Annotating eukaryotic and toxin-specific signal peptides using Razor

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, optimising recombinant protein expression, as well as for disease diagnosis and metagenomics.
Results: Here we explore the features of these signal sequences across eukaryotes. We find that different kingdoms have their characteristic distributions of signal peptide residues. Additionally, the signal peptides of secretory toxins have common features across kingdoms. We leverage these subtleties to build Razor, a simple yet powerful tool for annotating signal peptides, which additionally predicts toxin- and fungal-specific signal peptides based on the first 23 N-terminal residues. Finally, we demonstrate the usability of Razor by scanning all reviewed sequences from UniProt. Indeed, Razor is able to identify toxins using their signal peptide sequences only. Strikingly, we 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.
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 to 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). Several types of secretory pathways and their corresponding SPs, with different features and functions, have evolved across organisms (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 (Heijne and von Heijne, 1985; von Heijne, 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).

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). 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).

Machine learning methods such as support vector machines, hidden Markov models, and artificial neural networks are widely used for the prediction of SPs (Hiller *et al.*, 2004; Käll *et al.*, 2007; Viklund *et al.*, 2008; Nugent and Jones, 2009; Bagos *et al.*, 2010; Petersen *et al.*, 2011; Tsirigos *et al.*, 2015). Recent tools such as DeepSig and SignalP 5.0 use sophisticated architectures like convolutional neural networks (CNN) and recurrent neural networks (RNN), making them exceptionally accurate (Savojardo *et al.*, 2018; Almagro Armenteros *et al.*, 2019). However, the complexity of the algorithms behind these methods often obscures the underlying biology. Homology-based methods are more transparent and their performance is comparable with other methods (Frank and Sippl, 2008). A drawback of these is that they disregard the consensus-less property of SPs and the performance may drop significantly if the input sequence is not homologous to the training sequences (Almagro Armenteros *et al.*, 2019). Furthermore, 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; T. Jeffrey Cole, 2019). Moreover, these methods are based on the properties of the mature peptides (propeptides), rather than the SPs.

To address these issues, we first examine the features of SPs across the major eukaryotic kingdoms (Metazoa, Fungi, and Viridiplantae) and toxins. We then exploit those features to build Razor, a new tool for annotating SPs. Despite its simplicity, the performance metrics of Razor's SP prediction either outperformed or comparable to the tools built using complex architectures such as CNN and RNN. Furthermore, to assist with metagenomic studies, Razor also incorporates fungal SP prediction. This differs from the tools such as EffectorP because these tools are designed to identify effectors, a special class of secretory proteins, which often harbours a SP (Sperschneider *et al.*, 2018). We have optimised the command-line version of Razor for high-throughput analysis and used it to predict new SPs by scanning all reviewed sequences from UniProt (UniProt Consortium, 2019). In particular,

we predicted four new fungal SPs and novel toxins/defensive proteins across eukaryotes. Importantly, this prediction is possible just by utilising the first 23 residues of these SPs, without requiring the mature peptide itself, as validated by the protein family annotations.

MATERIALS AND METHODS

Datasets

We retrieved the training/benchmarking dataset from 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). These sequences were labelled as either SP or NO_SP (with or without a signal peptide). 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 for harbouring functional SPs. These sequences were used to build a generic SP classifier. For feature analysis, we clustered these sequences (60 N-terminal residues) at an identity threshold of 70% by kingdoms using CD-HIT v4.8.1 (Fu *et al.*, 2012). Only the representative sequences were retained to reduce sequence redundancy (Supplementary Table 1).

To annotate fungal SPs, we built a separate dataset using fungal SPs as a positive set. The SPs from other eukaryotic kingdoms were used as a negative set. Similarly, to build a classifier specialised in annotating toxin-specific SPs (SPT), we manually curated a separate positive set using the datasets from SignalP 5.0 and the animal toxin annotation project (Jungo *et al.*, 2012). Other SPs were assigned as a negative set (SPNT). We then clustered the fungal and toxin datasets as above and analysed the representative sequences (Supplementary Table 1).

The generic SP classifier was evaluated using the eukaryotic sequences from the benchmark set of SignalP 5.0, whereas fungal- and toxin-specific SP classifiers were evaluated using five-fold cross-validations. Furthermore, to demonstrate the usability of fungal- and toxin-specific SP classifiers, we used all reviewed fungal sequences (N=34,901) and all reviewed sequences (N=561,776) from UniProt, respectively (retrieved on 2 September 2020).

Bit score

The bit scores of the N-terminal residues in Metazoa, Fungi, Viridiplantae, and toxins were computed as:

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

For Metazoa, Fungi, and Viridiplantae, the positive set and the background set were SP and NO_SP, respectively. For toxins, the positive set and the background set were SPT and SPNT, 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 includes the Solubility-Weighted Index (SWI; (Bhandari *et al.*, 2020).

Generic SP classifier

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, S, T, and G) of the first 30 N-terminal residues. The residues were chosen such that they maximised the median 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 (Kyte and Doolittle, 1982) of each residue 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 the aligned set encoded using the optimised weight matrix. However, a limitation of our approach is that we are unable to detect the correct cleavage site if it is located before the 15th position.

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 shows the highest MCC on the benchmark set.

Toxin- and fungal-specific SP classifiers

We also built models specialised in annotating the SPs of toxins and fungal proteins. The toxin-specific SP classifier was built by taking the features hydrophobicity, SWI, flexibility, and turn. These features were selected such that they maximised the median MCC using five-fold cross-validations. In contrast, the fungal-specific SP classifier was built using counts per residue. The N-terminal lengths of 23 and 22 were found to generate the maximum median MCC scores for toxin- and fungal-specific SP classifiers, respectively (MCC=0.741 and 0.506, respectively, see also Supplementary Table 2). Similar to the SP prediction models, both the toxin- and fungal-specific SP classifiers consist of five models each.

Performance measures

We use MCC as a measure of performance to correctly identify 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 SP using the generic SP classifier, as well as the toxin- and fungal-specific SP classifiers. 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 1 and 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, 2019). Our web tools assist users in annotating SPs and protein domains, and making the decisions from gene cloning to protein expression and purification.

Statistical analysis

Data analysis was performed using pandas v1.0.3 (McKinney and Others, 2010). Hydrophobicity and SWI were smooth 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_2020. The source code for our SP annotation server (Razor) can be found at https://github.com/Gardner-BinfLab/TISIGNER-ReactJS.

RESULTS

The SPs of fungal secretory proteins and toxins have distinct sequence properties

We investigated the sequence composition of SPs by first aligning the sequences from the N-termini (Fig 1A, 2A) or by centering at the cleavage sites (Fig 1B, 2B), followed by computing bit scores for each amino acid residue. These approaches provide sufficient leverage to enumerate the tripartite domains of SPs (N-, H-, and C-domains). In general, hydrophilic residues are enriched towards the N-termini, which is a characteristic feature of SPs, i.e. the H-region (von Heijne, 1990). Notably, we identified several distinct patterns of SPs from Metazoa, Fungi, and Viridiplantae. More interestingly, the SPs of toxins also show some unique patterns (Fig 2). Therefore, we aimed to build generic and specific SP classifiers based on these features.

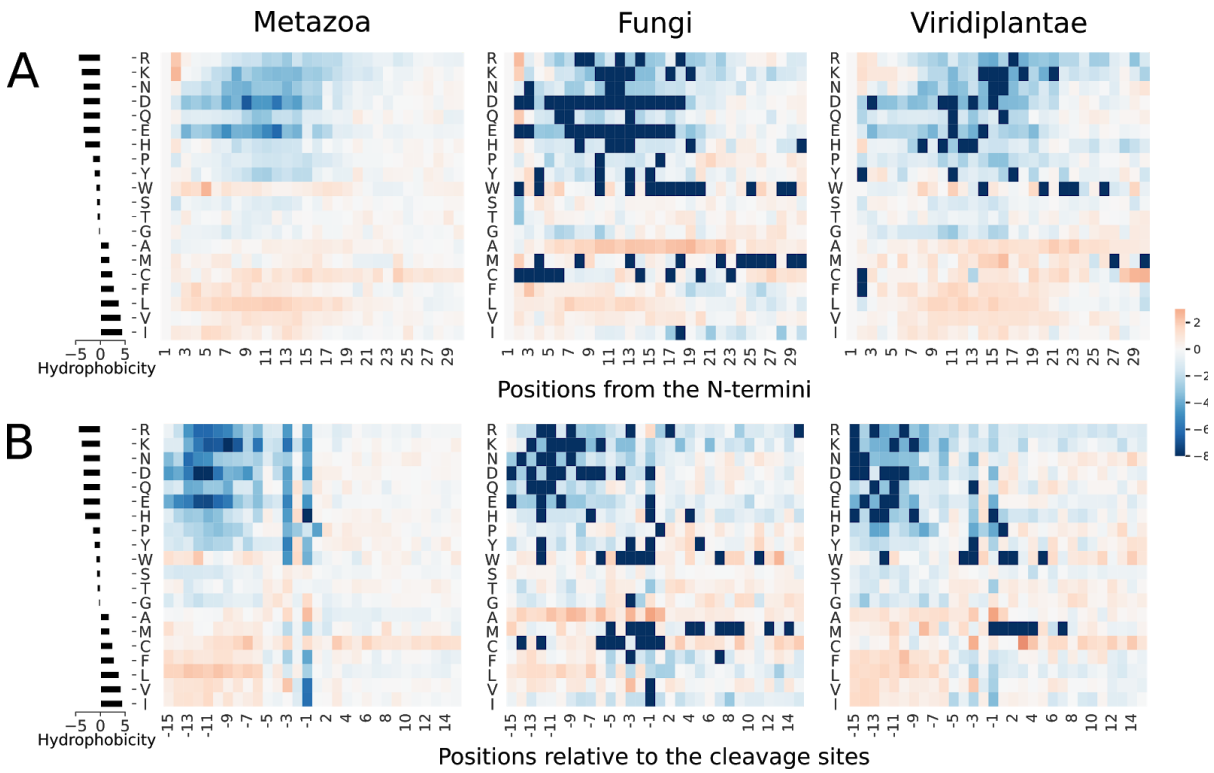

Fig 1. Amino acid enrichment in SPs across Metazoa (SP=1,594, NO_SP= 6,321), Fungi (SP=121, NO_SP=4,000) and Viridiplantae (SP=220, NO_SP=2,754). All groups show a strong hydrophobic property at the N-termini. The bar plots adjacent to the heatmaps show the Kyte and Doolittle's hydrophobicity scale of residues. Enrichment of residues as shown by bit scores for eukaryotic kingdoms when aligned at the **(A)** N-termini and **(B)** cleavage sites. The (-3, -1) rule for the cleavage site motif is visible in all groups when aligned at the cleavage sites. SP, Signal Peptide; NO_SP, Not a SP.

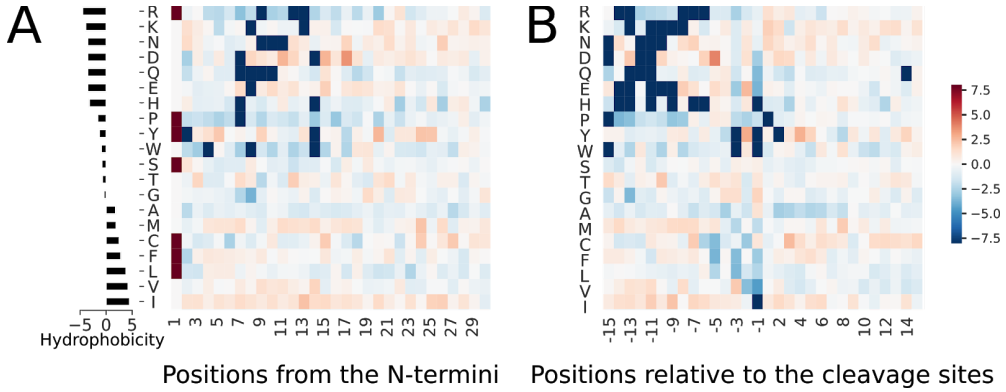

Fig 2. Amino acid enrichment in toxin-specific SPs in contrast to other eukaryotic SPs (SPT=261, SPNT=1,738). The distribution of residues in toxin-specific SPs when aligned at the **(A)** N-termini and **(B)** cleavage sites. SP, Signal Peptide; SPT, Toxin-specific SP; SPNT, Non-toxin SP.

The N-termini of SP and NO_SP sequences differ significantly in terms of residue composition and features (Fig 1A and 2A). These features can be recognised by the signal recognition particle, which binds to the SP and targets the translocon (Akopian *et al.*, 2013).

By aligning the sequences from the N-termini, we noted positively charged residues such as Arginine (R) and Lysine (K) that constitutes the N-domain (von Heijne, 1990; Guo *et al.*, 2018). A consecutive stretch of hydrophobic residues forms the H-domain, which probably is the most distinctive feature of SPs (Yarimizu *et al.*, 2015; Owji *et al.*, 2018). These characteristic distributions of residues are also supported by the AUC analysis of the abundance of each residue within all possible subsequences of the first 60 N-terminal residues (Supplementary Fig S1A, 2A, 3A, and 4A). However, the residue composition of this domain is remarkably different between the kingdoms (Fig 1A). In particular, the SPs of Metazoa and Viridiplantae have a preference towards Leucine (L), whereas the SPs of Fungi show a higher enrichment of Alanine (A). In contrast, the SPs of toxins show a strong abundance of Isoleucine (I) and are almost completely devoid of L and A (Fig 2A).

Next, we studied the protein sequences features within the first 60 N-terminal residues, i.e., GRAVY, structural flexibility, helix, sheet and turn propensities, instability index, aromaticity, isoelectric point, and SWI. Indeed, we observed several unique sequence properties across the different groups, in which GRAVY, flexibility and SWI showed the highest AUC scores (Supplementary Fig S1B, S2B, S3B, and S4B). Interestingly, we also found a higher AUC score in isoelectric point for the toxin dataset.

The cleavage sites mark the end of SPs and the beginning of the propeptide, which is a unique feature of SPs (Fig 1B and 2B). By aligning the sequences at the cleavage sites, we observed a clear emergence of (-3,-1) rule around the cleavage as a distinctive presence of small and charged residues such as A and Valine (V) (von Heijne, 1983). Notably, there's also an occasional presence of Threonine (T) and Glycine (G). The helix breaker, Proline (P), is absent at positions -3 to + 1, which has also been noted previously (von Heijne, 1983; Choo and Ranganathan, 2008). Polar residues such as Asparagine (N) and Cysteine (C) are sparsely present throughout all three domains, although their presence can be seen even after the cleavage sites and into the propeptides.

Razor shows comparable performance to the state-of-the-art programs

We made use of these important features of SPs and their cleavage sites to build SP classifiers and annotate eukaryotic SPs and fungal- and toxin-specific SPs. A recently published method SignalP 5.0 has benchmarked most of the publicly available methods for the prediction of SP. We used the eukaryotic sequences from their benchmarking dataset to compare our method with the previously benchmarked methods. Our method achieved MCC of 0.831, whereas our Cleavage Site precision and recall were 0.607 and 0.578 (Fig 3, CS precision and recall; see also Supplementary Fig S5 and Table S3 and S4). Our CS precision, in particular, increases significantly when the window size around the cleavage site is increased by up to 3 residues. This less restrictive definition of cleavage site might be useful in the cases where the cleavage sites have been incorrectly annotated (Almagro Armenteros *et al.*, 2019).

For fungal- and toxin-specific SP classifiers, we used five-fold cross-validations to evaluate the performance (Supplementary Fig S6 and S7). The median MCC scores for fungal- and toxin-specific SP classifiers across the five-fold cross-validations were 0.506 and 0.741, respectively.

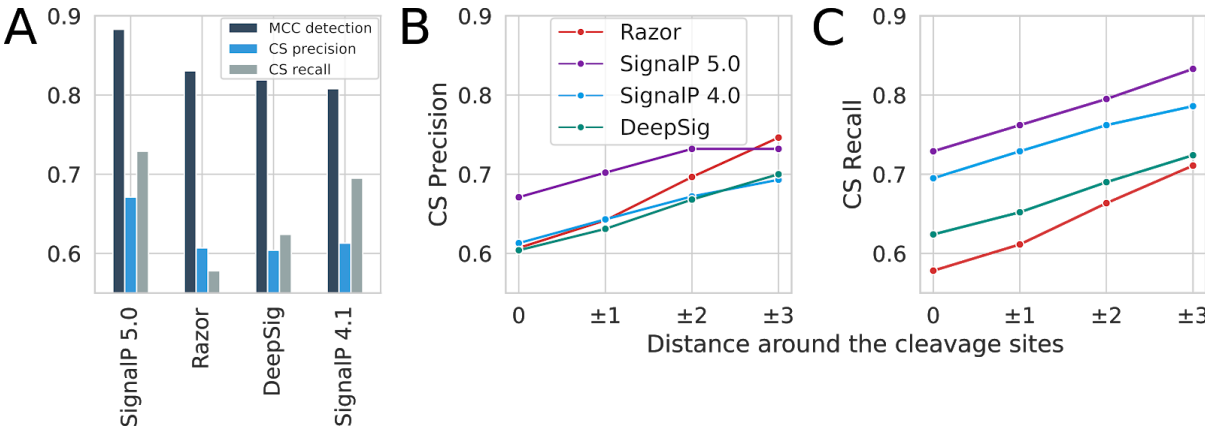

Fig 3. Benchmarking of SP prediction methods (A) MCC, CS precision and recall. **(B, C)** CS precision and recall of tools when different window sizes around the cleavage sites are used. Benchmarking is done using the eukaryotic sequences from the benchmark set of SignalP 5.0 (SP=211, NO_SP=7,246). MCC, CS precision, CS recall for SignalP 5.0, DeepSig, and SignalP 4.1 were taken from Almagro Armenteros et al. (2019). Data in Supplementary Tables 3 and 4. MCC, Matthew's Correlation Coefficient; CS, Cleavage Site; SP, Signal Peptide; NO_SP, Not a SP.

Defensive proteins could harbour a toxin-like SP

The toxin prediction model of Razor uses toxin-specific SPs with experimental evidence. A large proportion of our training sequences contained the SPs of animal toxins such as snake three-finger toxins, scorpion toxins, and phospholipase A2, followed by the SPs of ribosome-inactivating proteins from plants (Fig 4A). The features of these SPs were used to build models for the prediction of SPT in Razor.

To further assess our new toxin-specific SP classifier, we scanned all reviewed sequences in UniProt (N=561,776). A total of 910 sequences were predicted positive from all five SP detection models and toxin detection models. We further excluded false positive hits from bacteria (N=47), archaea (N=3), bacteriophages (N=3), and computationally annotated transmembrane proteins by Uniprot (N=8). The remaining sequences were divided into two groups based on the presence or absence of toxin annotation. From these probable SPT, 759 sequences had annotations for toxins (Supplementary Table S5). They included protein families such as scorpion toxin, phospholipase A2 and ribosome-inactivating protein (Fig 4B). The remaining 90 sequences had no annotations for toxins. These sequences were clustered at an identity threshold of 70%, which gave rise to 84 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 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 (*Crotalus, Pogona*) and mammalia (*Ornithorhynchus*) (Fry *et al.*, 2009, 2010; Whittington *et al.*, 2008). This suggests why their SPs show some remote similarity with toxin-specific SPs.

We also used the fungal SP classifier from Razor to scan all reviewed fungal proteins (N=34,901). We found four unannotated fungal SPs. Our results matched with the predictions from the most recent SP prediction tools SignalP 5.0 and DeepSig, except for a different cleavage site predicted by DeepSig (Supplementary Table S6).

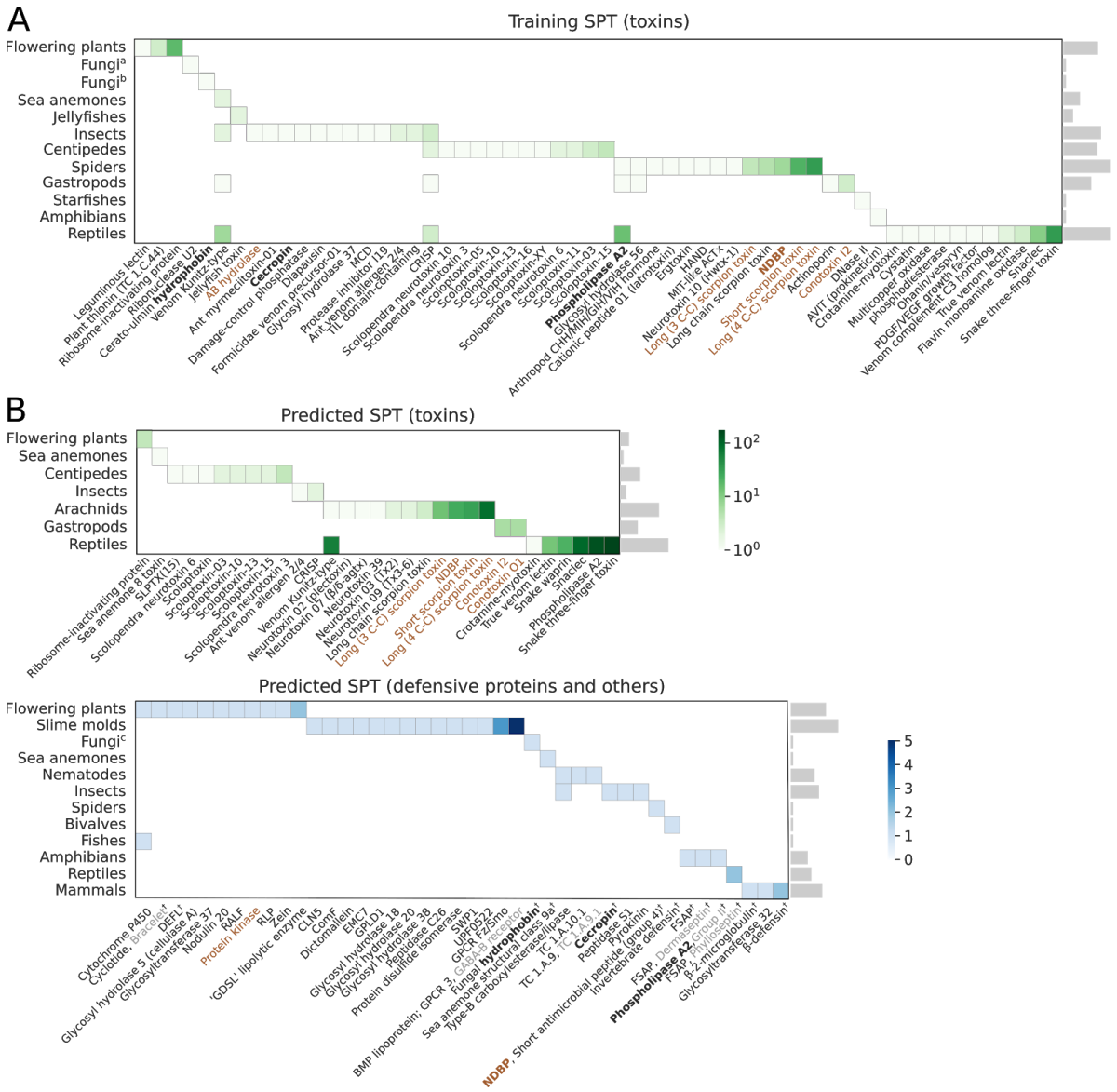

Fig 4. Razor identifies toxin-specific SPs along with several defensive proteins. All reviewed sequences from UniProt were examined (N=561,776). **(A)** Heatmap shows the abundance of protein families in the training SPT sequences by taxa. A total of 237 of 261 training SPT sequences had protein family annotations. **(B)** Heatmaps show the abundance of protein families in the predicted SPT sequences, with and without toxin annotations, by taxa (top and bottom panels, respectively). A total of 753 of 759 toxins in predicted SPT sequences had protein family annotations (top panel), whereas 53 of 90 remaining predicted SPT sequences had protein family annotations (bottom panel). Protein families in the bottom panel that show some defensive properties are marked with †. Protein subfamily, family and superfamily are shown in grey, black and brown, respectively. Protein families that are in common between the training and predicted SPT sequences are bolded (bottom panel).

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 across the major eukaryotic kingdoms. While SPs share a common hydrophobic nature, we have found several differences in their residue composition and consequently the sequence properties. We have used these features to build Razor for annotating eukaryotic SPs, which have specialised functionalities in annotating fungal- and toxin-specific SPs. Razor's performance was comparable to the other sophisticated methods. Using our tool, we have predicted four novel fungal SPs as well as several probable toxins, which are yet to be annotated, using the reviewed sequences from UniProt (Fig 4). Our predicted results consist of toxins/defensive proteins from diverse species, which gives us an overview of the source of toxins.

Since toxins/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 S8). 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; Pathogenesis-related proteins and peptides as promising tools for engineering plants with multiple stress tolerance, 2018; Boccardo *et al.*, 2019). Similarly, the cytotoxic activity of phospholipase A2 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 was able to identify previously unannotated SPs and a spectrum of eukaryotic toxins/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. Other related tools such as EffectorP are restricted to identifying effectors and are not suitable for whole-proteome scans (Sperschneider *et al.*, 2016). In contrast, Razor is designed to scan large datasets and therefore applicable to genomic, metagenomic and proteomic studies. Razor might also be useful in other research areas such as recombinant protein expression, toxicology, transgenics, and drug design.

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 her feedback on the Razor web application and figures.

FUNDING
This work was supported by the Ministry of Business, Innovation and Employment, New Zealand [MBIE grant: UOOX1709].

CONFLICT OF INTEREST
None declared.

REFERENCES

Akopian,D. *et al.* (2013) Signal recognition particle: an essential protein-targeting machine. *Annu. Rev. Biochem.*, **82**, 693–721.
Almagro Armenteros,J.J. *et al.* (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.*, **37**, 420–423.
Bagos,P.G. *et al.* (2010) Combined prediction of Tat and Sec signal peptides with hidden Markov models. *Bioinformatics*, **26**, 2811–2817.
Bhandari,B.K. *et al.* (2019) Highly accessible translation initiation sites are predictive of successful heterologous protein expression. *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.
Choo,K.H. and Ranganathan,S. (2008) Flanking signal and mature peptide residues

influence signal peptide cleavage. *BMC Bioinformatics*, **9 Suppl 12**, S15.
Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
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.
Frank,K. and Sippl,M.J. (2008) High-performance signal peptide prediction based on sequence alignment techniques. *Bioinformatics*, **24**, 2172–2176.
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.
Guo,H. *et al.* (2018) Positive charge in the n-region of the signal peptide contributes to efficient post-translational translocation of small secretory preproteins. *J. Biol. Chem.*, **293**, 1899–1907.
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. (1990) The signal peptide. *J. Membr. Biol.*, **115**, 195–201.
Heijne,G. von and von Heijne,G. (1985) Signal sequences. *Journal of Molecular Biology*, **184**, 99–105.
Hiller,K. *et al.* (2004) PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.*, **32**, W375–9.
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.
Käll,L. *et al.* (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.*, **35**, W429–32.
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**.
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. and Others (2010) Data structures for statistical computing in python. In, *Proceedings of 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.
Nugent,T. and Jones,D.T. (2009) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**, 159.
Owji,H. *et al.* (2018) A comprehensive review of signal peptides: Structure, roles, and applications. *Eur. J. Cell Biol.*, **97**, 422–441.
Pathogenesis-related proteins and peptides as promising tools for engineering plants with multiple stress tolerance (2018) *Microbiol. Res.*, **212-213**, 29–37.
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.
Sperschneider,J. *et al.* (2016) EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytol.*, **210**, 743–761.
Sperschneider,J. *et al.* (2018) Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Mol. Plant Pathol.*, **19**, 2094–2110.
Stotz,H.U. *et al.* (2009) Plant defensins: defense, development and application. *Plant Signal. Behav.*, **4**, 1010–1012.
T. Jeffrey Cole,M.S.B. (2019) TOXIFY: a deep learning approach to classify animal venom proteins. *PeerJ*, **7**.
Tsirigos,K.D. *et al.* (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.*, **43**, W401–7.
UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
Viklund,H. *et al.* (2008) SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics*, **24**, 2928–2929.
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).
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.
Yarimizu,T. *et al.* (2015) Synthetic signal sequences that enable efficient secretory protein production in the yeast Kluyveromyces marxianus. *Microb. Cell Fact.*, **14**, 20.