species. Razor was able to identify previously unannotated SPs and a spectrum of toxins and defensive proteins simply using the first 23 N-terminal residues. This also suggests a possible evolutionary constraint on SPs driven by the specialisation of the toxin secretory systems (or convergent evolution), and supports the idea of horizontal gene transfer of several toxin gene classes (Undheim and Jenner, 2021). Therefore, accurate annotation of toxin SPs can enhance comparative genomics analysis and genome sequencing projects. Razor might also be useful in other research areas such as recombinant protein expression, toxicology, transgenics, and drug design.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

AUTHORS CONTRIBUTIONS

CSL conceived the study. BKB carried out the analysis, built Razor, and drafted the manuscript. CSL and PPG supervised the study. All authors reviewed, edited, and approved the manuscript.

ACKNOWLEDGEMENTS

The authors thank Dr Astra Heywood for providing feedback on the figures and the Razor web application.

FUNDING

This work was supported in part by the Ministry of Business, Innovation and Employment [MBIE Smart Idea grant: UOOX1709 and MBIE Data Science Programmes grant: UOAX1932], and the Royal Society of New Zealand Te Apārangi [Marsden grant: 19-UOO-040].

CONFLICT OF INTEREST

None declared.

REFERENCES

Ali,S. *et al.* (2018) Pathogenesis-related proteins and peptides as promising tools for engineering plants with multiple stress tolerance. *Microbiol. Res.*, **212-213**, 29–37.

Almagro Armenteros,J.J. *et al.* (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.*, **37**, 420–423.

Bhandari,B.K. *et al.* (2021) Protein yield is tunable by synonymous codon changes of translation initiation sites. *bioRxiv*.

Bhandari,B.K. *et al.* (2020) Solubility-Weighted Index: fast and accurate prediction of protein solubility. *Bioinformatics*, **36**, 4691–4698.

Bidondo,L.F. *et al.* (2019) The overexpression of antifungal genes enhances resistance to rhizoctonia solani in transgenic potato plants without affecting arbuscular mycorrhizal symbiosis. *Crop Protection*, **124**, 104837.

Boccardo,N.A. *et al.* (2019) Expression of pathogenesis-related proteins in transplastomic tobacco plants confers resistance to filamentous pathogens under field trials. *Sci. Rep.*, **9**, 2791.

Borrego,F. *et al.* (1998) Recognition of Human Histocompatibility Leukocyte Antigen (HLA)-E Complexed with HLA Class I Signal Sequence–derived Peptides by CD94/NKG2 Confers Protection from Natural Killer Cell–mediated Lysis. *Journal of Experimental Medicine*, **187**, 813–818.

Casewell,N. (2020) Solenodon genome reveals convergent evolution of venom in eulipotyphlan mammals (15 min). *Toxicon*, **177 Suppl 1**, S18.

Casewell,N.R. *et al.* (2013) Complex cocktails: the evolutionary novelty of venoms. *Trends Ecol. Evol.*, **28**, 219–229.

Cho,H.J. *et al.* (2019) Efficient Interleukin-21 Production by Optimization of Codon and Signal Peptide in Chinese Hamster Ovarian Cells. *J. Microbiol. Biotechnol.*, **29**, 304–310.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Cole,T.J. and Brewer,M.S. (2019) TOXIFY: a deep learning approach to classify animal venom proteins. *PeerJ*, **7**.

Datta,R. *et al.* (2007) Signal sequence mutation in autosomal dominant form of hypoparathyroidism induces apoptosis that is corrected by a chemical chaperone. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 19989–19994.

Estrada,G. *et al.* (2007) Spider venoms: a rich source of acylpolyamines and peptides as new leads for CNS drugs. *Nat. Prod. Rep.*, **24**, 145–161.

Fry,B.G. *et al.* (2010) Novel venom proteins produced by differential domain-expression strategies in beaded lizards and gila monsters (genus Heloderma). *Mol. Biol. Evol.*, **27**, 395–407.

Fry,B.G. *et al.* (2009) The Toxicogenomic Multiverse: Convergent Recruitment of Proteins Into Animal Venoms. *Annual Review of Genomics and Human Genetics*, **10**, 483–511.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Futatsumori-Sugai,M. and Tsumoto,K. (2010) Signal peptide design for improving recombinant protein secretion in the baculovirus expression vector system. *Biochem. Biophys. Res. Commun.*, **391**, 931–935.

Gacesa,R. *et al.* (2016) Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions. *PeerJ Computer Science*, **2**, e90.

Gupta,S. *et al.* (2013) In silico approach for predicting toxicity of peptides and proteins. *PLoS One*, **8**, e73957.

Hegde,R.S. and Bernstein,H.D. (2006) The surprising complexity of signal sequences. *Trends Biochem. Sci.*, **31**, 563–571.

von Heijne,G. (1983) Patterns of amino acids near signal-sequence cleavage sites. *Eur. J. Biochem.*, **133**, 17–21.

von Heijne,G. (1985) Signal sequences. *Journal of Molecular Biology*, **184**, 99–105.

von Heijne,G. (1990) The signal peptide. *J. Membr. Biol.*, **115**, 195–201.

Hiu,J.J. and Yap,M.K.K. (2020) Cytotoxicity of snake venom enzymatic toxins: phospholipase A2 and l-amino acid oxidase. *Biochem. Soc. Trans.*, **48**, 719–731.

Hunter,J.D. (2007) Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, **9**, 90–95.

Jungo,F. *et al.* (2012) The UniProtKB/Swiss-Prot Tox-Prot program: A central hub of integrated venom protein data. *Toxicon*, **60**, 551–557.

Karyolaimos,A. *et al.* (2019) Enhancing Recombinant Protein Yields in the E. coli Periplasm by Combining Signal Peptide and Production Rate Screening. *Frontiers in Microbiology*, **10**.

King,G.F. (2011) Venoms as a platform for human drugs: translating toxins into therapeutics. *Expert Opin. Biol. Ther.*, **11**, 1469–1484.

Kirkpatrick,S. *et al.* (1987) Optimization by Simulated Annealing. *Readings in Computer Vision*, 606–615.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Lacerda,A.F. *et al.* (2014) Antifungal defensins and their role in plant defense. *Front. Microbiol.*, **5**, 116.

Li,L. *et al.* (2018) Snake Venoms in Cancer Therapy: Past, Present and Future. *Toxins*, **10**, 346.

Lomonte,B. and Rangel,J. (2012) Snake venom Lys49 myotoxins: From phospholipases A(2) to non-enzymatic membrane disruptors. *Toxicon*, **60**, 520–530.

McKinney,W. (2010) Data structures for statistical computing in python. In, *Proceedings of*

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.30.405613; this version posted March 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the 9th Python in Science Conference. Austin, TX, pp. 51–56.

Naamati,G. *et al.* (2010) A predictor for toxin-like proteins exposes cell modulator candidates within viral genomes. *Bioinformatics*, **26**, i482–8.

Nielsen,H. and Krogh,A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 122–130.

Owji,H. *et al.* (2018) A comprehensive review of signal peptides: Structure, roles, and applications. *Eur. J. Cell Biol.*, **97**, 422–441.

Pedregosa,F. *et al.* (2011) Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, **12**, 2825–2830.

Peng,C. *et al.* (2019) Factors Influencing Recombinant Protein Secretion Efficiency in Gram-Positive Bacteria: Signal Peptide and Beyond. *Front Bioeng Biotechnol*, **7**, 139.

Petersen,T.N. *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, **8**, 785–786.

Samy,R.P. *et al.* (2017) Animal venoms as antimicrobial agents. *Biochemical Pharmacology*, **134**, 127–138.

Savojardo,C. *et al.* (2018) DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics*, **34**, 1690–1696.

Stotz,H.U. *et al.* (2009) Plant defensins: defense, development and application. *Plant Signal. Behav.*, **4**, 1010–1012.

Undheim,E.A.B. and Jenner,R.A. (2021) Phylogenetic analyses suggest centipede venom arsenals were repeatedly stocked by horizontal gene transfer. *Nat. Commun.*, **12**, 818.

UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.

Virtanen,P. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.

Waskom,M. *et al.* (2020) mwaskom/seaborn: v0.10.0 (January 2020). http://dx.doi.org/10.5281/zenodo.3629446

Whittington,C.M. *et al.* (2008) Defensins and the convergent evolution of platypus and reptile venom genes. *Genome Res.*, **18**, 986–994.

Wong,E.S.W. *et al.* (2013) SVM-based prediction of propeptide cleavage sites in spider toxins identifies toxin innovation in an Australian tarantula. *PLoS One*, **8**, e66279.

Wu,J. *et al.* (2016) Overexpression of a Pathogenesis-Related Protein 10 Enhances Biotic and Abiotic Stress Tolerance in Rice. *Plant Pathol. J.*, **32**, 552–562.

Xiao,H. *et al.* (2017) Snake Venom PLA, a Promising Target for Broad-Spectrum Antivenom Drug Development. *Biomed Res. Int.*, **2017**, 6592820.