






ScrepYard: An online resource for disulfide-stabilized tandem repeat peptides

Junyu Liu¹  | Michael Maxwell¹ | Thom Cuddihy^{2,3} | Theo Crawford¹ |
Madeline Bassetti² | Cameron Hyde^{2,4}  | Steve Peigneur⁵  | Jan Tytgat⁵  |
Eivind A. B. Undheim^{1,6}  | Mehdi Mobli¹

¹Centre for Advanced Imaging, The University of Queensland, St. Lucia, Queensland, Australia

²Queensland Cyber Infrastructure Foundation Ltd., The University of Queensland, St. Lucia, Queensland, Australia

³Centre for Clinical Research, The University of Queensland, St. Lucia, Queensland, Australia

⁴University of the Sunshine Coast, Maroochydore, Queensland, Australia

⁵Toxicology and Pharmacology, University of Leuven (KU Leuven), Leuven, Belgium

⁶Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway

Correspondence

Mehdi Mobli, Centre for Advanced Imaging, Building 57, Research Road, The University of Queensland, St. Lucia, QLD 4072, Australia.

Email: m.mobli@uq.edu.au

Eivind A. B. Undheim, Department of Biosciences, Centre for Ecological and Evolutionary Synthesis, University of Oslo, PO box 1066, Blindern, 0316 Oslo, Norway.

Email: e.a.b.undheim@ibv.uio.no

Funding information

Australian Research Council, Grant/Award Number: DP190101177; Fonds Wetenschappelijk Onderzoek, Grant/Award Numbers: 12W7822N, GOA4919N, GOC2319N, GOE7120N; Norges Forskningsråd, Grant/Award Number: 287462; Katholieke Universiteit Leuven, Grant/Award Number: PDM/19/164

Review Editor: Aitziber L. Cortajarena

Abstract

Receptor avidity through multivalency is a highly sought-after property of ligands. While readily available in nature in the form of bivalent antibodies, this property remains challenging to engineer in synthetic molecules. The discovery of several bivalent venom peptides containing two homologous and independently folded domains (in a tandem repeat arrangement) has provided a unique opportunity to better understand the underpinning design of multivalency in multimeric biomolecules, as well as how naturally occurring multivalent ligands can be identified. In previous work, we classified these molecules as a larger class termed secreted cysteine-rich repeat-proteins (SCREPs). Here, we present an online resource; ScrepYard, designed to assist researchers in identification of SCREP sequences of interest and to aid in characterizing this emerging class of biomolecules. Analysis of sequences within the ScrepYard reveals that two-domain tandem repeats constitute the most abundant SCREP domain architecture, while the interdomain “linker” regions connecting the functional domains are found to be abundant in amino acids with short or polar sidechains and contain an unusually high abundance of proline residues. Finally, we demonstrate the utility of ScrepYard as a virtual screening tool for discovery of putatively multivalent peptides, by using it as a resource to identify a previously uncharacterized serine protease inhibitor and confirm its predicted activity using an enzyme assay.

Junyu Liu and Michael Maxwell authors contributed equally.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

KEYWORDS

bioactive, bivalent, disulfide-rich, multivalent, peptide, SCREPs, secreted proteins, tandem-repeat

1 | INTRODUCTION

Multivalency is a common property of biomolecules that describes the interaction between two molecules through multiple nonoverlapping binding interfaces. The advantage of multivalency is two-fold (1) higher specificity due to a larger interaction interface, and (2) enhanced binding kinetics and thermodynamics that result in high avidity (Mammen et al., 1998; Vauquelin & Charlton, 2013). Nowhere is this better recognized than in the adaptive immune system where antibodies use multivalency as a key mechanism in responding to infections through the dimeric nature of the antigen recognizing regions and the symmetry in the Y-shaped structure (Uvyn & De Geest, 2020). Mimicry of this process has resulted in the field of antibody therapeutics, which have had a tremendous impact on contemporary pharmaceutical development (Miller & Lanthier, 2015). Despite the success of antibody therapeutics, there are a number of limitations; these include a requirement of a good (unique and accessible) antigen, relatively poor thermal and chemical stability, and that antigen recognition may or may not lead to the desired (or any) functional outcome (Rodgers & Chou, 2016). Where antibodies are limited, small molecules often excel, with the caveat of poor selectivity that can potentially lead to serious side-effects. Peptides offer an attractive middle ground, providing higher specificity than small molecules due to their larger binding interface whilst being as functionally potent as antibodies. Indeed, peptides have received substantial attention over the past few decades, demonstrating an exceptional capacity for use as molecular probes which target many therapeutically relevant biomolecules (Dutertre & Lewis, 2010; Muttenthaler et al., 2021; Pennington et al., 2018). They are also increasingly being developed into novel therapeutics, with approximately 80 peptide drugs now approved for use, and over 150 peptides currently undergoing clinical trials (Muttenthaler et al., 2021).

Disulfide-rich peptides (DRPs) have emerged as a particularly attractive class of peptides due to their covalent intramolecular disulfide bonds. These bonds act as cross-braces to increase structural stability and backbone rigidity, resulting in resistances to proteolysis and extreme physicochemical conditions (i.e., extremes of pH and temperature) (Gongora-Benitez et al., 2014). The majority of DRPs characterized to date are highly potent neurotoxins isolated from animal venoms and consist of a

single domain (Mobli et al., 2017). However, the therapeutic potential of many potent single domain DRPs are limited due to poor selectivity. For example, the analgesic potential of several voltage-gated sodium channel inhibitors is overshadowed by their effect on other physiologically crucial ion channels (Deuis et al., 2017; Zhang et al., 2018). Interestingly, there are several reports of naturally occurring multi-domain DRPs that display a multivalent mode-of-action (Bohlen et al., 2010; Chassagnon et al., 2017; Guyot et al., 2020; Van de Locht et al., 1995; Van de Locht et al., 1996). All of these characterized multi-domain DRPs contain a tandem repeat (TR) architecture, where the individual domains share high internal sequence homology. Previous bioinformatics studies of these TR-DRPs revealed that they belong to the larger molecular class that we have defined as secreted cysteine-rich repeat proteins (SCREPs) (Maxwell et al., 2018).

To date, three venom derived TR-DRPs have been characterized in detail; including two spider derived ion channel modulating toxins; DkTx (τ -theraphotoxin-Hs1a; UniProtKB ID P0CH43) from *Cyriopagopus schmidti* (Bohlen et al., 2010) and π -hexatoxin-Hi1a (henceforth Hi1a; UniProtKB ID A0A1L1QJU3) from *Hadronyche infensa* (Chassagnon et al., 2017), and the serine protease inhibitor rhodniin (UniProtKB ID Q06684) from *Rhodnius prolixus* (Van de Locht et al., 1995). All three TR-DRPs use bivalency—simultaneously binding to two receptor sites—as a mechanism to enhance and prolong their pharmacological effects (Bohlen et al., 2010; Chassagnon et al., 2017; Van de Locht et al., 1995). The larger interaction interface observed in the bivalency of SCREPs (Bae et al., 2016; Gao et al., 2016) demonstrates their capacity for improved target selectivity compared to their single domain counterparts, such as the improved selectivity of DkTx compared with τ -theraphotoxin-Pc1b (Vanillotoxin-2; UniProt ID P0C245) (Bohlen et al., 2010). This provides an opportunity to leverage existing knowledge of venom derived DRPs in the search for peptides with higher specificity toward therapeutic targets. Additionally, the relatively slow dissociation rates of bivalent DRPs make them ideal molecular probes for studying channel structure (Gao et al., 2016).

However, despite their attractiveness, there is currently no resource designed for mining or browsing SCREPs. Common databases dedicated to sequence repeats often focus on genomic DNA sequences (Boby

et al., 2005; Gelfand et al., 2007; Le Fleche et al., 2001), or short amino acid repeats (Kalita et al., 2006). For example, PRDB (Jorda & Thierry, 2012) defines repeats as short periodic amino acid sequences that are directly adjacent to one another, while RepeatsDB (Paladin et al., 2017) uses structural information obtained from the Protein Data Bank to define protein repeats. Databases containing large numbers of DRPs such as ConoServer (Kaas et al., 2010) and the Knottin database (Postic et al., 2018) are likely to contain examples of SCREPs, but they do not include any curation relating to peptide domain organization (architecture). Here, we present SrepYard, an online database of SCREPs extracted using a refined and automated datamining pipeline. Open access to this database is provided to facilitate discovery and further investigation of SCREPs, extending the available resources for uncovering the underlying mechanisms that drive their fascinating multivalent activity.

2 | DATABASE CONSTRUCTION

The following section will outline the construction of the SrepYard database. This process is comprised of three distinct stages: SCREP datamining, SCREP architecture annotation, and the upload of data to SrepYard (Figure 1).

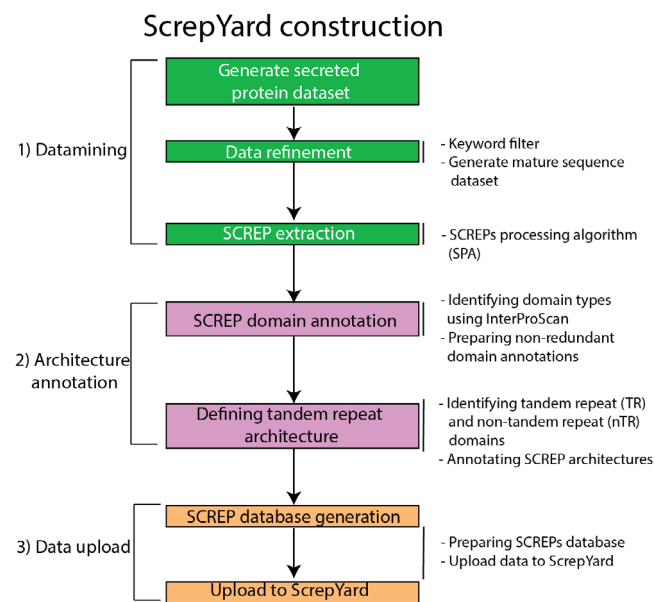


FIGURE 1 Flowchart outlining the construction of the SrepYard database. The process can be divided into three stages, (1) SCREP datamining (green), (2) SCREP architecture annotation (purple), and (3) the compiling and upload of SCREP data to SrepYard (orange). Key processes for each step shown on the right.

2.1 | Datamining

2.1.1 | Generating the initial dataset

Three datasets are downloaded from UniProtKB (Bairoch et al., 2005) (1) those that are manually curated (Swiss-Prot subset) and non-Swiss-Prot sequences (UniProt-TrEMBL) that contain either (2) the annotation “signal peptide” or (3) have a subcellular annotation of “secreted”. All three datasets are filtered to exclude sequences with subcellular location annotation of “intramembrane,” “topological domain,” and “transmembrane.” The outputs of the three filters are merged and used as the initial dataset.

2.1.2 | Data refinement and SCREP extraction

The initial dataset is subsequently refined by applying a keyword filter to remove known non-SCREPs based on their annotations. Currently the keyword list consists of “intracellular” (Maxwell et al., 2018), “disulfide isomerase” (Wilkinson & Gilbert, 2004), “double CXXCH motif” (Chivers et al., 1996), “ferredoxin” (Schurmann & Buchanan, 2008), “sulfur” (Baghshani & Abadi, 2014; Lill, 2009), “zinc” (Brandt et al., 2009), “iron” (Lill, 2009), “cytochrome” (Meunier et al., 2004), “thioredoxin” (Arner & Holmgren, 2000), and “dehydrogenase” (Brandt et al., 2009). This heuristic approach allows for continual optimization of the pipeline with the addition of new keywords as these are identified and ongoing updates to the database.

After the keyword filter, SignalP [v-5.0 (Armenteros et al., 2019)] is used to recognize and remove the signal peptide from each sequence in the dataset, generating mature protein sequences. In some cases, secreted proteins are sequenced from native material such as venom secretions, and do not contain a signal peptide region, for example, the spider toxin DkTx (UniProt ID P0CH43). Proteins that are not recognized by SignalP are directly grouped together with the mature sequences. Finally, the SCREP processing algorithm (SPA) is applied to remove all sequences that contain >500 or <20 amino acids (AAs) and sequences that contain <4 cysteine residues. The upper limit is set to avoid collecting much larger proteins, for example, transmembrane receptors, and the lower limit is set to avoid possible false identification of small non-domain repetitive elements. All remaining sequences are then processed to identify regions with internal sequence homology by use of an iterative BLAST function [see also Maxwell et al. (Maxwell et al., 2018)].

2.2 | Architecture annotation

2.2.1 | Generation of domain information

The dataset of extracted SCREP sequences requires further processing to accurately characterize each SCREP architecture including the specific domain types occurring in each SCREP, the order in which they appear, the sequence length of each domain, and the inter-domain linkers. The first step in SCREP annotation generates domain information. We utilize InterProScan (v-5.48), a consortium of several protein databases that predict domains using sequence-based recognition methods (Jones et al., 2014). As InterProScan consists of multiple databases, a single domain may be identified multiple times with slight differences in domain boundaries. To refine the InterProScan output data, the series of identified domains for each SCREP sequence is clustered by the database used, for example, Pfam, Prosite, and so on. In each cluster, identified regions are sorted according to their start and end positions. If an overlap exists between annotated domains, preference is given to the smallest recognized domain. The database-cluster with the highest number of recognized domains is then selected as the representative series of domain annotations for the SCREP candidate. If the number of domain annotations are identical, the database-cluster is selected according to a database preference list: Pfam (Mistry et al., 2021) > Prosite (Sigrist et al., 2010) > SMART (Letunic et al., 2021) > CDD (Lu et al., 2020) > SUPERFAMILY (Gough et al., 2001). After extracting the nonredundant domain annotations, each domain within a SCREP is numbered in sequential order according to its location from N- to C-terminus.

2.2.2 | Defining TR architecture

For each SCREP, a series of internal BLAST functions (default parameter, e-value <10) are performed between all identified domains to determine interdomain sequence homology. Domains are defined as TR if a BLAST alignment can be found between neighboring domains (e-value <10) or are deemed nonhomologous (e-values >10) and defined as non-TR (nTR) domains. After defining the number of domains and whether they are TR or nTR domains, the SCREP architecture is annotated according to the sequential order of TR / nTR domains, (i.e., all three domain SCREPs may be annotated as TR1-TR2-TR3, TR1-TR2-nTR3, and nTR1-TR2-TR3) distinguishing between all possible combinations of TR and nTR domains. Finally, the sequence

length of various SCREP elements including the N- and C-termini, the individual domains, and the interdomain linker regions are calculated based on the identified domain boundaries. In SCREPs containing more than one linker, that is, containing ≥ 3 domains, each linker is sequentially numbered in the same way as the domains described above. Finally, we note that our approach to generate SCREP architecture annotation relies on the use of InterProScan, and in instances where ordered regions are not recognized by this tool, no annotations are produced in the SrepYard output.

2.3 | Data upload

2.3.1 | SCREP database generation

To remove any duplicate SCREPs from SrepYard, CD-HIT [v-4.8.1 (Li et al., 2001)] is used with a threshold of 0.999. CD-HIT is only applied to sequences that originate from TrEMBL (Boeckmann et al., 2003) as they have not been manually curated and may contain errors resulting in sequence duplication and fragmentation. All manually curated SCREPs that originate from SwissProt (Boutet et al., 2016) are maintained without applying CD-HIT. All SCREP domain annotations and other relevant information, such as taxonomy and cysteine content, is compiled, formatted, and uploaded to SrepYard.

2.3.2 | SrepYard updates

The content in SrepYard is automatically updated every 3 months. For each update, all newly released and recently modified sequences from UniProtKB are processed. Any existing SCREPs that are found as new entries in the updated UniProtKB dataset are removed from SrepYard and re-processed (this is to account for any slightly modified SCREP sequences). The newly processed data are then merged with the existing SCREPs database. Previous database iterations are archived on the Nectar Research Cloud (Barker et al., 2019) for 1 year, after which archived data are stored on a local server at the Centre for Advanced Imaging, University of Queensland, Australia.

The SCREP recognition process relies heavily on existing third-party software, including blast+, InterProScan, SignalP, and CD-HIT. To ensure the accuracy of SCREP datamining and annotation, we also perform software updates as required. After any software updates, the entire SrepYard database is rebuilt.

3 | DATABASE UTILITY AND DISCUSSION

3.1 | Database content—SCREP architectures

In the latest update of SrepYard (Dec 2022), 183,518 sequences were identified as putative SCREPs from the total secreted protein dataset (18,791,263 sequences)—three times as many SCREPs as the previous published extraction (May 2018), which comprised 60,935 putative

SCREPs from 8,006,061 secreted protein sequences (Maxwell et al., 2018). The growth in number of sequences shows the remarkably rapid expansion of available sequences within UniProtKB, further emphasizing the need for automated processing tools to extract sequences of interest.

The data can be broadly broken down into two categories based on their putative domain annotations, “InterProScan-identified” (49.3%) and “unknown architecture” (50.7%) (Figure 2a). The large proportion of unknown domains reflect the abundance of

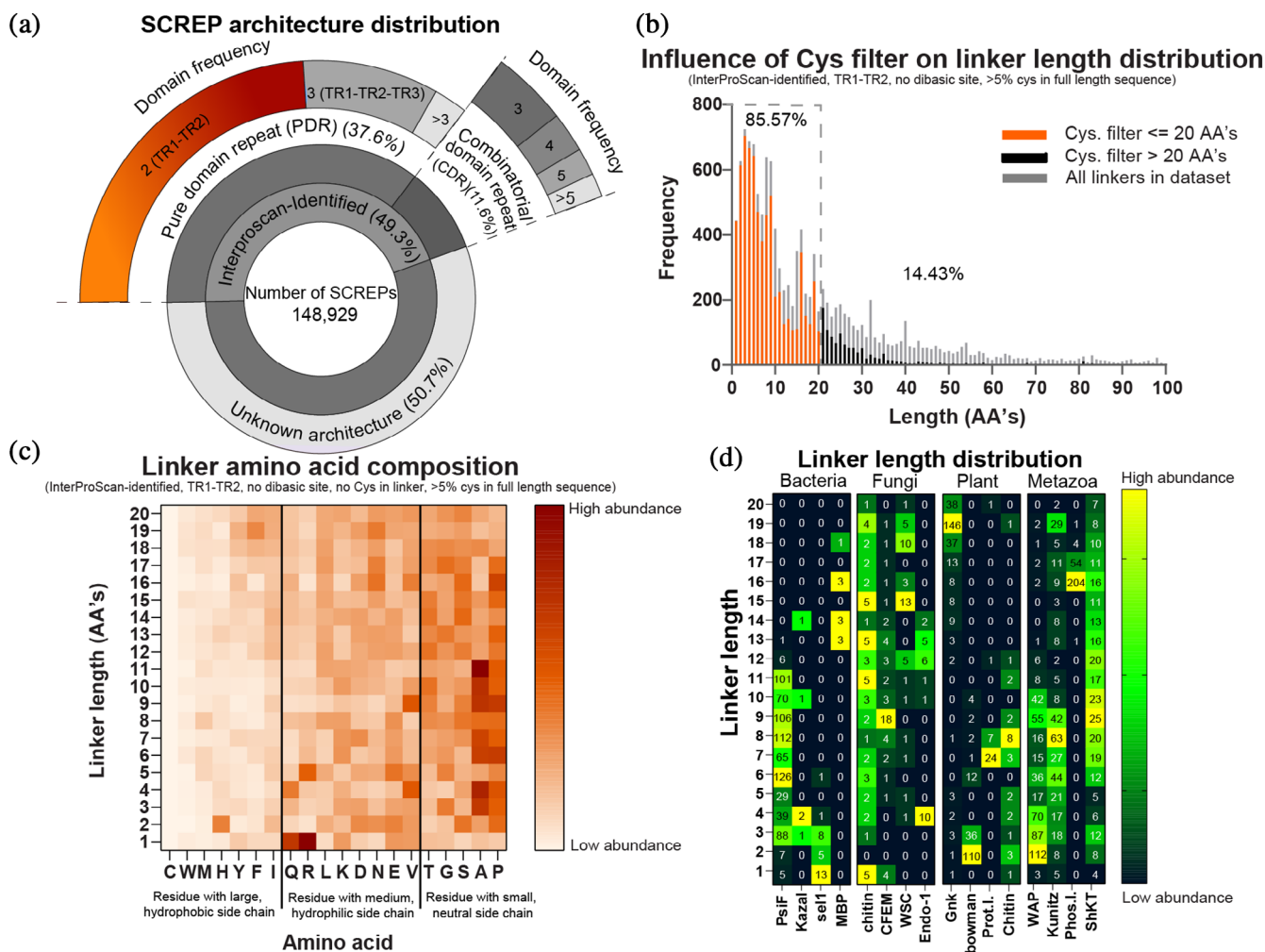


FIGURE 2 Distribution of SCREP architectures and linker analysis of two-domain SCREPs. (a) The inner circle demonstrates the two major clusters of SCREPs; “InterProScan-identified” (SCREPs with predicted domain types) and “unknown architectures” (SCREPs with unknown domain types). The outer circle demonstrates the different architecture types; unknown architectures (50.7%), pure domain repeats (PDR) (37.6%), and combinatorial domain repeats (CDR) (11.6%), dividing the PDR’s and CDR’s into a distribution based on the length of repeating domains. (b) The frequency distribution of linker lengths within the TR1-TR2 dataset. Most linkers are <20 AAs in length (85.57%), with the remaining linkers (14.43%) extending between 20 and 100 AAs. (c) A heat map of amino acid composition for all known two-domain SCREP for linker lengths between 1–20 AAs. AAs are sorted left-to-right in order of decreasing side hydrophobicity (Monera et al., 1995). (d) A grouped heatmap displaying the abundance of domain specific linker lengths in bacteria, fungi, plants, and metazoans. Each kingdom contains the four highest occurring domain types, with the frequency of each linker length displayed. The coloring indicates the relative level of abundance for each domain type

uncharacterized domain types within UniProtKB, sometimes referred to as the “dark proteome” (Perdigao et al., 2015). Within the “unknown architecture” dataset we find a taxonomic bias, where 69.5% of prokaryotic SCREPs contain unrecognizable domains compared to just 39.1% of eukaryotic SCREPs (Figure S1). Given the uncertainties associated with the “unknown architecture” dataset, we have restricted the below analysis of the database to the “InterProScan-identified” dataset only.

The “InterProScan-identified” dataset can be further broken down into two groups based on their architecture. A pure domain repeat (PDR) is defined as having an architecture that contains only TR domains (e.g., TR1-TR2 or TR1-TR2-TR3), while combinatorial domain repeats (CDRs) represent all other architectures (e.g., nTR1-TR2-TR3 or TR1-TR2-nTR3; see also Figure 2a). Within PDRs, the TR1-TR2 architecture is most common, accounting for 24.9% of the entire database. CDRs make a smaller fraction of the database accounting for 11.7% of all sequences. For both PDRs and CDRs the most abundant architectures are those with the fewest number of domains, and in general there is a decrease in the number of SCREPs with a certain architecture, as a function of decreasing number of domains within that architecture (Figure S2).

The cysteine density (percentage of cysteine residues) within a SCREP is a defining feature of this class of molecules, and in general we would expect that a higher cysteine density would correlate with disulfide-directed, and hence thermodynamically more stable, folds. Overall, we find that SCREPs from eukaryotic kingdoms have a higher cysteine density (Figure S3), consistent with the more sophisticated disulfide processing machinery in these higher organisms (Gruber et al., 2006; Van Anken & Braakman, 2005). Given the central importance of this feature, we have made it possible to directly refine search results within SrepYard by defining a minimum and maximum cysteine density. We note that while we have taken the inclusive approach of retaining any sequences within SrepYard with a potential domain repeat that contains a single disulfide bond, this does not necessarily satisfy the requirement of “cysteine-rich”. The cysteine-density filter, thus, allows the user to search or view a subset of the database (default values within the advanced search are >4 cysteines and >10% cysteine density).

3.2 | Database content—Linker analysis

An important yet poorly characterized aspect of multivalency is how multiple domains are linked together, and what effect the “linker” region has on binding and

function. The peptide linker is crucial in ensuring that each domain is positioned for optimal engagement with their molecular target (Bae et al., 2016; Bobrovnik, 2007; Klein et al., 2014; Mack et al., 2012). Elements of the linker, such as flexibility/rigidity and its effect on spatial positioning of the domains, play an important role in defining the intermolecular binding kinetics (Soler & Fortuna, 2017). Under evolutionary pressure, naturally occurring multivalent ligands have yielded linkers of a specific length that have a suitable amount of structural rigidity for enhanced target engagement (Bohlen et al., 2010; Chassagnon et al., 2017; Van de Locht et al., 1995; Van de Locht et al., 1996). These evolved linkers are consequently also likely to be dependent on the molecular target of the peptide. SrepYard has been designed to be enriched in sequences that contain multivalent ligands. Analysis of linker sequences in SrepYard may thus provide insights into the basic design principles that have emerged as a product of an evolutionary process in naturally occurring multivalent peptide ligands, thereby aiding rational engineering of synthetic multivalent peptides.

To further investigate the potential of SrepYard to provide insights into linker properties of multivalent peptide ligands, we selected a subset of SCREPs with a two-domain TR architecture (Figure 2a). We subsequently filtered this subset to remove any sequences that contain a cysteine residue in the inter-domain linker sequence as this may indicate incorrectly defined domain boundaries and/or unrecognized domain regions (Figure 2b). Although there may be some genuine cases of cysteine containing linkers, to verify this requires individual assessment of existing experimental data. Our preliminary analysis of two-domain SCREPs from SwissProt (Boutet et al., 2016) with linkers ≤ 20 AAs that contain a cysteine, reveal a total of 40 SCREPs. One interesting example where four cysteine residues form two disulfide bonds within the linker was observed for the double-antistatin like peptide (UniProt ID P15358) (Lapatto et al., 1997). This suggests that there may be a small population of SCREPs with functionally relevant cysteine residues within the linker region, warranting further analysis which extends beyond the scope of our current investigation. Additionally, SCREPs that are posttranslationally processed into two separate domains, via protease cleavage, were identified and removed from our analysis (2153 SCREPs in total). Sequences containing a dibasic site, that is, “KK,” “KR,” “RK,” “RR,” within the linker region indicate cleavage from subtilisin-like proprotein convertases (SPCs) (Rholam & Fahy, 2009). An example is human endothelin (UniProt ID P05305), which harbors two homologous endothelin-like domains. During posttranslation modification, the gene product is

cleaved by a Furin enzyme at a K91-R92 motif (between the two TR domains), yielding mature endothelin peptide and a second endothelin domain with unknown function (Turner & Murphy, 1996). This example highlights cases where TRs are posttranslationally cleaved to yield monovalent disulfide rich peptides, and as such contain linkers that do not contribute to multivalent binding.

Next, we restricted the data to proteins that had a cysteine density $\geq 5\%$, to enrich for potential disulfide-stabilized protein structures. This dataset is here simply referred to as the *two-domain SCREPs*. Within this dataset we found that the linker length (number of AAs) has an asymmetric Gaussian distribution with a maximum at approximately 10 AAs in length and the majority of peptides containing a linker between 1 and 20 AAs (85.57%) (Figure 2b). As the potential for the existence of an unrecognized domain within the linker increases with linker length (regardless of the presence of a cysteine), subsequent analysis of the amino acid composition (Figure 2c) and the distribution of linker lengths within various taxonomic groups and domain types (Figure 2d) was performed using a subset of peptides containing a linker of ≤ 20 AAs. The linkers of these two-domain SCREPs appear to consist primarily of amino acids with short or polar sidechains highly enriched in proline and alanine residues (Figure 2c). The observed linker composition of these SCREPs aligns with previous findings of AA occurrence within naturally occurring linker regions (Chen et al., 2013; George & Heringa, 2002).

The secondary structure prediction tool [MobiDBLite (Necci et al., 2017)] was then used to predict the presence of disorder within these linkers. Overall, we find that disordered linkers are more prevalent in SCREPs with a bacterial origin (12.3% of bacterial linkers compared with 0.2% of eukaryotic linkers) (Table S1). Although the exact functional purposes of these disordered linkers are unknown, their presence demonstrates natural variability of structural rigidity. We can only speculate that increased disorder would lead to lower avidity, or higher receptor promiscuity, which may reflect the differences observed between the prokaryotes and the more complex eukaryotic organisms.

Next, we investigated if there was a relationship between the linker length and the domain type. It is known that some DRP domain types are associated with specific functions (e.g., protease-inhibiting Kunitz domains). In these cases, if the second domain has evolved to bind to a common and adjacent receptor site, there may be evolutionary pressure to restrict the length and composition of the interdomain linker (Handl et al., 2007; Tran et al., 2020). In this analysis we find three general patterns, (1) domains with a broad distribution of linker lengths, (2) domains that have either short

or longer linkers, or (3) domain types with a highly conserved linker length (with a sharp distribution, i.e., length ± 1 residue). Examples of the three types are as follows:

1. The fungal chitin domains and the metazoan ShKt domain types appear to have a broad distribution of linker lengths between 1–20 AAs.
2. The PsiF bacterial domains, and the metazoan Kunitz and WAP domains appear to favor shorter linker lengths (<12 AA's), while the Gnk2 plant domain has a cluster of linkers with a longer length (>14 AAs).
3. Domains with highly conserved linker lengths include; the CFEM fungal domain, the short (1–3 AA linkers) bacterial sell-like repeats, the 2-residue linkers in Bowman-Birk plant domain, the 7-residue linkers in proteinase inhibitor plant domain (Prot.I.), and the 16-residue linkers in phospholipase inhibitor (Phos.I.) domain.

Examples of where a correlation between linker length and molecular target may exist can also be found in the bivalent serine-protease inhibitors rhodniin (a Kazal-type SCREP; UniProt ID Q06684) and ornithodorin (a peptide with two Kunitz domains with low sequence similarity; UniProt ID P56409). Despite their domains being structurally different, both rhodniin and ornithodorin bind to the same two regions of thrombin and have very similar linker lengths of 9 and 10 AAs, respectively (Van de Locht et al., 1995; Van de Locht et al., 1996). Therefore, we speculate that in some circumstances linker length may be indicative of molecular target (in this case more so than the 3D structure of the individual domains). Domain types with broad linker-length distributions may indicate that these domains have undergone functional divergence, interacting with structurally diverse targets. Conversely, the highly conserved lengths such as that observed within the phospholipase inhibitor domain (Phos.I.), suggest interactions with either a limited number of molecular targets, or a family of targets with a high degree of structural similarity. Evidently, the elucidation of correlations between linker length and molecular target may serve as a powerful method in discovering novel multivalent ligands of known receptors.

3.2.1 | Identifying bioactive SCREPs

In addition to annotating the SCREP architectures of natural multivalent peptides, ScrepYard has been devised to aid researchers to mine SCREP sequences to identify multivalent versions of their well-characterized single-

domain counterparts. Our approach relies on the observation that the individual domains of two-domain bivalent SCREPs reported to date align well with existing single domain DRPs (Bohlen et al., 2010; Chassagnon et al., 2017). In addition, evidence suggests DRPs that target the same receptor tend to convergently evolve similar primary structures (Undheim et al., 2016), meaning that within a fold type, there is a high probability that a SCREP with a particular function shares a relatively high degree of sequence similarity with a single-domain DRP with the same function. For example, there is high sequence identity between the single-domain PcTx-1 isolated from the venom of the spider *Psalmopoeus cambridgei* (Escoubas et al., 2003) and the two-domain SCREP Hi1a isolated from the distantly related spider *Hadronyche infensa* (Chassagnon et al., 2017) (71% and 56% sequence identity with TR1 and TR2 of Hi1a, respectively). Both PcTx-1 and Hi1a have been confirmed to modulate the acid sensing ion channel 1a (ASIC1a) (Berkut et al., 2015; Chassagnon et al., 2017; Escoubas et al., 2003), with Hi1a exhibiting higher avidity than PcTx-1 due to a bivalent mode-of-action (Chassagnon et al., 2017). To apply this evolution-guided mining approach, we propose that the wealth of functional data available for single-domain DRPs [such as those curated in ToxProt (Jungo et al., 2012)] may serve as an ideal starting point to identify SCREPs with a putative multivalent mode-of-action.

As proof of principle, we employed a batch sequence analysis method aimed at identifying toxins with known activity that share sequence identity with SCREPs. A dataset of experimentally validated bioactive toxins was extracted from the ToxProt (Jungo et al., 2012) database, using this as a set of query sequences we performed a BLAST search between the known toxins and the SCREPs database. Using this method, we identified 9325 SCREPs which display varying degrees of sequence similarity with known single domain toxins. From these data, we selected the single-domain DRP Kaliclude-3, a dual-function toxin isolated from the sea anemone *Anemonia sulcata* (UniProt ID Q9TWF8) that inhibits trypsin—a serine protease from the PA clan superfamily—and voltage-sensitive potassium channels (Schweitz et al., 1995). Kaliclude-3 was subsequently used as a query sequence to further demonstrate the utility of the SrepYard BLAST search.

The output shows that Kaliclude-3 has high sequence homology with d-Gs1a; a putative double domain SCREP from the marine gastropod *Gemmula speciosa* (UniProt ID A0A098LW49) (Figure 3a). Thus, to determine if d-Gs1a shares the same bioactivity as Kaliclude-3, a d-Gs1a gene was synthesized and cloned into an *E. coli* expression vector for recombinant production (Figure S4). Following successful production, we used NMR spectroscopy to assess the folding of the peptide, and found a highly dispersed NH-fingerprint region,

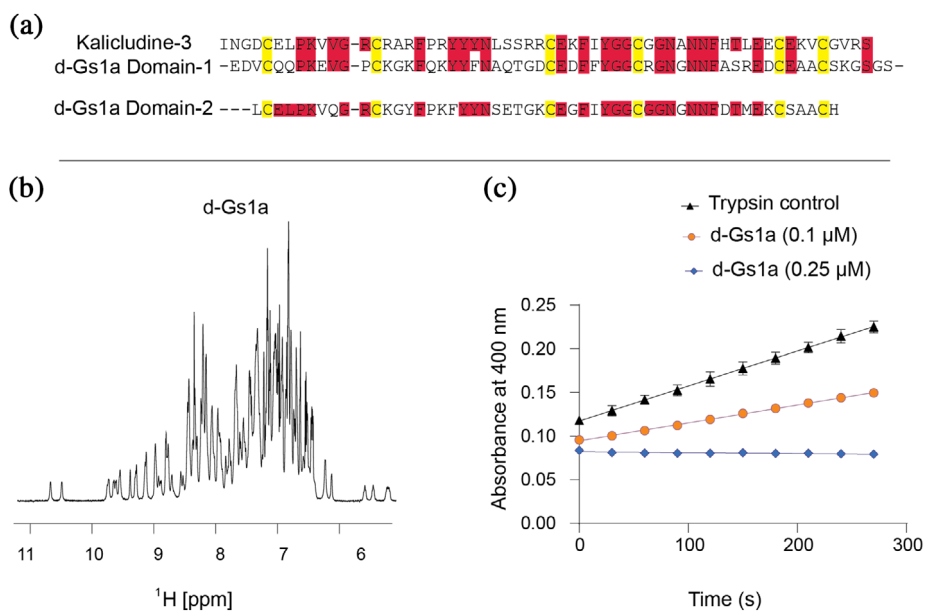


FIGURE 3 Sequence based identification, NMR confirmation of structural order and trypsin inhibition assay of d-Gs1a (A0A098LW49). (a) Alignment between Kaliclude-3 with each domain of d-Gs1a. Conserved residues between Kaliclude-3 and d-Gs1a are highlighted in red, while cysteines are highlighted in yellow. (b) 1D ^1H -NMR spectrum of d-Gs1a demonstrating well resolved and dispersed signal within the NH region, a characteristic feature of a well-defined globular fold. (c) Trypsin assay in the presence of d-Gs1a (0.1 μM and 0.25 μM) demonstrating inhibition of digestion of a trypsin substrate which fluoresces upon enzymatic cleavage (increased absorbance correlates with enzyme activity). All trypsin assays were performed in triplicate with 0.5 μM trypsin.

consistent with a well-defined globular fold (Figure 3b). As Kaliclude-3 is a known serine protease inhibitor, a trypsin inhibition assay was performed to test the function of d-Gs1a. As suspected, we find that the recombinant d-Gs1a peptide shares activity with Kaliclude-3, showing potent trypsin inhibition in a concentration dependent manner, and being able to achieve full inhibition at sub-stoichiometric ratios (Figure 3c). However, screening for activity against a wide panel of voltage-gated potassium channels, TRPV1, and ASIC1a channels (Supplementary methods), revealed that d-Gs1a does not share the dual functionality of other, venom-derived, Kunitz-type peptides (Figure S5).

4 | CONCLUSION

Naturally occurring multivalent peptides represent a valuable source of bioactive ligands, with a potential to be developed into novel biologics in the pharmaceutical and agrochemical industries. These molecules benefit from an evolutionary refinement process that offers unique insights into the underlying design principles of multivalency in peptides (Bohlen et al., 2010; Chassagnon et al., 2017). SrepeYard has been designed to be enriched for multivalent peptide ligands and provides researchers with the necessary tools to mine this resource using a variety of search and browse functions. To demonstrate the utility of this resource, we show how analyses of sequences within the database provide new insights into the significance of interdomain peptide sequences in defining peptide function. We further outline a targeted mining approach that enables the identification of novel SCREPs using the known sequence and bioactivity of previously studied receptor ligands. Using this approach, we identify a previously unknown two-domain protease inhibitor from the marine gastropod *Gemmula speciosa*. The construction and demonstrated utility of this resource promises to improve our understanding of multivalency while uncovering molecules of pharmaceutical and agricultural relevance.

AUTHOR CONTRIBUTIONS

Junyu Liu: Conceptualization (supporting); data curation (lead); formal analysis (lead); investigation (equal); methodology (supporting); writing – original draft (equal); writing – review and editing (equal). **Michael J Maxwell:** Conceptualization (supporting); formal analysis (supporting); investigation (equal); writing – original draft (equal); writing – review and editing (equal). **Thom Cuddihy:** Methodology (supporting); software (equal); writing – review and editing (supporting). **Theo**

Crawford: Investigation (equal); writing – original draft (supporting); writing – review and editing (equal). **Madeleine Bassetti:** Software (supporting); visualization (equal); writing – review and editing (supporting). **Cameron Hyde:** Methodology (supporting); software (equal); writing – review and editing (supporting). **Steve Peigneur:** Funding acquisition (supporting); investigation (equal); methodology (supporting); writing – review and editing (equal). **Jan Tytgat:** Funding acquisition (supporting); investigation (equal); methodology (supporting); writing – review and editing (equal). **Eivind Undheim:** Conceptualization (equal); formal analysis (supporting); funding acquisition (equal); project administration (equal); supervision (supporting); writing – review and editing (equal). **Mehdi Mobli:** Conceptualization (equal); formal analysis (supporting); funding acquisition (equal); project administration (equal); supervision (lead); writing – original draft (supporting); writing – review and editing (equal).

ACKNOWLEDGMENTS

We would like to thank A/Prof. Mikael Boden for guidance and insightful discussion relating to this project. We would also like to thank Mr Alan Hockings for implementing the virtual machine for SCREP datamining on the Nectar Research Cloud. In addition, we thank the Queensland NMR network (QNN) for access to the NMR facilities. This work was supported by the Australian Research Council (DP190101177 to Mehdi Mobli and Eivind A. B. Undheim), the Norwegian Research Council (FRIPRO-YRT Fellowship no. 287462 to EU), the University of Queensland Postgraduate Research Scholarship (to Michael J. Maxwell) and the University of Queensland Research Training Stipend Scholarship (to Junyu Liu). Jan Tytgat was supported by FWO-Vlaanderen grants GOA4919N, GOE7120N, and GOC2319N. Steve Peigneur was supported by KU Leuven funding (PDM/19/164) and F.W.O. Vlaanderen Grant 12W7822N. Open access publishing facilitated by The University of Queensland, as part of the Wiley - The University of Queensland agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The SrepeYard database is freely available at <http://www.srepeyard.org>. The SrepeYard web application is independent and supports most browsers. The datasets generated and analysed during the current study are available in the SCREP repository at <https://srepeyard.org/database>.

ORCID

Junyu Liu  <https://orcid.org/0000-0002-8926-0161>
Cameron Hyde  <https://orcid.org/0000-0002-5913-9766>
Steve Peigneur  <https://orcid.org/0000-0003-0504-5702>
Jan Tytgat  <https://orcid.org/0000-0003-1778-6022>
Eivind A. B. Undheim  <https://orcid.org/0000-0002-8667-3999>

REFERENCES

- Armenteros JJA, Tsirigos KD, Sonderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol.* 2019; 37:420–3.
- Arner ES, Holmgren A. Physiological functions of thioredoxin and thioredoxin reductase. *Eur J Biochem.* 2000;267:6102–9.
- Bae C, Anselmi C, Kalia J, Jara-Oseguera A, Schwieters CD, Krepkiy D, et al. Structural insights into the mechanism of activation of the TRPV1 channel by a membrane-bound tarantula toxin. *Elife.* 2016;5.
- Baghshani H, Abadi MS. Thiosulphate: cyanide sulphur transferase activity in some species of helminth parasites. *J Parasit Dis.* 2014;38:181–4.
- Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The universal protein resource (UniProt). *Nucleic Acids Res.* 2005;33:D154–9.
- Barker M, Wilkinson R, Treloar A. The Australian research data commons. *Data Sci J.* 2019;18:44.
- Berkut AA, Peigneur S, Myshkin MY, Paramonov AS, Lyukmanova EN, Arseniev AS, et al. Structure of membrane-active toxin from crab spider *Heriades melloteei* suggests parallel evolution of sodium channel gating modifiers in Araneomorphae and Mygalomorphae. *J Biol Chem.* 2015;290:492–504.
- Bobrovnik SA. The influence of rigid or flexible linkage between two ligands on the effective affinity and avidity for reversible interactions with bivalent receptors. *J Mol Recognit.* 2007;20: 253–62.
- Boby T, Patch AM, Aves SJ. TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics.* 2005;21:811–6.
- Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 2003; 31:365–70.
- Bohlen CJ, Priel A, Zhou S, King D, Siemens J, Julius D. A bivalent tarantula toxin activates the capsaicin receptor, TRPV1, by targeting the outer pore domain. *Cell.* 2010;141:834–45.
- Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol Biol.* 2016;1374:23–54.
- Brandt EG, Hellgren M, Brinck T, Bergman T, Edholm O. Molecular dynamics study of zinc binding to cysteines in a peptide mimic of the alcohol dehydrogenase structural zinc site. *Phys Chem Chem Phys.* 2009;11:975–83.
- Chassagnon IR, McCarthy CA, Chin YK, Pineda SS, Keramidis A, Mobli M, et al. Potent neuroprotection after stroke afforded by a double-knot spider-venom peptide that inhibits acid-sensing ion channel 1a. *Proc Natl Acad Sci U. S. A.* 2017;114:3750–5.
- Chen X, Zaro JL, Shen WC. Fusion protein linkers: property, design and functionality. *Adv Drug Deliv Rev.* 2013;65:1357–69.
- Chivers PT, Laboissiere MC, Raines RT. The CXXC motif: imperatives for the formation of native disulfide bonds in the cell. *EMBO J.* 1996;15:2659–67.
- Deuis JR, Dekan Z, Wingerd JS, Smith JJ, Munasinghe NR, Bhola RF, et al. Pharmacological characterisation of the highly NaV1.7 selective spider venom peptide Pn3a. *Sci Rep.* 2017;7: 40883.
- Dutertre S, Lewis RJ. Use of venom peptides to probe ion channel structure and function. *J Biol Chem.* 2010;285:13315–20.
- Escoubas P, Bernard C, Lambeau G, Lazdunski M, Darbon H. Recombinant production and solution structure of PcTx1, the specific peptide inhibitor of ASIC1a proton-gated cation channels. *Protein Sci.* 2003;12:1332–43.
- Gao Y, Cao E, Julius D, Cheng Y. TRPV1 structures in nanodiscs reveal mechanisms of ligand and lipid action. *Nature.* 2016;534: 347–51.
- Gelfand Y, Rodriguez A, Benson G. TRDB—the tandem repeats database. *Nucleic Acids Res.* 2007;35:D80–7.
- George RA, Heringa J. An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng Des Sel.* 2002;15:871–9.
- Gongora-Benitez M, Tulla-Puche J, Albericio F. Multifaceted roles of disulfide bonds. Peptides as therapeutics. *Chem Rev.* 2014; 114:901–26.
- Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol.* 2001;313:903–19.
- Gruber CW, Cemazar M, Heras B, Martin JL, Craik DJ. Protein disulfide isomerase: the structure of oxidative folding. *Trends Biochem Sci.* 2006;31:455–64.
- Guyot N, Meudal H, Trapp S, Iochmann S, Silvestre A, Jousset G, et al. Structure, function, and evolution of Gga-AvBD11, the archetype of the structural avian-double-beta-defensin family. *Proc Natl Acad Sci U. S. A.* 2020;117:337–45.
- Handl HL, Sankaranarayanan R, Josan JS, Vagner J, Mash EA, Gillies RJ, et al. Synthesis and evaluation of bivalent NDP-alpha-MSH(7) peptide ligands for binding to the human melanocortin receptor 4 (hMC4R). *Bioconjug Chem.* 2007;18:1101–9.
- Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30:1236–40.
- Jorda J, Thierry B, Andrey VK. PRDB: Protein Repeat Data Base. *Proteomics.* 2012;12:1333–6.
- Jungo F, Bougueleret L, Xenarios I, Poux S. The UniProtKB/Swiss-Prot tox-Prot program: a central hub of integrated venom protein data. *Toxicon.* 2012;60:551–7.
- Kaas Q, Westermann JC, Craik DJ. Conopeptide characterization and classifications: an analysis using ConoServer. *Toxicon.* 2010;55:1491–509.
- Kalita MK, Ramasamy G, Duraisamy S, Chauhan VS, Gupta D. ProTRepeatsDB: a database of amino acid repeats in genomes. *BMC Bioinform.* 2006;7:336.

- Klein JS, Jiang S, Galimidi RP, Keeffe JR, Bjorkman PJ. Design and characterization of structured protein linkers with differing flexibilities. *Protein Eng des Sel*. 2014;27:325–30.
- Lapatto R, Kregel U, Schreuder HA, Arkema A, de Boer B, Kalk KH, et al. X-ray structure of antistasin at 1.9 Å resolution and its modelled complex with blood coagulation factor Xa. *EMBO J*. 1997;16:5151–61.
- Le Fleche P, Hauck Y, Onteniente L, Prieur A, Denoed F, Ramisse V, et al. A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiol*. 2001;1:2.
- Letunic I, Khedkar S, Bork P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res*. 2021;49:D458–60.
- Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*. 2001;17:282–3.
- Lill R. Function and biogenesis of iron-sulphur proteins. *Nature*. 2009;460:831–8.
- Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*. 2020;48:D265–8.
- Mack ET, Snyder PW, Perez-Castillejos R, Bilgicer B, Moustakas DT, Butte MJ, et al. Dependence of avidity on linker length for a bivalent ligand-bivalent receptor model system. *J Am Chem Soc*. 2012;134:333–45.
- Mammen M, Choi SK, Whitesides GM. Polyvalent interactions in biological systems: implications for design and use of multivalent ligands and inhibitors. *Angew Chem Int ed Engl*. 1998;37:2754–94.
- Maxwell M, Undheim EAB, Mobli M. Secreted cysteine-rich repeat proteins "SCREPs": a novel multi-domain architecture. *Front Pharmacol*. 2018;9:1333.
- Meunier B, de Visser SP, Shaik S. Mechanism of oxidation reactions catalyzed by cytochrome p450 enzymes. *Chem Rev*. 2004;104:3947–80.
- Miller KL, Lanthier M. Regulatory watch: innovation in biologic new molecular entities: 1986–2014. *Nat Rev Drug Discov*. 2015;14:83.
- Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. *Nucleic Acids Res*. 2021;49:D412–9.
- Mobli M, Undheim EA, Rash LD. Chapter seven—modulation of ion channels by cysteine-rich peptides: from sequence to structure. In: Geraghty DP, Rash LD, editors. *Advances in pharmacology*. Cambridge, MA: Academic Press; 2017. p. 199–223.
- Monera OD, Sereda TJ, Zhou NE, Kay CM, Hodges RS. Relationship of sidechain hydrophobicity and alpha-helical propensity on the stability of the single-stranded amphipathic alpha-helix. *J Pept Sci*. 1995;1:319–29.
- Muttenthaler M, King GF, Adams DJ, Alewood PF. Trends in peptide drug discovery. *Nat Rev Drug Discov*. 2021;20:309–25.
- Necci M, Piovesan D, Dosztanyi Z, Tosatto SCE. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*. 2017;33:1402–4.
- Paladin L, Hirsh L, Piovesan D, Andrade-Navarro MA, Kajava AV, Tosatto SC. RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures. *Nucleic Acids Res*. 2017;45:D308–12.
- Pennington MW, Czerwinski A, Norton RS. Peptide therapeutics from venom: current status and potential. *Bioorg Med Chem*. 2018;26:2738–58.
- Perdigao N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, et al. Unexpected features of the dark proteome. *Proc Natl Acad Sci U. S. A*. 2015;112:15898–903.
- Postic G, Gracy J, Perin C, Chiche L, Gelly JC. KNOTTIN: the database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling. *Nucleic Acids Res*. 2018;46:D454–8.
- Rholam M, Fahy C. Processing of peptide and hormone precursors at the dibasic cleavage sites. *Cell Mol Life Sci*. 2009;66:2075–91.
- Rodgers KR, Chou RC. Therapeutic monoclonal antibodies and derivatives: historical perspectives and future directions. *Bio-technol Adv*. 2016;34:1149–58.
- Schurmann P, Buchanan BB. The ferredoxin/thioredoxin system of oxygenic photosynthesis. *Antioxid Redox Signal*. 2008;10:1235–74.
- Schweitz H, Bruhn T, Guillemare E, Moinier D, Lancelin JM, Beress L, et al. Two different classes of sea anemone toxins for voltage sensitive K⁺ channels. *J Biol Chem*. 1995;270:25121–6.
- Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, et al. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res*. 2010;38:D161–6.
- Soler MA, Fortuna S. Influence of linker flexibility on the binding affinity of bidentate binders. *J Phys Chem B*. 2017;121:3918–24.
- Tran HNT, Tran P, Deuis JR, Agwa AJ, Zhang AH, Vetter I, et al. Enzymatic ligation of a pore blocker toxin and a gating modifier toxin: creating double-knotted peptides with improved sodium channel NaV1.7 inhibition. *Bioconjug Chem*. 2020;31:64–73.
- Turner AJ, Murphy LJ. Molecular pharmacology of endothelin converting enzymes. *Biochem Pharmacol*. 1996;51:91–102.
- Undheim EAB, Mobli M, King GF. Toxin structures as evolutionary tools: using conserved 3D folds to study the evolution of rapidly evolving peptides. *Bioessays*. 2016;38:539–48.
- Uvyn A, De Geest BG. Multivalent antibody-recruiting macromolecules: linking increased binding affinity with enhanced innate immune killing. *Chembiochem*. 2020;21:3036–43.
- Van Anken E, Braakman I. Versatility of the endoplasmic reticulum protein folding factory. *Crit Rev Biochem Mol Biol*. 2005;40:191–228.
- Van de Locht A, Bauer M, Huber R, Friedrich T, Kroger B, Hoffken W, et al. Two heads are better than one crystal structure of the insect derived double domain Kazal inhibitor rhodniin in complex with thrombin. *EMBO J*. 1995;14:5149–57.
- Van de Locht A, Stubbs MT, Bode W, Friedrich T, Bollschweiler C, Hoffken W, et al. The ornithodorin-thrombin crystal structure, a key to the TAP enigma? *EMBO J*. 1996;15:6011–7.
- Vauquelin G, Charlton SJ. Exploring avidity: understanding the potential gains in functional affinity and target residence time of bivalent and heterobivalent ligands. *Br J Pharmacol*. 2013;168:1771–85.
- Wilkinson B, Gilbert HF. Protein disulfide isomerase. *Biochim Biophys Acta Proteins Proteom*. 2004;1699:35–44.

Zhang Y, Peng D, Huang B, Yang Q, Zhang Q, Chen M, et al. Discovery of a novel Nav1.7 inhibitor from *Cyriopagopus albostratus* venom with potent analgesic efficacy. *Front Pharmacol.* 2018;9:1158.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu J, Maxwell M, Cuddihy T, Crawford T, Bassetti M, Hyde C, et al. SrepYard: An online resource for disulfide-stabilized tandem repeat peptides. *Protein Science.* 2023;32(2):e4566. <https://doi.org/10.1002/pro.4566>