Published online 1 February 2022

Nucleic Acids Research, 2022, Vol. 50, No. 9 ***e50***
https://doi.org/10.1093/nar/gkac029

SEQUENCE SLIDER: integration of structural and genetic data to characterize isoforms from natural sources

Rafael J. Borges [1,2,*], Guilherme H. M. Salvador[1], Daniel C. Pimenta[3], Lucilene D. dos Santos[4,5], Marcos R. M. Fontes[1] and Isabel Usón[2,6]

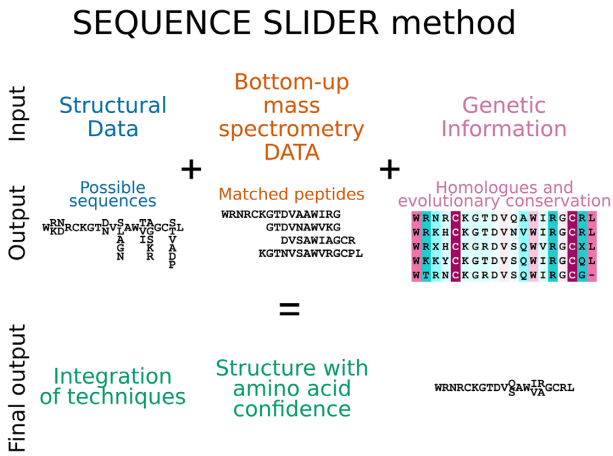
[1]Departament of Biophysics and Pharmacology, Biosciences Institute, São Paulo State University (UNESP), Botucatu, São Paulo 18618-689, Brazil, [2]Crystallographic Methods, Institute of Molecular Biology of Barcelona (IBMB–CSIC), Barcelona 08028, Spain, [3]Biochemistry and Biophysics Laboratory, Butantan Institute, São Paulo, São Paulo 05503-900, Brazil, [4]Graduate Program in Tropical Diseases, Botucatu Medical School (FMB), São Paulo State University (UNESP), Botucatu, São Paulo 18618-687, Brazil, [5]Biotechnology Institute (IBTEC), São Paulo State University (UNESP), Botucatu, São Paulo 18607-440, Brazil and [6]ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

Received September 20, 2021; Revised January 05, 2022; Editorial Decision January 06, 2022; Accepted January 30, 2022

ABSTRACT

Proteins isolated from natural sources can be composed of a mixture of isoforms with similar physicochemical properties that coexist in the final steps of purification. Yet, even where unverified, the assumed sequence is enforced throughout the structural studies. Herein, we propose a novel perspective to address the usually neglected sequence heterogeneity of natural products by integrating biophysical, genetic and structural data in our program SEQUENCE SLIDER. The aim is to assess the evidence supporting chemical composition in structure determination. Locally, we interrogate the experimental map to establish which side chains are supported by the structural data, and the genetic information relating sequence conservation is integrated into this statistic. Hence, we build a constrained peptide database, containing most probable sequences to interpret mass spectrometry data (MS). In parallel, we perform MS *de novo* sequencing with genomic-based algorithms to detect point mutations. We calibrated SLIDER with *Gallus gallus* lysozyme, whose sequence is unequivocally established and numerous natural isoforms are reported. We used SLIDER to characterize a metalloproteinase and a phospholipase A_2-like protein from the venom of *Bothrops moojeni* and a crotoxin from *Crotalus durissus collilineatus*. This integrated approach offers a more realistic structural descriptor to characterize macromolecules isolated from natural sources.

GRAPHICAL ABSTRACT



INTRODUCTION

Structural biology has been invaluable to unravel the molecular mechanisms of biological macromolecules and their complexes. The observation of different states and chemical reaction intermediates has provided the base for major breakthroughs in the last six decades with X-ray crystallography being the most used structural technique. A validated crystallographic model derived from high-resolution diffraction data with global indicators conforming to the norm is usually unquestioned by the scientific community. It will be adopted to interpret biochemical data and to perform theoretical studies, such as homology modelling, molecular dynamics, docking and virtual screening, even if some local features included in the model

*To whom correspondence should be addressed. Tel: +55 14 3880 0262; Fax: +55 14 3880 0238; Email: rafael.borges@unesp.br

may be undetermined (1). Since the resolution revolution, single-particle cryogenic electron microscopy (cryoEM) has reached atomic resolution with apoferritin (2). Homogeneous samples, favoured by high symmetry and low flexibility represent particularly favourable cases, whereas the typical resolution in a cryoEM reconstruction varies locally between 3 and 20 Å. In regions with resolution better than 3 Å, side chain modelling becomes possible, but it is subject to specific issues leading to side chains strongly affected by radiation damage or carrying net negative charges frequently becoming invisible in the electrostatic potential map (3,4).

CryoEM obviates the need for crystallization, relieving the difficulty to study complex structures such as ribosomes, ionic channels and viral proteins. Various of these samples are obtained directly from the natural source, such as the trimeric spike protein isolated from the severe acute respiratory syndrome coronavirus 2 (5), ryanodine receptor isoform 1 purified from rabbit skeletal muscle (6–8) and its isoform 2 from dog heart ventricles (9), inositol 1,4,5-trisphosphate receptor type 1 from rat cerebellum (EMD-6369) and Nav1.4-β1 Complex from electric eel (10).

Prior knowledge of the sample composition is essential for building an atomic model of the macromolecule in the typical scenario where structural data do not reach atomic resolution. For proteins produced recombinantly, sequence knowledge comes beforehand, whereas investigating structures obtained from natural sources, isoforms that share physicochemical properties may coexist even after extended purification. Lysozyme, the most studied protein in crystallography (11), has been seen in at least four isoforms in bovine cartilage (12) and three in the medicinal leech (13). Most relevant is the case of natural products where the concerted action of isoforms enhances their function, such as for natural venoms. Venoms are composed of a cocktail of toxins characterized by one of the most rapid evolutionary divergence and variability seen in any category of proteins (14). In fact, variability in venom composition within a single species may be related to ontogeny (15), diet (16), seasonality (17), geographical location (18) and gender (19,20). In the case of snake venom phospholipases A_2 (PLA_2s), at least 16 isoforms of β-bungarotoxin (21) and 16 isoforms of crotoxin (22,23) have been identified and characterized. Importantly, the precise isoform may determine biological activity: as an example, two point mutations F124I and K128E in the ammodytoxins are enough to change their toxicity, anticoagulant properties and toxin targets (24). Researchers may assume their sample coincides with one entry on the sequence and structural repositories containing crystal and cryoEM structure; therefore, bias propagation of established sequences is a concern in the field of natural products. Moreover, repositories are not error-free as they are curated at a level relying mostly on global validating statistics, and the information they contain ultimately depends on depositors.

For non-recombinant samples, crystallographic and cryoEM determination can be effectively complemented by independent genetic information. A sophisticated analysis of electron density was implemented in the method *findMySequence*, which aims to identify the most plausible protein sequence in a database given a density map (25). It implements a sequence alignment based on a machine-learning residue-type classifier (25). Relevant genetic information can be derived from evolutionary conservation and amino acid frequencies from phylogenetic relations between homologues (26,27), as residues related to function are conserved and structural stability requires concerted pairwise changes of interacting residues (28–31). Such information has also proven useful for predicting protein structures (implemented in AlphaFold (32,33), RoseTTaFold (34) and 3Dseq (35)) and using fragments from these for *ab initio* phasing of crystallographic structures using Molecular Replacement (implemented in Ample (36)), as they infer nearby residues analysing evolutionary covariance (32,35,36).

Mass spectrometry (MS), Edman degradation and cDNA sequencing are the complementary experimental techniques that allow determination of protein sequences directly purified from natural sources (37). MS does not require samples of high purity and has been a breakthrough characterizing complex samples (38,39). Venomics, the particular application of omics fields to characterize venom, has advanced to the point of allowing one to characterize the variability between different isoforms contributing as little as 0.05% (39,40). Conventional identification approaches require concordance between the fragmented peptides from MS/MS with the predicted digested sequences from a database, although unknown sequences or proteins with strong polymorphism are not revealed (reviews in (41–43)). Alternatives are *de novo* sequencing, which relies only on the experimental data, usually at the price of compromising coverage, and algorithms that tolerate mismatches, point mutations and post-translation modifications (PTM). Determining side chain composition in structures of samples containing multiple isoforms will not be trivial and will require the integration of information from different biophysical techniques, including MS (44).

The integrated use of MS and structural determination has proven invaluable in multiple ways comprising protein-construct design, purification, crystallization, phasing and model building (45). MS can define compact domains and mobile regions guiding truncation to improve sample preparation, assess the components present in a crystal, measure the number of heavy atoms in a crystal and address issues in modelling, topology and side-chain proximity (45). Diemer *et al.* characterized a phosphate binding protein purified from human plasma, but absent in eukaryotic genome databases by tandem use of crystallography and MS data combining multiple alternative digestions to obtain the correct sequence (46). Guo *et al.* (47) characterized the sequence of haemoglobins from two endangered felines by use of crystallography, MS data from single proteolysis and evaluation of sequences from homologues (47).

The side chain evaluator method called SEQUENCE SLIDER (SLIDER) provides a novel framework to coordinate this complex task of integrating structural and genetic information. SLIDER (48) was first created in the *ab initio* phasing scope of ARCIMBOLDO methods (49), in which small fragments are identified in the asymmetric unit using molecular replacement (50) and the rest of the structure is revealed through density modification and automatic map interpretation (51,52). SLIDER uses available sequence information, secondary structure prediction or alignment be-

tween remote homologues, to build the most probable side-chain atoms into a partial solution usually composed of polyalanine fragments. Discrimination in global statistics of the agreement between data and model may reveal the true sequence hypothesis (48).

Herein, we describe the implementation of SLIDER to integrate structure determination, mass spectrometry and genetic information to address the heterogeneity of complex samples purified from natural sources building a probability of amino acids per residue basis. SLIDER was first carefully calibrated using a crystallographic dataset of the known hen egg-white lysozyme and further refined with the basement membrane-specific heparan sulphate proteoglycan core protein (BaM). Subsequently, we apply it to elucidate a metalloproteinase (BmooMP-I) (53), a PLA_2-like protein (MjTX-I) from *Bothrops moojeni* and a crotoxin from *Crotalus durissus collineatus* (CBCol), revealing sequences not yet seen in the literature. We offer the structural community a methodology that consistently assigns amino acids integrating available information for macromolecules purified from natural source.

MATERIALS AND METHODS

Protein purification

Lyophilized lysozyme from *Gallus gallus* was purchased from Sigma-Aldrich. Freeze-dried *Bothrops moojeni* crude venom was purchased from Centro de Extração de Toxina Animais (CETA), Morungaba – SP, Brazil and freeze-dried *Crotalus durissus collineatus* crude venom was purchased from Serpentarium Bioagents of Batatais – SP, Brazil. BmooMP-I was isolated from *B. moojeni* snake venom by cation-exchange chromatography on a CM-FF column (5 ml) followed by reverse-phase using a C-18 column as described by Salvador *et al* (53). CBCol was isolated from *C. d. collilineatus* snake venom by cation-exchange chromatography on a CM-Sepharose column (2 × 20 cm), followed by the dissociation of subunits according to Hendon & Fraenkel-Conrat (54).

Kidney tissue lysis preparation

Two hundred microliters of lysis buffer (8 M urea, 75 mM NaCl, 25 mM Tris-HCl pH 8, 2 mM $MgCl_2$, protease inhibitor (Roche) 1×, benzonase 1 U) were added to a microtube containing 20 mg of healthy mouse (*Mus musculus*) kidney (Ethics Committee Ceau 07150321–0). The sample was submitted to Sonics Vibra-Cell VCX-600 Ultraconic Processor equipment using 5 pulses by minute (25 Hz of power/20% amplitude) with 1 min interval within pulses. Following centrifugation to 14 000 *g* for 30 min, the supernatant was collected.

Crystallographic experiments

Crystals were obtained by the hanging-drop vapour diffusion method at 18°C (McPherson, 2009). Lysozyme crystals were grown in crystallization drops composed of 1 μl of protein solution (50 mg/ml) and 1 μl of the precipitant solution, equilibrated against the reservoir solution of 1.2 M $MgCl_2$ and 0.1 M Tris HCl, pH 7.6. BmooMP-I crystals were obtained by mixing 0.5 μl protein solution (20 mg/ml) and 0.5 μl reservoir solution equilibrated against a 50 μl reservoir containing 30% (w/v) PEG 400, 0.05 M Tris HCl, pH 8.5, 0.05 M lithium sulphate and 0.05 M sodium sulphate (53). CBCol crystals were obtained from a crystallization drop composed of 1 μl of protein (10 mg/ml) and 1 μl of precipitant solution equilibrated against reservoir solution of 2.0 M ammonium sulphate and 0.1 M Tris HCl, pH 9.0 (55).

Crystals were mounted in nylon loops and flash cooled in liquid nitrogen. Diffraction data were collected at the MX2 beamline, Laboratório Nacional de Luz Síncrotron (LNLS, Campinas, Brazil). The BmooMP-I dataset was indexed, integrated and scaled using HLK2000 (56), and all other data with XDS (57). The BmooMP-I, MjTX-I, CBCol and lysozyme structures were solved by molecular replacement with PHASER (50), using the BmooMPalpha-I (PDB code: 3GBO), MjTX-I (6CE2), crotoxin basic subunit (2QOG), lysozyme (6G8A) coordinates, respectively. Refinement was performed with phenix.refine (58), and the model was manually built with Coot (59) using the σ_A-weighted F_{obs}-F_{calc} and $2F_{obs}$-F_{calc} maps. The crystallographic data and molecular structure of the BaM were extracted from the PDB (1GL4) (60). Possible improvement of main and side chain positions was probed with the PDB_REDO server (61). The final models were evaluated using the validation analysis from Molprobity (62).

Sequence assignment

Crystallography. SEQUENCE SLIDER for natural compounds was used to assign the sequence of lysozyme, BaM, BmooMP-I, MjTX-I and CBCol. A protein model with good agreement to the experimental map based on global indicators conforming to the norm, R_{free} below 25% and Molprobity score below the resolution of the data, was used as starting point. Rotamers for each one of the 20 possible amino acids were generated using the coot function 'auto-fit-best rotamer' for each residue in the protein model. Water molecules within 2.5 Å were removed with the coot 'delete_atoms' command. All atoms within 5 Å distance of the focus residue were refined using coot 'refine residues' (63). Side-chain atoms had their *B*-factors set to 30, and their real-space correlation coefficient (RSCC) was calculated against phenix.polder omit maps (64). Each residue with its omit map from phenix.polder was reviewed manually to assess the possibility of sequence or conformational heterogeneity consistent with its chemical surrounding. Amino acids with top RSCC values were included in the probability distribution to generate a database of sequences to be used against MS data. All potential peptides from proteolytic cleavage, which in the cases of this article were for Arg/Lys (trypsin), were generated. Images were generated using Coot (63) and Raster3D (65).

Mass spectrometry. intact proteins. About 5 μl aliquots of lysozyme previously diluted in 50% acetonitrile (ACN) and 1% formic acid were analysed by direct infusion into the Electrospray Ionization Time-of-Flight mass spectrometer (Shimadzu, Kyoto, Japan), under 0.2 ml/min constant flow of 50% of solution A&B (solution A: water: DMSO:

formic acid (949: 50: 1) and solution B: ACN: DMSO: water: formic acid (850: 50: 99: 1)). The MS interface was kept at 4.5 kV and 275 ºC. Detector operated at 1.95 kV. MS spectra were acquired in positive mode, in the 350–1400 m/z range and the multiply charged ions were manually deconvoluted for molecular mass determination.

Digested proteins. Tryptic digests of the samples were solubilized in 0.1% (v/v) formic acid (solution A) and subjected to nano-ESI-LCMS/MS analysis, using an Ultimate 3000 HPLC (Dionex), coupled to a Q Exactive Orbitrap™ mass spectrometer (Thermo Fisher Scientific). Peptides were loaded on a Trap Column with nanoViper Fitting (P/N 164649, C_{18}, 5 mm × 30 μm, Thermo Fisher Scientific) and eluted at a flow rate of 300 nl/min using an isocratic gradient of 4% solution B (100% (v/v) acetonitrile containing 0.1% (v/v) formic acid) for 3 min. Thereafter, peptides were loaded on a C18 PicoChip column (Reprosil-Pur®, C18-AQ, 3 μm, 120 Å, 105 mm, New Objective, Woburn, MA, USA) using a segmented concentration gradient from 4–55% B for 30 min, 55–90% B for 1 min, 90% B over 5 min and then returning to 4% B over 20 min at a flow rate of 300 nl/min. Ion polarity was set to positive ionization mode using data-dependent acquisition (DDA) mode. Mass spectra were acquired with a scan range of m/z 200–2000, resolution of 70 000 and injection time of 100 ms. The fragmentation chamber was conditioned with collision energy between 29 and 35% with a resolution of 17 500, 50 ms injection time, 4.0 m/z isolation window and dynamic exclusion of 10 s. The spectrometric data were acquired through the *Thermo Xcalibur*™ software v.4.0.27.19 (Thermo Fisher Scientific).

All MS raw data were compared against the UniProt (66) database with the *PatternLab* software v.4.0.0.84 (67) or MASCOT (68). A database with the most probable sequences from the crystallographic evaluation was also used. The following search parameters were used: semi-tryptic cleavage products (two tryptic missed cleavages allowed), carbamidomethylation of cysteine as fixed modification and oxidation of methionine as variable modification. Parent mass tolerance error was set at 40 ppm and fragment mass error at 0.02 ppm. Proteins were identified considering a minimum of one fragment ion per peptide, five fragment ions per protein, two peptides per protein and a false discovery identification rate set to 1%, estimated by a simultaneous search against a reversed database.

MS raw data were evaluated using PTM, point mutation and *de novo* algorithms with PEAKS/SPIDER (69,70). *De novo* analyses were performed with tolerances of 0.01 Da for precursor and fragment ions; methionine oxidation and carbamidomethylation of cysteine were considered variable modifications and fixed, respectively; maximum of 3 PTMs; *de novo* score (ALC%) threshold of 15 and peptide hit threshold (-10log$P$): 30.0.

Phylogenetic analysis. The most probable sequences rendered by MS and crystallographic analysis, along with their structural coordinates, were input to ConSurf (26). For the multiple sequence alignment, HMMER homolog search (71), which implements profile hidden Markov models to find remote homologues, was parameterized with a single iteration, E-value cutoff of 0.0001 and using the UniProt Reference Clusters 90 database (UniRef90) (66). The alignment was built using MAFFT-L-INS-I (72), software designed for the alignment of hundreds to thousands of sequences. For each sample, the 300 most similar sequences (within an interval of 100% to 35% sequence identity) were used to calculate a percentage of amino acid occurrences for each residue position in the output file 'query_msa.aln'.

SEQUENCE SLIDER is an open-source initiative written in Python3 and distributed through a GitHub repository (https://github.com/LBME/slider) and within ARCIMBOLDO (http://chango.ibmb.csic.es/), through the package installer for Python (PyPI) and Collaborative Computational Project No. 4 (CCP4) (73).

RESULTS

Method description

SLIDER integrates crystallography, mass spectrometry and phylogenetic data to assign sequence and its heterogeneity in complex samples (Figure 1). All data gathered from these three techniques can be jointly used to restrict hypothesis generation or independently to validate one another. In crystallography, SLIDER starts from a reflection file containing experimental intensities or amplitudes and an initial protein file with global indicators conforming to the norm, R_{free} below 25% and Molprobity score below the resolution of the data, and, if not supplied, it generates a map with phenix.maps (74). It builds the rotamer with the best fit to the electron density map for every one of the 20 natural amino acids in each residue position, removes water molecules within a 2.5 Å radius and refines the positions of atoms within 5.0 Å distance using the software Coot (59). Their side chain atoms are used to estimate the agreement between their calculated electron density with the observed one using the real-space correlation coefficient (RSCC) in phenix.polder (64), a methodology that creates an omit map excluding the bulk solvent around the omitted region. This strategy does not use a precomputed library or templates of expected side chain electron densities, as it enforces for each residue a new electron density calculated from side chain coordinates fitted in the experimental map. Alternatively, instead of searching for all 20 amino acids, restrictions may be extracted from genetic information (homologues) or from MS data (illustrated in Figure 1 by arrow 'Restricting seq hyp' connecting Genetic information or Mass spectrometry data ellipses to Structural data ellipse). For each residue position, single or multiple amino acid possibilities may be accommodated in the observed map, measured herein with RSCC (structural data ellipse in Figure 1). To discriminate among them automatically, we use a statistical indicator referred to as Δcontrast, which is calculated by the subtraction of the RSCC of each amino acid with the next best possibility (as an example, the Δcontrast of the best scoring amino acid would be the subtraction of its RSCC with the second best RSCC value). We visually inspected the omit maps relating to the Δcontrast values and established empirically that a Δcontrast of 3.0% is an appropriate threshold to distinguish one amino acid from all others. Seven residues from three groups that share similar scattering and shape (and thus are approximately isosteric) were

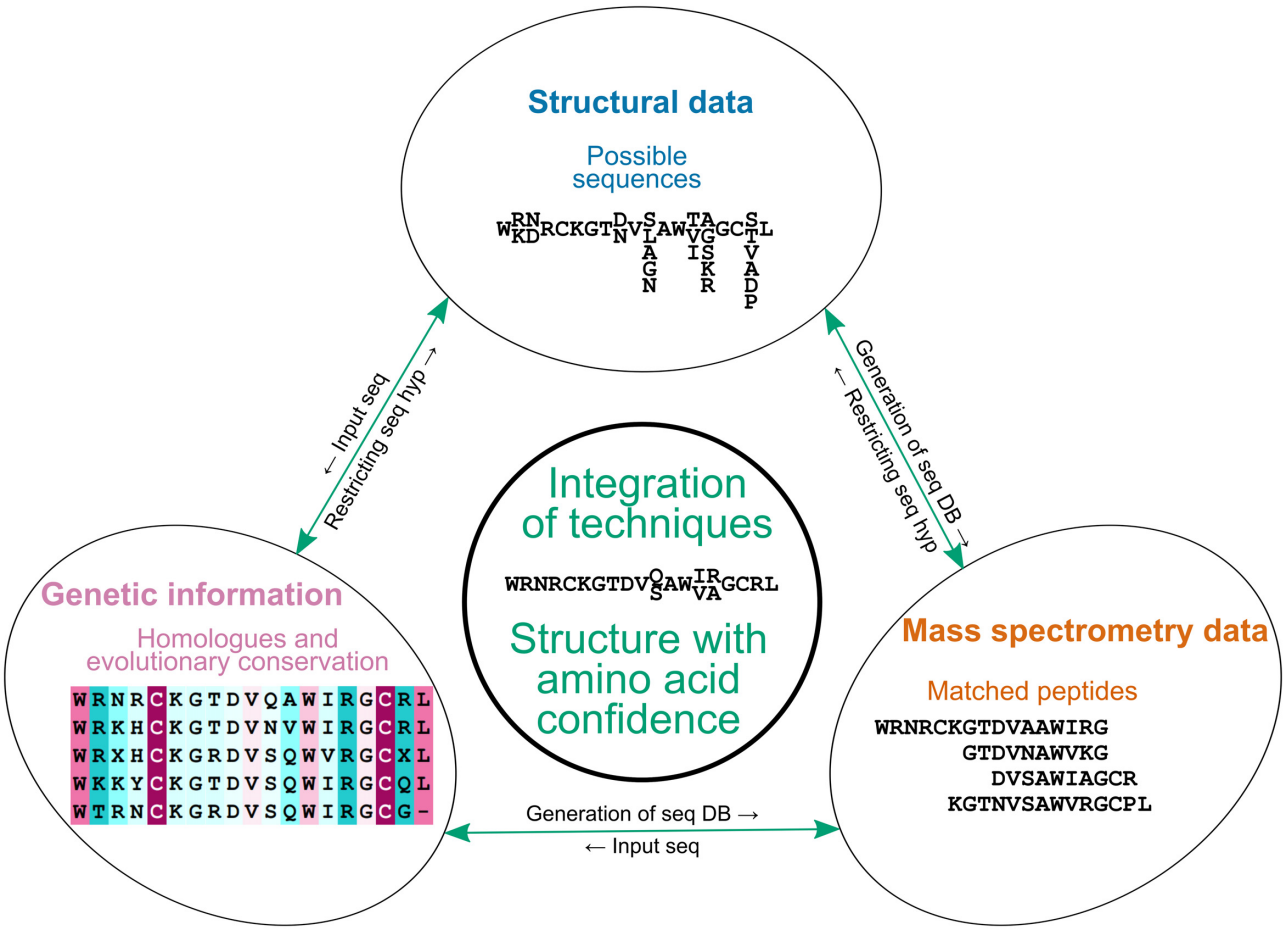


Figure 1. Scheme of how SEQUENCE SLIDER integrates structural, mass spectrometry and phylogenetic data for sequence assignment.

ambiguous: (i) valine and threonine; (ii) asparagine, aspartic acid and leucine; (iii) glutamic acid and glutamine. These amino acids, however, can be unambiguously identified by analysing the chemical environment. Hydrophobic residues will likely be buried and surrounded by other hydrophobic side chains, and hydrophilic residues will form a distinctive H-bonding pattern.

From the single and multiple possibilities of amino acids in each residue position, a database of sequences in FASTA format can be built (arrow 'Generation of seq database' connecting Structural data ellipse to Mass spectrometry ellipse in Figure 1). To reduce the exponential growth of the combination of amino acid sequences, possibilities are generated based on the expected peptide pattern from proteolytic cleavage. This database also incorporates the reversed order of each sequence as a baseline to estimate true and false positives by the available mass spectrometry software, which is known as false discovery rate. Against this database built from crystallographic data, we used Patternlab for Proteomics (http://www.patternlabforproteomics.org/), the Brazilian integrated computational environment for analysing shotgun proteomic data conceived by Dr Paulo C. Carvalho (67). Alternatively, the MS/MS spectrum can be searched against *in silico* digested sequences of proteins available in public databases, such as NCBI (75) and UniProt databases (66) (illustrated in Figure 1 by arrow 'Generation of seq database' connecting the Genetic information ellipse to the Mass spectrometry data ellipse). Additionally, the possibility of PTM and homology peptide mutations can be evaluated with *de novo* sequencing, which is available in PEAKS/SPIDER (69,70) (arrow 'Restricting seq hyp' connecting the Mass spectrometry ellipse to the Structural data ellipse in Figure 1).

Estimation of sequence conservation and the amino acid frequency of each residue position based on homologue sequences (illustrated by Genetic information ellipse in Figure 1) is extracted by SLIDER from the results from ConSurf, a bioinformatics tool to estimate evolutionary conservation based on phylogenetic relations among homologues (26,27), giving as input sequence the most probable sequence using results from the previous techniques (arrow 'Input seq' connecting either Structural data or Mass spectrometry data ellipses to the Genetic information ellipse in Figure 1). Various tools are available for remote homologue sequence detection (76), e.g. HH-suite3 (77). Currently, Consurf was chosen for this task because it builds phylogenetic tree, calculates sequence conservation by residue position and incorporates visualization of this conservation score in an atomic model in 3D, functionalities that could be later incorporated in SLIDER. In Consurf, the HMMER ho-

molog search algorithm (71) is used, and the multiple sequence alignment of the 300 most similar sequences (within an interval of 100% and 35% sequence identity) extracted from the UniProt database (66) are built using MAFFT-L-INS-I (72). From the multiple sequence alignment output file, the percentages of occurrence of the 20 amino acids at each residue position are calculated. Highly conserved regions, where a single amino acid is observed (100% occurrence), are related to function or structural stability (26,27) and therefore should be conserved in the sample sequence. Mutations should be expected in regions with a high degree of variability. This analysis is useful to confirm assignments from crystallography and MS data and to distinguish among groups of approximately isosteric amino acids, as discussed above. The phylogenetic analysis could restrict hypothesis generation from the crystallographic data instead of considering all 20 possible amino acids in each residue position. However, this strategy was not applied here to prevent bias from known sequences; as well the crystallographic resolution of 2.0 Å or better, attained in the work described here, should already allow sufficient distinction. For the BaM case where MS data were collected for a mixture of proteins, phylogenetic analysis was extended to protein regions not represented in the experimental spectra.

For some residues, information gathered may point to a single possibility; for others, more than one amino acid variation may be possible, which can be built with partial occupancy or even as an unknown residue (described as UNK). Alternative identifications for amino acids at a particular sequence position, known as microheterogeneity (78), are allowed in the PDB deposition. In summary, SLIDER balances different sources of information to either assign a single amino acid if unambiguously identified, multiple possibilities or even unknown when the evidence is not conclusive, in an approach that derives sequence hypothesis from data and prevents bias propagation from information known *a priori*.

The computing requirements for SLIDER depend on the resolution of the crystallographic data and the number of residues in the asymmetric unit. The crystallographic calculations vary between 5 and 20 minutes per residue, but as the method takes advantage of multiprocessing, this time is divided by the number of physical cores available. An initial test estimates memory usage given the number of cores; the latter may be reduced to prevent RAM memory overload. The MS and phylogenetic analyses take around 10 min per dataset. The calculations for cases described here were run on a laptop equipped with Intel i7-10710U (6 physical cores @ 3.67 GHz) and 16 Gb of RAM. The lysozyme case took 2.0 h, 2.8 h for BaM, 3.6 h for BmooMP-I, 20.5 h for MjTX-I and 5.7 h for Cb.

Lysozyme

Lysozyme is one of the most studied structures, as it is readily obtained (e.g. from egg white) and crystallized, totalling over 5000 entries in the PDB (11). Still, these represent but a fraction of the thousands of sequences registered for different species in multiple isoforms in the NCBI (75) and UniProt databases (66). As the structure and sequence of this enzyme have been extensively characterized, we chose it to calibrate the SLIDER methodology. Using the NCBI database against the MS spectrum of *Gallus gallus* lysozyme yields a high coverage of 117 residues out of 129 (Figure 2A). We collected diffraction data to a resolution of 1.9 Å and concluded model-building and refinement once agreement between experimental and calculated data (R_{work}/R_{free} of 18.5/21.7%) and stereochemistry (Molprobity score is 1.71, corresponding to the 87th percentile for structures in the PDB) were conventionally acceptable. The automatic main and side chain rebuilding method PDB_REDO (61) led to lower $R_{factors}$ by about 0.7%, with 3 side chain rotamers changed.

For the SLIDER sequence evaluation using crystallographic data based on the RSCC and Δcontrast calculations, the groups of (i) valine and threonine, (ii) asparagine, aspartic acid and leucine, (iii) glutamic acid and glutamine had possibilities restricted by evaluating their chemical environment. Twenty-eight solvent-exposed residues (21.7%) were deemed to be hydrophilic, while 13 buried residues (10.1%) were restricted to the hydrophobic group (Figure 2B). Hence, we generated 9648 sequence hypotheses involving the residues that were favoured by the electron density, as evaluated by the RSCC, and the previous physicochemical restrictions (Figure 2B).

In applying PEAKS/SPIDER against the lysozyme MS data and UniProt database restricted to *Gallus gallus* sequences, we found 845 peptide-spectrum matches with a false discovery rate of 0.2%, which is the expected percentage of peptides where a reversed sequence, and therefore are false positive, was selected. The known single sequence (P00698) showed good coverage, leading to 3, 113, 12 and 1 residues with none, one, two and four amino acid possibilities, respectively (sequences shown in PEAKS/SPIDER description of Figure 2A). The sequence database built from crystallographic data evaluated with PatternLab presented more variation, showing 5, 96, 23, 3 and 2 residues with none, one, two, three and four amino acid possibilities, respectively (sequences shown in PatternLab description of Figure 2A). Comparing both evaluations, 83 residues (64.3%) were given the same assignment while 46 (35.7%) were different.

Most residues in the crystallographic model clearly favoured a single amino acid, discriminated by the Δcontrast. These calculations were confirmed by manual evaluation of omit maps, by MS and/or by phylogenetic analysis (Figure 2A and Supplementary Table S1). The ambiguity of 39 residues (30.2%) was resolved by the MS spectrum (orange residues in Figure 2A and Supplementary Table S2) and ambiguity in the remaining 21 residues (16.3%) was resolved with phylogenetic analysis (Supplementary Table S3).

The isomorphous template (PDB code: 6G8A) used to phase our lysozyme crystallographic model has the presumably correct sequence. It was confirmed by measuring the intact mass of our sample using mass spectrometry as the experimental molecular mass of 14306.31 ± 2.81 Da agrees with the theoretical mass (14 303.88 Da). PEAKS/SPIDER identified 113 residues correctly and unambiguously (87.6%), and 13 including the correct choice among alternative amino acid possibilities (10.0%) but failed to find 3 (2.4%). PatternLab using the database

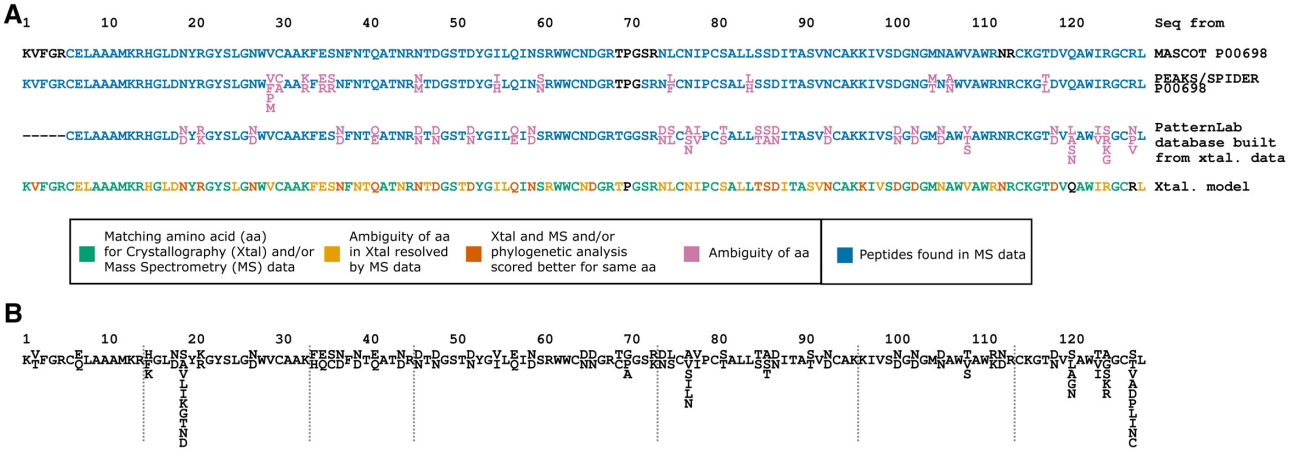


Figure 2. Sequence assignment of lysozyme. (**A**) Mass spectrometry data are used against an NCBI and UniProt database of lysozyme sequences, against PEAKS/SPIDER evaluation and against PatternLab with database built from crystallographic data and the sequence from the crystal model. (**B**) Database constructed by SEQUENCE SLIDER from crystallographic data.

built from crystallographic data identified 95 residues correctly (73.6%), 27 correct but ambiguous (20.9%), 3 residues were incorrectly assigned (2.4%) and failed to find 5 (3.9%). Pro70, Gln121 and Arg128 were not seen in this database; instead a Gly70, Leu/Asn/Ser/Ala/Gly121 and Asn/Pro/Val129 were observed (Supplementary Table S4). In fact, the residues Gln121 and Arg129 could not have been found by PatternLab, being absent from the sequences generated (Figure 2B), as their electron density omit maps support serines (Figure 3). Also, the omit map for residue 70 supports a Gly (Figure 3). Considering the PEAKS/SPIDER amino acids, residues 70–72 were not seen, but Gln121 and Arg129 were observed. Overall, the lysozyme data allowed us to calibrate SEQUENCE SLIDER, to establish the minimum Δcontrast to distinguish a particular side chain, to integrate MS and genetic information to previous assignments.

Basement membrane protein BaM

To further assess SEQUENCE SLIDER in the less ideal scenario of a membrane protein with limited experimental data, we evaluated the crystallographic, phylogenetic and MS data of the basement membrane-specific heparan sulphate proteoglycan core protein BaM. We identified BaM in bottom-up MS data of mouse kidney, but it is also present in specialized extracellular matrix underlying epithelia and surrounding peripheral nerve axons, muscle and fat cells (60), and BaM is conserved in all metazoa. The chain B of PDB entry 1GL4 is a part of BaM. The structure along with experimental diffraction data to 2.0 Å resolution were deposited with good refinement (R_{work}/R_{free} of 21.7/24.7%) and stereochemistry statistics (Molprobity score is 1.56, which corresponds to the 96th percentile). When this model is submitted to PDB_REDO (61), 5 side chain rotamers are changed and the $R_{factors}$ are reduced by about 2.5%.

Applying SLIDER evaluation of the BaM crystallographic data calculating RSCC and Δcontrast, the amino acids that fit the side chain electron density at each residue position are shown in Figure 4. Twelve residues were restricted to hydrophobic amino acids and 17 to hydrophilic due to their chemical environment. An Asp was discriminated from Asn based on salt bridges. Sixty-one, 22, 1, 2, 2 and 2 residues had 1, 2, 3, 4, 5 and 6 amino acid(s) possibilities, respectively.

As the MS data were collected from a mixture of proteins present in the mouse kidney, only two non-overlapping peptides (31.5% of total residues) were found for the sequence of the BaM crystal structure by PatternLab (67) and PEAKS/SPIDER (69,70) (Figure 4). In the case of mouse, given that its proteome and genome are known (79), identification of unique peptides from the BaM sequence in the MS spectrum allows us to infer the rest of its sequence where no peptides were matched. In those cases, where no complete genome or proteome data are available, which is common for most venomous animals, the missing information could be inferred approximately from remote homologue primary structures. Therefore, testing this strategy using the phylogenetic analysis from Consurf (26), we obtained the 300 most similar sequences from UniRef90 (66), and we calculated amino acid occurrence (percentage) per residue position with SLIDER. The most abundant amino acids were considered as possibilities for each residue position (Figure 4). The information from all three techniques was gathered to assign the most probable sequence. Forty-eight residues shared a common assignment in crystallography and/or MS and/or phylogenetic analysis (Supplementary Table S5). The lingering ambiguity from the crystallographic evaluation was resolved by phylogenetic analysis in one residue. Conversely, the ambiguities affecting 35 residues based on phylogenetic analysis were resolved by crystallography (Supplementary Table S6). Thirteen residues remained ambiguous (Supplementary Table S7). Therefore, we show that SLIDER can be applied to membrane proteins and missing information from MS of a purified sample can be substituted by extension of phylogenetic analysis.

Metalloproteinase BmooMP-I

The first novel crystallographic and mass spectrometry data to which SLIDER was applied for natural compounds were

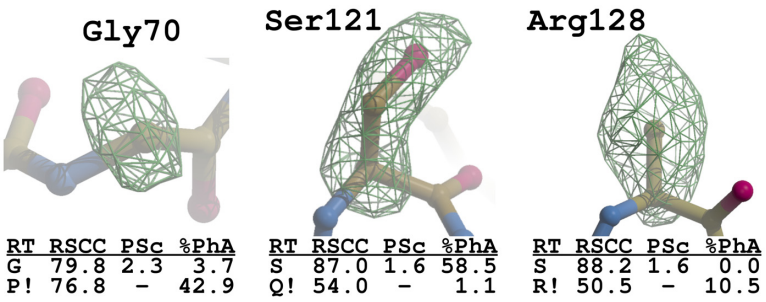


Figure 3. Electron density of omit map of different lysozyme residues in comparison to the PDB code: 6G8A with mass spectrometry and phylogenetic analysis statistics. RT stands for residue type; RSCC for real-space correlation coefficient, PSc for Primary Score, %PhA frequency from phylogenetic analysis; and ! for SLIDER choice for given amino acid in a single letter code.

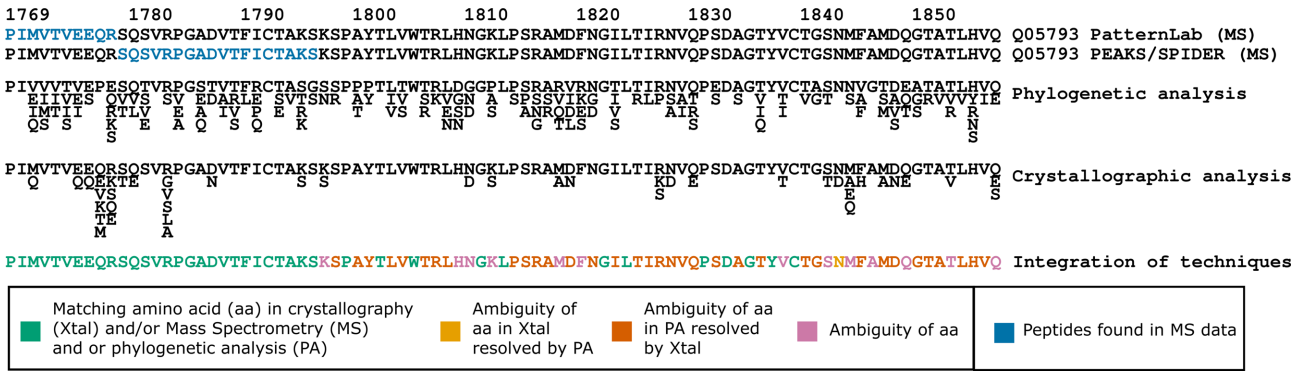


Figure 4. Sequence modelling of BaM. Three techniques, mass spectrometry (with PatternLab and PEAKS/SPIDER evaluation), phylogenetic and crystallographic analysis are integrated to assign the BaM sequence.

from the metalloproteinase BmooMP-I from the Brazilian Lancehead viper *B. moojeni*. BmooMP-I exhibits fibrinogenolytic and gelatinase activities that are important in the immobilization and digestion of the prey (53).

The MS data analysis in a typical strategy using a single digestion procedure, trypsin, against the NCBI database for the Bothropic genus reached a coverage of 131 residues (65.5%) out of 200. The sequences with highest scores were five bothropic metalloproteinases with 57–78 residues matching in the MS spectrum (Figure 5A). With 34.5% of the sequence missing, this was an ideal case for application of the SLIDER methodology to assign sequence heterogeneity.

The BmooMP-I crystallographic model was deposited under PDB code: 6X5X, at 1.92 Å resolution with excellent agreement to the data (R_{work}/R_{free} of 16.0/18.1%) and to the stereochemistry (Molprobity score of 0.89, which corresponds to the 100th percentile) (53). Sequence heterogeneity was removed from the structure to evaluate improvements generated with PDB_REDO (61); it reduced $R_{factors}$ by about 0.9%, changed 4 rotamers and improved the density fit of one residue. When considering the structural environment, the side chain possibilities were restricted to hydrophobic residues for 19 residues (9.5%) that were mostly buried (Figure 5B). Hydrogen bonds were found for 33 residues (16.5%) that were mainly exposed (Figure 5B) and were assigned to the hydrophilic group. The resulting amino acid sequence possibilities and peptide cleavage sites used for database built from crystallographic data are illustrated in Figure 5B giving a total of 419 268 sequences combining the possibilities.

The peptides built from crystallographic data with a spectrum matching the MS data from PatternLab are shown in Figure 5A in PatternLab description with certain residues coloured in green and uncertain residues in magenta. Using the MS data and UniProt database restricted to taxonomy Serpentes with PEAKS/SPIDER, we found 1900 peptide-spectrum matches with a false discovery rate of 1.4%. In order to achieve good sequence coverage, we merged peptides from three sequences, all bothropic metalloproteases, and 6, 137, 42, 12, 2, 2 and 1 residues had none, one, two, three, four, five and six amino acid possibilities, respectively (sequences shown in PEAKS/SPIDER description of Figure 5A). On the other hand, the PatternLab sequence built from crystallographic data strategy had 6, 164 and 32 residues having none, one and two amino acid possibilities, respectively (sequences shown in PatternLab description of Figure 5A). Comparing these two MS evaluations, 103 residues (51.5%) were given the same assignment and 97 (48.5%) were different.

In the crystallographic model, out of 200 residues, 121 (60.5%) favoured a single, well discriminated assignment, confirmed by MS and/or phylogenetic analysis (coloured green in Figure 5A and shown in Supplementary Table S8). More than one possibility was found for 60 residues (30.0%), but their ambiguity was resolved by MS (coloured orange in Figure 5A and shown in Supplementary Table S9). Using the phylogenetic analysis from ConSurf, the

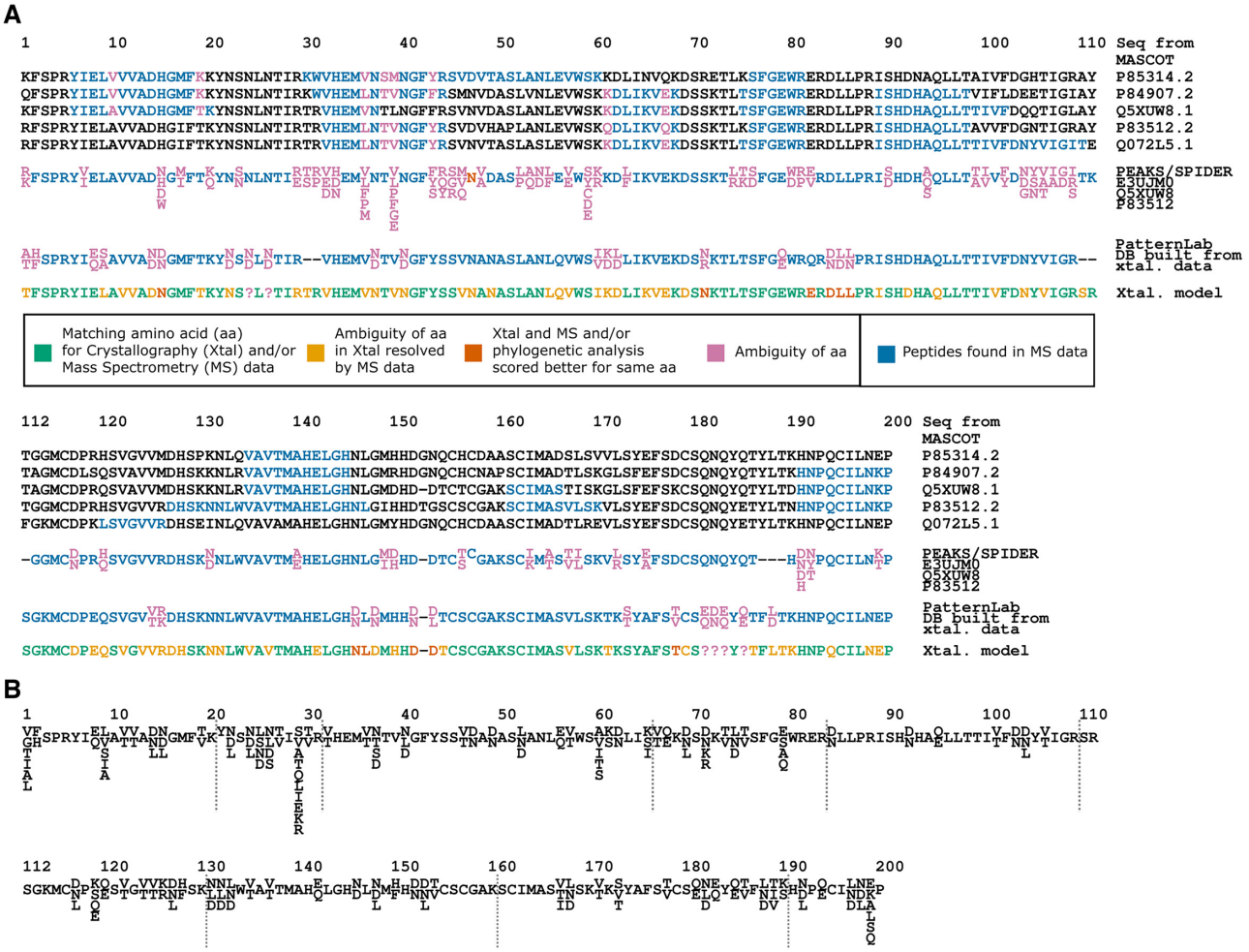


Figure 5. Sequence heterogeneity modelling of BmooMP-I. (**A**) Mass spectrometry data against an NCBI and UniProt database with *Bothrops* genus sequences, against PEAKS/SPIDER evaluation and against PatternLab with database built from crystallography and the sequence from the crystal model. (**B**) Database constructed by SEQUENCE SLIDER from crystallographic data. Adapted from (53).

amino acid 15, 71 and 146 were assigned as asparagine, 82 as glutamic acid, 84, 152 and 153 as aspartic acid, 85, 86 and 147 as leucine, and 178 as threonine (Supplementary Table S10). From the group of residues possessing atomic similarity, the mass spectrometry data and phylogenetic analysis were not enough to distinguish Asn/Asp24 (residue 24 which could be asparagine or aspartic acid), Asn/Asp26, Gln/Glu181, Asn/Asp182, Glu/Gln183 and Gln/Glu185 (coloured magenta in Figure 5A, illustrated in Figure 6 and shown in Supplementary Table S11). As two amino acids were equally likely for these previous residues, SLIDER identified this heterogeneity by leaving them partially occupied in the model. Therefore, SLIDER was able to infer a single amino acid possibility for 194 residues and two possibilities for the remaining 6.

PLA_2-like protein MjTX-I

The MjTX-I is an unique PLA_2-like protein due to an insertion in its C-terminus and 11 mutations that decrease its myotoxicity compared to other PLA_2-like proteins (80). The NCBI database of protein sequences contains various bothropic PLA_2s and using them against MjTX-I MS data, five bothropic PLA_2-like proteins sequences have 92–114 residues matched out of 122 (Figure 7A). There are ten residues identified as different (coloured in magenta in Figure 7A) that provide evidence for isoforms and SEQUENCE SLIDER should resolve these sequence uncertainties.

The MjTX-I crystallographic structure was deposited under the PDB code: 7LYE at 1.85 Å resolution with good agreement to the data (R_{work}/R_{free} of 19.0/23.0%) and to the stereochemistry (Molprobity score of 1.72, which corresponds to the 83rd percentile) (81). The biological implications of the structure were previously published by us (81). PDB_REDO (61) decreased $R_{factors}$ by about 0.2%, changed 4 rotamers, flipped 1 peptide and improved the density fit of 4 residues. Four chains were present in the asymmetric unit, so they were evaluated separately in SLIDER to allow the possibility to distinguish isoforms. For the MjTX-I calculations, side chains of 13 mostly exposed residues (10.7%) had charged atoms in their environment and/or formed one or more hydrogen bond(s), so their possibilities restricted to hydrophilic residues (Figure 7B). The side chains of 5

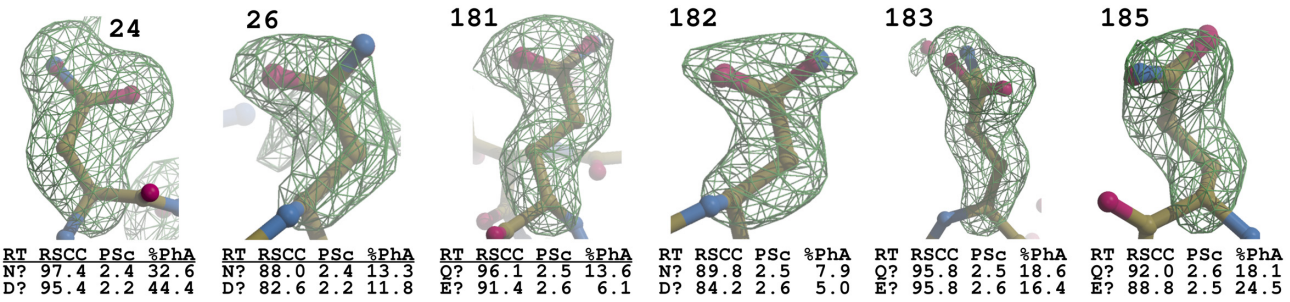


Figure 6. Electron density of omit map with possible residues given crystallographic, mass spectrometry and phylogenetic analysis statistics in metalloproteinase BmooMP-I. RT stands for residue type; RSCC for real-space correlation coefficient, PSc for Primary Score, %PhA frequency from phylogenetic analysis; and ! for SLIDER choice for given amino acid in a single letter code.

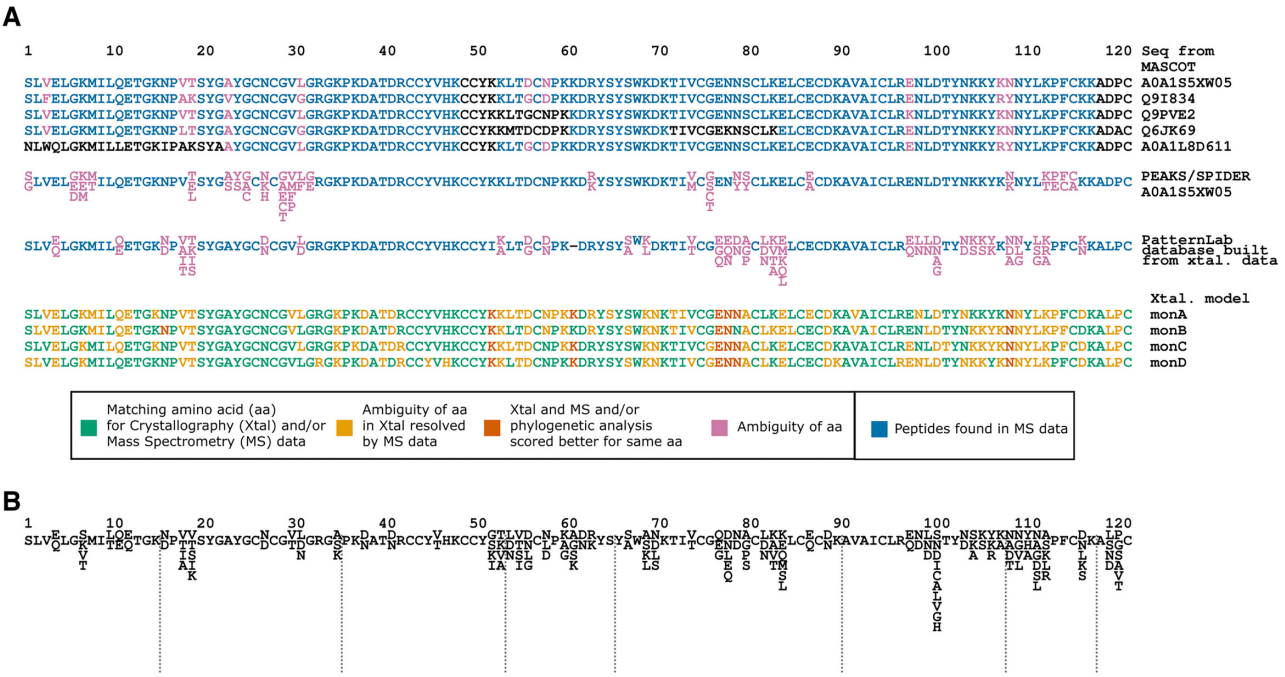


Figure 7. Sequence heterogeneity modelling of MjTX-I. (**A**) Mass spectrometry data against an NCBI database with *Bothrops* genus sequences, against PEAKS/SPIDER evaluation and against PatternLab with database built from crystallographic data and the sequence from the crystal model. (**B**) Database constructed by SEQUENCE SLIDER from crystallographic data.

buried residues (4.1%) were surrounded by hydrophobic residues and were therefore also restricted to the hydrophobic group (Figure 7B). The hypotheses of sequences extracted by SLIDER are shown in Figure 7B; combinations of possibilities generate a total of 2 236 248 sequences.

The application of PEAKS/SPIDER considering the MjTX-I MS data and the UniProt database restricted to Serpentes taxonomy identified 1489 peptide-spectrum matches with a false discovery rate of 1.5%. A single sequence was enough to obtain good sequence coverage, BomoTx (A0A1S5XW05) had 1, 98, 15, 6, 2 and 1 residues with none, one, two, three, four and five amino acid possibilities, respectively (sequences shown in the PEAKS/SPIDER description of Figure 7A). Sequences built from crystallographic data with PatternLab evaluation had more options, 1, 89, 20, 8, 4 and 1 residues with none, one, two, three, four and five amino acid possibilities, respectively (sequences shown in PatternLab description of Figure 7A). Comparing both evaluations, 71 residues (58.2%) were given the same assignment and 51 (41.8%) were different. Moreover, in residue number 116, a deletion was discovered by PEAKS/SPIDER evaluation, showing that MjTX-I isoforms contain at least one protein of 121 residues and another with 122; therefore, the SLIDER restricted database works well to restrict possibilities within a crystal structure but fails to capture overall complexity of isoforms.

Most of the residues in the crystallographic model were resolved, which was confirmed by either MS or phylogenetic analysis (coloured green in Figure 7A under Xtal. model description and shown in Supplementary Table S12). The residues coloured in orange in Figure 7A (Supplementary Table S13) had no single possibility based on RSCC but were resolved based on MS. The side chain choice for 23 residues, out of the 488 in the tetramer, were based on the conjunction of RSCC, MS and phylogenetic statistics

(vermilion residues in Figure 7A and Supplementary Table S14). We did not find clear evidence that the monomers in the tetramer had different sequences; the only variation in fact was lack of electron density for the residues Lys7C–D, Lys52D, Asp56C, Lys60C, Lys61A/C–D, Ser67D, Lys69A–D, Glu77D, Lys83B, Glu84C, Lys105C, Lys106B–D, Asn109C, Leu112A/C–D, Lys113A–C, Pro121C. Also, we generated omit maps in the C-terminal region to evaluate the deletion in residue 116 found in the PEAKS/SPIDER analysis, but electron density did not support this deletion.

By applying this novel methodology, we could identify that the sequence obtained was not yet seen in databases; its highest identity compared to other proteins in PDB is 92.6% with MtTX-II (PDB code: 4DCF) differing in 9 residues of 121. Compared to the crystallographic model of MjTX-I complexed with suramin (PDB code: 6CE2), sequence differences are found for Lys-Thr19, Asp-Ser67, Glu-Asn78, Pro-Ala80, Lys-Leu100, Gly-Asp101, an insertion in 107, Arg-Lys108, Val-Asn110, Gly-Ala119, Arg-Leu120 and Asp-Pro121 and ambiguous residues, Asn-Asp56, Gln-Glu84 and Asp-Asn109 (Figure 8). Compared to the non-redundant protein sequences in BLAST, BomoTx from *B. moojeni* (code A0A1S5XW05.1) shares 96.7% identity with 4 residues being different. Therefore, SLIDER was able to review MjTX-I sequence and curate the database.

Crotoxin CBCol

Crotoxin is one of the most studied snake venom toxins, due to its high abundance in Crotalus venom, and to its toxicity and pharmaceutical potential (82). The database of protein sequences contains various isoforms of the crotoxin, most of them being from *Crotalus durissus terrificus*. Against the MS data from *Crotalus durissus collilineatus*, the NCBI database of snakes gave a maximum of 114 matched residues out of 122 for CBc and CBa2 (both from *C. d. terrificus*), with 8 amino acids being different (magenta in Figure 9A), providing evidence for the presence of isoforms in the purified sample. Applying SEQUENCE SLIDER to the MS, crystallographic and phylogenetic data should improve even more this coverage and resolve uncertainty.

We obtained a crystallographic model of crotoxin crystallized from *C. d. collineatus* at 1.9 Å resolution with good agreement with the data (R_{work}/R_{free} of 20.2/24.3%) and with stereochemistry (Molprobity score of 1.6, which corresponds to the 91st percentile). PDB_REDO (61) reduced its R_{free} by 0.9%, R was not affected, changed 4 rotamers and improved the density fit of 1 residue. Two chains were present in the asymmetric unit and were evaluated separately with SLIDER. Possible choices for 14 mostly exposed residues (11.5%) were reduced to the group of hydrophilic residues supported by the presence of hydrogen bond(s) (Figure 9B). Four buried residues had their possibilities restricted to hydrophobic amino acids. The hypotheses of sequences extracted by SLIDER are shown in Figure 9B, giving a total of 27876 combinations.

In PEAKS/SPIDER application against the CBCol MS data and UniProt database restricted to Serpentes taxonomy, we found 1994 peptide-spectrum matches with a false discovery rate of 1.8%. Two sequences (CBa2 from *Crotalus durissus terrificus* and Cb3 from *Crotalus basiliscus*) had to be considered to obtain a better distribution of the amino acid possibilities that had 2, 4, 24, 92 and 5 residues with 4, 3, 2, 1 and none amino acid possibilities, respectively (sequences shown in PEAKS/SPIDER description of Figure 9A). Sequences built from crystallographic data evaluated with PatternLab had less variation: 1, 3, 11, 107 and 1 residues with 4, 3, 2, 1 and none amino acid possibilities, respectively (sequences shown in PatternLab description of Figure 9A). Comparing both evaluations, 89 residues (73.0%) were given the same assignment and 33 (27.0%) were different. Good agreement is seen between the two methodologies.

Most of the residues of the crystallographic model were resolved to a single possibility confirmed by either MS or phylogenetic analysis (coloured green in Figure 9A and Supplementary Table S15). Out of the dimer (244 residues), 47 residues (19.3%) had no single possibility based on RSCC but were resolved based on MS (orange residues in Figure 9A and Supplementary Table S16). The side chain choices of 9 residues (3.7%) were based on the best RSCC confirmed by phylogenetic analysis (vermilion residues in Figure 9A and Supplementary Table S17). Residue 102 had matches for Asp, Asn and Leu in MS spectrum, but only Leu was observed in remote homologous sequences with 47.5% frequency; therefore, this alternative was chosen. Residue 7A/B and 112A had their side chain atoms removed due to lack of clear electron density.

By applying this novel methodology, we could identify that residues 1, 18, 33, 37, 77, 98, 104 and 105 are different when chain A is compared to B (Figure 10) as SLIDER was able to distinguish these two isoforms of crotoxin and better characterize the crystal sample. These differences would have probably been neglected in a conventional structure elucidation practice. The monomer A is similar to CBa_2 isoform from *C. d. terrificus* (PDB code: 2QOG) sharing 92.6% identity and the monomer B is most comparable to CBd (PDB code: 6TMY) from this same specie with 97.5% identity. In fact, the crotoxin complex consists of the noncovalent association of a crotapotin (CA) acid component of one of its isoforms $CA_{1\text{-}4}$ (the product of different PTM events) with a basic CB, one of its isoforms CBa_2, CBb, CBc and CBd (the consequence of expression of different mRNAs (22)). CA/CBa_2 results in a less stable complex with a higher enzymatic activity and lower toxicity than the other associations. Our model has only basic components, CBa_2 and CBd. These two monomers are similar; their RMSD considering all Cαs is 1.5 Å, which is improved to 0.4 Å by rejecting 23 Cαs. The largest differences are found in loop residues 58–62, which participate in the crystal packing with residue 1 in monomer CBa_2, and with residue 77 in monomer CBd, supporting a hypothesis that isoform preference member of the dimer is favoured by the different crystal packing. Moreover, in monomer CBd, Trp61 is attracted by His1 of helix 1, while in CBa_2, Trp61 is oriented towards hydrophobic residues in the antiparallel helices 2 and 3; therefore, the point mutation in residue 1 seems responsible for the largest observed difference between the two monomers.

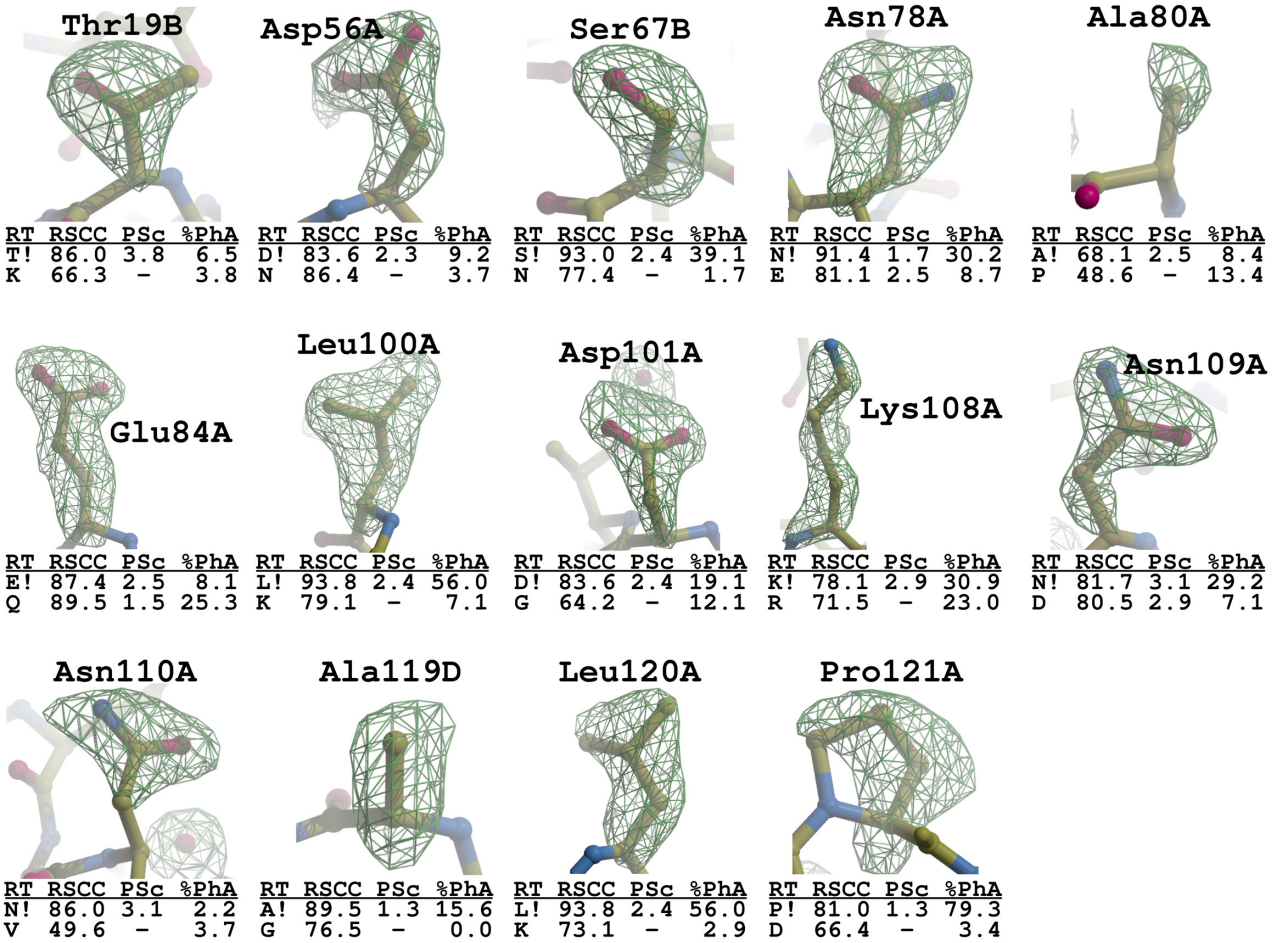


Figure 8. Electron density of omit map of different MjTX-I residues in respect to its previous crystallographic model (PDB code: 6CE2) with mass spectrometry and phylogenetic analysis statistics. RT stands for residue type; RSCC for real-space correlation coefficient, PSc for Primary Score, %PhA frequency from phylogenetic analysis; and ! for SLIDER choice for given amino acid in a single letter code.

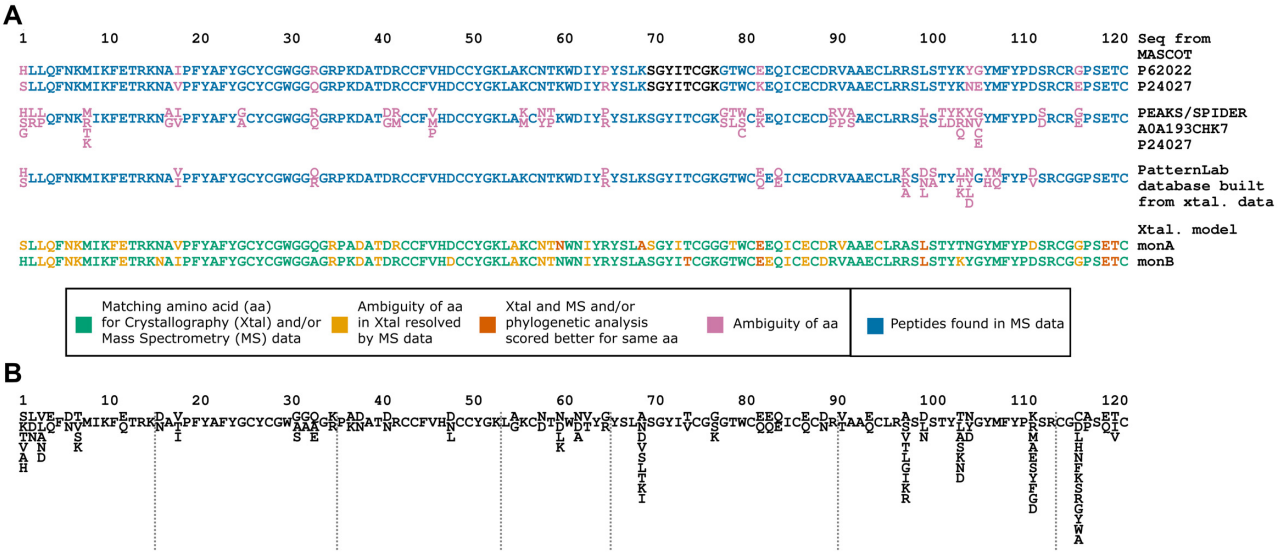


Figure 9. Sequence heterogeneity modelling of CBCol. (**A**) Mass spectrometry data are used against an NCBI and UniProt database with *Crotalus* genus sequences, against PEAKS/SPIDER evaluation and against PatternLab with database built from crystallographic data and the sequence from the crystal model. (**B**) Database constructed by SEQUENCE SLIDER from crystallographic data.

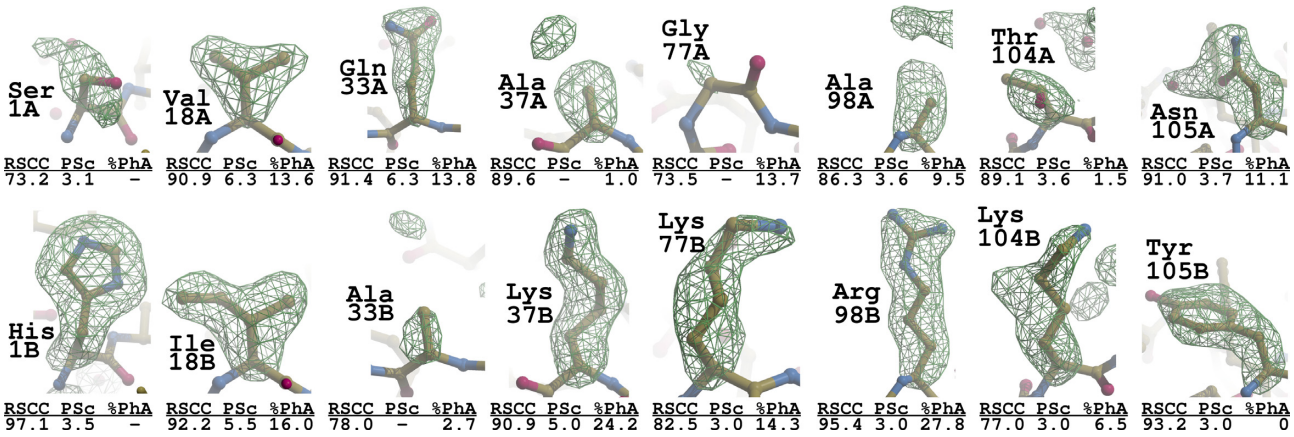


Figure 10. Electron density of omit map of different CbCol residues among its monomers with mass spectrometry and phylogenetic analysis statistics. RSCC stands for real-space correlation coefficient, PSc for Primary Score and %PhA frequency from phylogenetic analysis.

DISCUSSION AND CONCLUSION

The coexistence of isoforms with similar physicochemical properties may prevent complete purification of proteins from natural source. This is typical in toxinology, as toxins are usually obtained from extracted venom, and they are characterized by one of the most rapid evolutionary divergence and variability in any category of proteins (14). In the absence of atomic resolution crystallographic data, the structure elucidation from those samples requires the combination of complementary biophysical methods to account for sequence heterogeneity.

Herein, we repurposed our amino acid evaluator SEQUENCE SLIDER (48) from phasing to integrate crystallography, MS and phylogenetic analysis aimed at the characterization of these complex samples. We use the agreement between calculated and observed electron density (64) to foresee which amino acids are possible in each residue position. From this probability distribution, we generate a sequence database built from crystallographic data that is used against MS data to identify amino acids from the peptides present in the sample (67). Algorithms searching for point mutations in homologues and PTM can also be integrated (69,70). Our approach integrates those instances successfully exploited by Diemer *et al.,* who obtained a preliminary sequence from crystallography and then corrected it by MS data (46) and Guo *et al.,* who estimated residue conservation and frequency using phylogenetic analysis (26), to distinguish aspartic acid from asparagine and glutamic acid from glutamine using homolog sequences in regions where no MS data revealed the correct amino acids. We extend the integrative framework to estimate in an automated program the probability of establishing a single assignment or to acknowledge an existing ambiguity with the available experimental data and prior knowledge.

In the absence of atomic resolution, the small electron scattering difference between O, C and N renders three groups of residues approximately isosteric and indistinguishable in the crystallographic omit electron density maps: (i) valine and threonine; (ii) asparagine, aspartic acid and leucine; (iii) glutamic acid and glutamine (Figure 11) (46–47,83). Depending on the chemical environment of such residues, the ambiguity of groups A and B may be decreased to hydrophobic or hydrophilic residues (blue and orange description in Figure 11) (46). Such restrictions were essential to reduce the number of sequences in the database built from crystallographic data that, for the MjTX-I case as an example, reached more than two million combinations. Moreover, despite the stability of the protein structure and their multiple disulphide bonds, large hydrophilic residues exposed to the solvent may have flexible side chain with no clear electron density, making them difficult to distinguish from small residues, such as Ala and Ser (46). For the MjTX-I model, 28 out of 488 residues had their side chain omitted for lack of density. On the other hand, in high-resolution MS equipment, such as the orbitrap detector used here, the resulting spectrum allows the distinction of these residues; the only residues whose charge and molecular mass are the same are leucine and isoleucine (Figure 11) (46,47). Given the different ambiguities of crystallography and MS, the tandem of the two structural techniques, informed by phylogenetic analysis, should enhance identification of single sequences present in the purified simple, while realistically accounting for lingering ambiguity (46).

These assumptions were confirmed by the datasets collected here, as most of the residues were distinguished either by crystallography itself or by its combination with genetically informed MS. We compared two different strategies, first evaluating a database of sequences built from crystallographic data against MS data with PatternLab (67), and second using the MS spectrum to sequence *de novo* and against UniProt sequences considering possible point mutations and PTM with PEAKS/SPIDER (69). Considering the matched spectrum of peptides, the PatternLab evaluation had fewer possibilities than PEAKS/SPIDER for BmooMP-I and CbCol cases, while lysozyme and MjTX-I were the opposite. In this last case, the generality of PEAKS/SPIDER considering less restriction was able to find peptides in the C-terminus of MjTX-I supporting the presence of at least two different sized proteins, one containing 121 residues and another with 122. While our approach in PatternLab is dependent on crystallographic omit maps, a PEAKS/SPIDER run may discover potential deletions or insertions and is run as an independent analysis. The agreement between the two strategies ranged from half to three quarters of each residue assignment depending on the case.

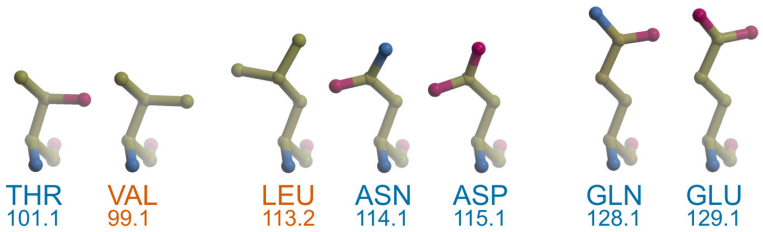


Figure 11. Three groups of residues approximately isosteric sharing atomic structure with their molecular mass, coloured in blue and red hydrophilic and hydrophobic residues, respectively.

The only possible appraisal in respect to the true sequence is for the lysozyme data; PEAKS/SPIDER, with its calculated lower false discovery rate of 0.2%, was slightly better than PatternLab, which had 2.4% of residues wrongly assigned. On the other hand, PatternLab is available to the community for free and may be used for all structural biology groups (67). Moreover, the use of a sequence database built from crystallographic data approximates to a *de novo* sequencing and does not require the presence of remote homologue sequences in databases.

The use of the conservation estimation based on remote homologues (26) provides an additional level of information, although caution is necessary to prevent introduction of bias from what is known from the database of sequences. In fact, the models of BmooMP-I, MjTX-I and CBCol were not identical to any previously known sequence, which highlights the importance of having a proper strategy to model sequence heterogeneity within a crystal model. The analysis for MjTX-I suggested that all four chains are identical differing in four residues compared to another protein from the venom of *B. moojeni* (BooMooTX). In contrast, CBCol was in fact a heterodimer, differing in 8 residues when the two chains are compared. This supports the hypothesis that two isoforms could be distinguished using crystallographic data together with MS by SLIDER. The preference of one isoform in the dimer over the other is supported by the different crystal packing environments. The database of metalloproteinase sequences is smaller compared to the other enzymes and BmooMP-I was also a novel sequence.

Validation is an essential step in structure deposition, and Molprobity (62) and PDB_REDO (61) are widely used to perform a series of evaluations to check if a model is physically plausible. For protein structures obtained from natural sources, SLIDER sub-routines could complement such analyses with sequence validation, as for some residue positions a single assignment may be given whereas for others it may be uncertain. Large and flexible residues with high B-factors may be confused with small amino acids in crystallography. A similar scenario is found in single particle electron cryomicroscopy, as maps show electrostatic potential and negatively charged side chains are not observed (84).

In natural, non-recombinant species, sequence variability should be acknowledged rather than brushed aside. We propose a tool to address this intrinsic characteristic, by integration of different experimental and bioinformatic analyses, thus effectively combining different sources of information. Furthermore, the model deposited should reflect this resulting probability in microheterogeneities. Moreover, the SLIDER approach to test multiple hypotheses of amino acid identities is not exclusive to toxins nor to crystallography. It can also be applied for single particle cryoEM maps, to assign nucleotides in DNA/RNA and to elucidate naturally purified proteins whose sequences cannot be retrieved from genomes.

DATA AVAILABILITY

SEQUENCE SLIDER is an open-source initiative available in the GitHub repository (https://github.com/LBME/slider), in ARCIMBOLDO (http://chango.ibmb.csic.es/) through the package installer for Python (pip) and CCP4.

Mass spectrometry data of BmooMP-I (doi: 10.25345/C55J77), MjTX-I (doi: 10.25345/C59B95), CBCol (doi: 10.25345/C5X22B), BaM (accession code MSV000088503) and lysozyme (doi: 10.25345/C51V5B) are available at Mass Spectrometry Interactive Virtual Environment (MassIVE) https://massive.ucsd.edu/.

Atomic coordinates and structure factors for the reported crystal structures have been deposited with the Protein Data bank under accession number 6X5X (BmooMP-I) and 7LYE (MjTX-I/varespladib).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the use of the Brazilian Synchrotron Light Laboratory (LNLS, Campinas, SP, Brazil) for the use of its facilities, the Institute of Biotechnology (IBTEC/UNESP - Botucatu, SP, Brazil) for the use of its MS facilities, Dr Paulo C. Carvalho for his aid processing the MS data using his software, PatternLab, Prof. Roberta Jeane Bezerra Jorge for sharing mouse kidney MS data, and Prof. Ehmke Pohl and Prof. Randy Read for critically reading the manuscript.

FUNDING

São Paulo Research Foundation [FAPESP, 2015/17286–0, 2019/05958-4, 2020/10143–7]; FAPESP [2016/24191–8 to R.J.B.]; Ministerio de Ciencia e Innovación [PGC2018-101370-BI00 AEI/FEDER/UE]; CNPq [301974/2019–5, 302883/2017–7]. Funding for open access charge: FAPESP [2016/24191–8 and 2020/10143–7]; Ministerio de Ciencia e Innovación [PGC2018-101370-BI00 AEI/FEDER/UE].
Conflict of interest statement. None declared.

REFERENCES

1. Wlodawer,A., Minor,W., Dauter,Z. and Jaskolski,M. (2008) Protein crystallography for non-crystallographers, or how to get the best (but

not more) from published macromolecular structures: protein crystallography for non-crystallographers. *FEBS J.*, **275**, 1–21.
2. Nakane,T., Kotecha,A., Sente,A., McMullan,G., Masiulis,S., Brown,P.M.G.E., Grigoras,I.T., Malinauskaite,L., Malinauskas,T., Miehling,J. *et al.* (2020) Single-particle cryo-EM at atomic resolution. *Nature*, **587**, 152–156.
3. Bartesaghi,A., Matthies,D., Banerjee,S., Merk,A. and Subramaniam,S. (2014) Structure of β-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy. *Proc. Natl Acad. Sci.*, **111**, 11709–11714.
4. Wang,J. and Moore,P.B. (2017) On the interpretation of electron microscopic maps of biological macromolecules: electric potentials involving negative charges. *Protein Sci.*, **26**, 122–129.
5. Ke,Z., Oton,J., Qu,K., Cortese,M., Zila,V., McKeane,L., Nakane,T., Zivanov,J., Neufeldt,C.J., Cerikan,B. *et al.* (2020) Structures and distributions of SARS-CoV-2 spike proteins on intact virions. *Nature*, **588**, 498–502.
6. Radermacher,M., Rao,V., Grassucci,R., Frank,J., Timerman,A.P., Fleischer,S. and Wagenknecht,T. (1994) Cryo-electron microscopy and three-dimensional reconstruction of the calcium release channel/ryanodine receptor from skeletal muscle. *J. Cell Biol.*, **127**, 411–423.
7. Liu,Z., Zhang,J., Li,P., Chen,S.R.W. and Wagenknecht,T. (2002) Three-dimensional reconstruction of the recombinant type 2 ryanodine receptor and localization of its divergent region 1. *J. Biol. Chem.*, **277**, 46712–46719.
8. Yan,Z., Bai,X., Yan,C., Wu,J., Li,Z., Xie,T., Peng,W., Yin,C., Li,X., Scheres,S.H.W. *et al.* (2015) Structure of the rabbit ryanodine receptor ryr1 at near-atomic resolution. *Nature*, **517**, 50–55.
9. Sharma,M.R., Penczek,P., Grassucci,R., Xin,H.-B., Fleischer,S. and Wagenknecht,T. (1998) Cryoelectron microscopy and image analysis of the cardiac ryanodine receptor *. *J. Biol. Chem.*, **273**, 18429–18434.
10. Yan,Z., Zhou,Q., Wang,L., Wu,J., Zhao,Y., Huang,G., Peng,W., Shen,H., Lei,J. and Yan,N. (2017) Structure of the Nav1.4-β1 complex from electric eel. *Cell*, **170**, 470–482.
11. Strynadka,N.C.J. and James,M.N.G. (1996) Lysozyme: a model enzyme in protein crystallography. In: Jollès,P. (ed). Lysozymes: Model Enzymes Biochem. Biol. Experientia Supplementum. Birkhäuser Basel, Basel, Vol. **75**, pp. 185–222.
12. Moss,J.M., Van Damme,M.-P.I., Murphy,W.H., Stanton,P.G., Thomas,P. and Preston,B.N. (1997) Purification, characterization, and biosynthesis of bovine cartilage lysozyme isoforms. *Arch. Biochem. Biophys.*, **339**, 172–182.
13. Zavalova,L.L., Artamonova,I.I., Berezhnoy,S.N., Tagaev,A.A., Baskova,I.P., Andersen,J., Roepstorff,P. and Egorov,Ts.A. (2003) Multiple forms of medicinal leech destabilase-lysozyme. *Biochem. Biophys. Res. Commun.*, **306**, 318–323.
14. Calvete,J.J. (2009) Venomics: digging into the evolution of venomous systems and learning to twist nature to fight pathology. *J. Proteomics*, **72**, 121–126.
15. Guércio,R.A., Shevchenko,A., Shevchenko,A., López-Lozano,J.L., Paba,J., Sousa,M.V. and Ricart,C.A. (2006) Ontogenetic variations in the venom proteome of the amazonian snake bothrops atrox. *Proteome Sci.*, **4**, 11.
16. Daltry,J.C., Wüster,W. and Thorpe,R.S. (1996) Diet and snake venom evolution. *Nature*, **379**, 537–540.
17. Williams,V. and White,J. (1992) Variation in the composition of the venom from a single specimen of pseudonaja textilis (common brown snake) over one year. *Toxicon*, **30**, 202–206.
18. Massey,D.J., Calvete,J.J., Sánchez,E.E., Sanz,L., Richards,K., Curtis,R. and Boesen,K. (2012) Venom variability and envenoming severity outcomes of the crotalus scutulatus scutulatus (Mojave rattlesnake) from southern arizona. *J. Proteomics*, **75**, 2576–2587.
19. Menezes,M.C., Furtado,M.F., Travaglia-Cardoso,S.R., Camargo,A.C.M. and Serrano,S.M.T. (2006) Sex-based individual variation of snake venom proteome among eighteen bothrops jararaca siblings. *Toxicon*, **47**, 304–312.
20. Amorim,F., Costa,T., Baiwir,D., De Pauw,E., Quinton,L. and Sampaio,S. (2018) Proteopeptidomic, functional and immunoreactivity characterization of bothrops moojeni snake venom: influence of snake gender on venom composition. *Toxins*, **10**, 177.
21. Chu,C.-C., Li,S.-H. and Chen,Y.-H. (1995) Resolution of isotoxins in the β-bungarotoxin family. *J. Chromatogr. A*, **694**, 492–497.
22. Faure,G., Guillaume,J.L., Camoin,L., Saliou,B. and Bon,C. (1991) Multiplicity of acidic subunit isoforms of crotoxin, the phospholipase A2 neurotoxin from crotalus durissus terrificus venom, results from posttranslational modifications. *Biochemistry*, **30**, 8074–8083.
23. Shih-Hsiung,W., Fu-Hsiung,C. and Mu-Chin,T. (1983) Separation of the subunits of crotoxin by high-performance liquid chromatography. *J. Chromatogr. A*, **259**, 375–377.
24. Saul,F.A., Prijatelj-Znidarsic,P., Vulliez-le Normand,B., Villette,B., Raynal,B., Pungercar,J., Krizaj,I. and Faure,G. (2010) Comparative structural studies of two natural isoforms of ammodytoxin, phospholipases A2 from vipera ammodytes ammodytes which differ in neurotoxicity and anticoagulant activity. *J. Struct. Biol.*, **169**, 360–369.
25. Chojnowski,G., Simpkin,A.J., Leonardo,D.A., Seifert-Davila,W., Vivas-Ruiz,D.E., Keegan,R.M. and Rigden,D.J. (2022) findMySequence: a neural-network-based approach for identification of unknown proteins in X-ray crystallography and cryo-EM. *IUCrJ*, **9**, 86–97.
26. Ashkenazy,H., Abadi,S., Martz,E., Chay,O., Mayrose,I., Pupko,T. and Ben-Tal,N. (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.*, **44**, W344–W350.
27. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, W299–W302.
28. Marks,D.S., Hopf,T.A. and Sander,C. (2012) Protein structure prediction from sequence variation. *Nat. Biotechnol.*, **30**, 1072–1080.
29. Figliuzzi,M., Barrat-Charlaix,P. and Weigt,M. (2018) How pairwise coevolutionary models capture the collective residue variability in proteins?*Mol. Biol. Evol.*, **35**, 1018–1027.
30. Cocco,S., Feinauer,C., Figliuzzi,M., Monasson,R. and Weigt,M. (2018) Inverse statistical physics of protein sequences: a key issues review. *Rep. Prog. Phys.*, **81**, 032601.
31. de Juan,D., Pazos,F. and Valencia,A. (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.
32. Senior,A.W., Evans,R., Jumper,J., Kirkpatrick,J., Sifre,L., Green,T., Qin,C., Žídek,A., Nelson,A.W.R., Bridgland,A. *et al.* (2020) Improved protein structure prediction using potentials from deep learning. *Nature*, **577**, 706–710.
33. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with alphafold. *Nature*, **596**, 583–589.
34. Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Schaeffer,R.D. *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
35. Stiffler,M.A., Poelwijk,F.J., Brock,K.P., Stein,R.R., Riesselman,A., Teyra,J., Sidhu,S.S., Marks,D.S., Gauthier,N.P. and Sander,C. (2020) Protein structure from experimental evolution. *Cell Syst.*, **10**, 15–24.
36. Simkovic,F., Thomas,J.M.H., Keegan,R.M., Winn,M.D., Mayans,O. and Rigden,D.J. (2016) Residue contacts predicted by evolutionary covariance extend the application of ab initio molecular replacement to larger and more challenging protein folds. *IUCrJ*, **3**, 259–270.
37. Saraswathy,N. and Ramalingam,P. (2011) Protein sequencing techniques. In: *Concepts and Techniques in Genomics and Proteomics*. Woodhead Publishing, Oxford, pp. 193–201.
38. Domon,B. (2006) Mass spectrometry and protein analysis. *Science*, **312**, 212–217.
39. Wilson,D. and Daly,N.L. (2018) Venomics: a mini-review. *High Throughput*, **7**, 19.
40. Calvete,J.J., Juárez,P. and Sanz,L. (2007) Snake venomics. Strategy and applications. *J. Mass Spectrom.*, **42**, 1405–1414.
41. Liska,A.J. and Shevchenko,A. (2003) Expanding the organismal scope of proteomics: Cross-species protein identification by mass spectrometry and its implications. *Proteomics*, **3**, 19–28.
42. Liska,A.J. and Shevchenko,A. (2003) Combining mass spectrometry with database interrogation strategies in proteomics. *TrAC Trends Anal. Chem.*, **22**, 291–298.
43. Standing,K. (2003) Peptide and protein de novo sequencing by mass spectrometry. *Curr. Opin. Struct. Biol.*, **13**, 595–601.
44. Borges,R.J., Salvador,G.H.M., Campanelli,H.B., Pimenta,D.C., de Oliveira Neto,M., Usón,I. and Fontes,M.R.M. (2021) BthTX-II from

bothrops jararacussu venom has variants with different oligomeric assemblies: an example of snake venom phospholipases A2 versatility. *Int. J. Biol. Macromol.*, **191**, 255–266.
45. Cohen,S.L. and Chait,B.T. (2001) Mass spectrometry as a tool for protein crystallography. *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 67–85.
46. Diemer,H., Elias,M., Renault,F., Rochu,D., Contreras-Martel,C., Schaeffer,C., Dorsselaer,A.V. and Chabriere,E. (2008) Tandem use of X-ray crystallography and mass spectrometry to obtain ab initio the complete and exact amino acids sequence of HPBP, a human 38-kDa apolipoprotein. *Protein. Struct. Funct. Bioinf.*, **71**, 1708–1720.
47. Guo,J., Uppal,S., Easthon,L.M., Mueser,T.C. and Griffith,W.P. (2012) Complete sequence determination of hemoglobin from endangered feline species using a combined ESI-MS and X-ray crystallography approach. *Int. J. Mass spectrom.*, **312**, 70–77.
48. Borges,R.J., Meindl,K., Triviño,J., Sammito,M., Medina,A., Millán,C., Alcorlo,M., Hermoso,J.A., Fontes,M.R., de,M. *et al.* (2020) *SEQUENCE SLIDER*: expanding polyalanine fragments for phasing with multiple side-chain hypotheses. *Acta Crystallogr D. Struct. Biol.*, **76**, 221–237.
49. Sammito,M., Millán,C., Frieske,D., Rodríguez-Freire,E., Borges,R.J. and Usón,I. (2015) ARCIMBOLDO_LITE: single-workstation implementation and use. *Acta Crystallogr. D Biol. Crystallogr.*, **71**, 1921–1930.
50. McCoy,A.J., Grosse-Kunstleve,R.W., Adams,P.D., Winn,M.D., Storoni,L.C. and Read,R.J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.*, **40**, 658–674.
51. Thorn,A. and Sheldrick,G.M. (2013) Extending molecular-replacement solutions with SHELXE. *Acta Crystallogr. D Biol. Crystallogr.*, **69**, 2251–2256.
52. Usón,I. and Sheldrick,G.M. (2018) An introduction to experimental phasing of macromolecules illustrated by *SHELX*; new autotracing features. *Acta Crystallogr. Sect. D Struct. Biol.*, **74**, 106–116.
53. Salvador,G.H.M., Borges,R.J., Eulálio,M.M.C., Santos,L.D. and Fontes,M.R.M. (2020) Biochemical, pharmacological and structural characterization of BmooMP-I, a new P–I metalloproteinase from bothrops moojeni venom. *Biochimie*, **179**, 54–64.
54. Hendon,R.A. and Fraenkel-Conrat,H. (1971) Biological roles of the two components of crotoxin. *Proc. Natl. Acad. Sci.*, **68**, 1560–1563.
55. Salvador,G.H.M., Fernandes,C.A.H., Corrêa,L.C., Santos-Filho,N.A., Soares,A.M. and Fontes,M.R.M. (2009) Crystallization and preliminary X-ray diffraction analysis of crotoxin b from *crotalusdurissus collilineatus* venom. *Acta Crystallogr. F. Struct. Biol. Cryst. Commun.*, **65**, 1011–1013.
56. Otwinowski,Z. and Minor,W. (1997) In: *Processing of X-ray diffraction data collected in oscillation mode*. Academic Press, NY, Vol. **276**, pp. 307–326.
57. Kabsch,W. (2010) xds. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 125–132.
58. Afonine,P.V., Grosse-Kunstleve,R.W., Echols,N., Headd,J.J., Moriarty,N.W., Mustyakimov,M., Terwilliger,T.C., Urzhumtsev,A., Zwart,P.H. and Adams,P.D. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.*, **68**, 352–367.
59. Emsley,P., Lohkamp,B., Scott,W.G. and Cowtan,K. (2010) Features and development of coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **66**, 486–501.
60. Kvansakul,M. (2001) Structural basis for the high-affinity interaction of nidogen-1 with immunoglobulin-like domain 3 of perlecan. *EMBO J.*, **20**, 5342–5346.
61. Joosten,R.P., Long,F., Murshudov,G.N. and Perrakis,A. (2014) The *PDB_REDO* server for macromolecular structure model optimization. *IUCrJ*, **1**, 213–220.
62. Davis,I.W., Leaver-Fay,A., Chen,V.B., Block,J.N., Kapral,G.J., Wang,X., Murray,L.W., Arendall,W.B., Snoeyink,J., Richardson,J.S. *et al.* (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.*, **35**, W375–W383.
63. Emsley,P. and Cowtan,K. (2004) *Coot*: model-building tools for molecular graphics. *Acta. Crystallogr. D Biol. Crystallogr.*, **60**, 2126–2132.
64. Liebschner,D., Afonine,P.V., Moriarty,N.W., Poon,B.K., Sobolev,O.V., Terwilliger,T.C. and Adams,P.D. (2017) Polder maps: improving OMIT maps by excluding bulk solvent. *Acta Crystallogr. Sect. D Struct. Biol.*, **73**, 148–157.
65. Merritt,E.A. and Bacon,D.J. (1997) Raster3D: photorealistic molecular graphics. In: *Methods in Enzymology*. Academic Press, NY, Vol. **277**, pp. 505–524.
66. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
67. Carvalho,P.C., Lima,D.B., Leprevost,F.V., Santos,M.D.M., Fischer,J.S.G., Aquino,P.F., Moresco,J.J., Yates,J.R. and Barbosa,V.C. (2016) Integrated analysis of shotgun proteomic data with patternlab for proteomics 4.0. *Nat. Protoc.*, **11**, 102–117.
68. Koenig,T., Menze,B.H., Kirchner,M., Monigatti,F., Parker,K.C., Patterson,T., Steen,J.J., Hamprecht,F.A. and Steen,H. (2008) Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. *J. Proteome Res.*, **7**, 3708–3717.
69. Han,X., He,L., Xin,L., Shan,B. and Ma,B. (2011) PeaksPTM: mass spectrometry-based identification of peptides with unspecified modifications. *J. Proteome Res.*, **10**, 2930–2936.
70. Han,Y., Ma,B. and Zhang,K. (2005) Spider: software for protein identification from sequence tags with *de novo* sequencing error. *J. Bioinform. Comput. Biol.*, **03**, 697–716.
71. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
72. Rozewicki,J., Li,S., Amada,K.M., Standley,D.M. and Katoh,K. (2019) MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res.*, **47**, W5 –W10.
73. Winn,M.D., Ballard,C.C., Cowtan,K.D., Dodson,E.J., Emsley,P., Evans,P.R., Keegan,R.M., Krissinel,E.B., Leslie,A.G.W., McCoy,A. *et al.* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **67**, 235–242.
74. Pražnikar,J., Afonine,P.V., Guncar,G., Adams,P.D. and Turk,D. (2009) Averaged kick maps: less noise, more signal… and probably less bias. *Acta Crystallogr. D Biol. Crystallogr.*, **65**, 921–931.
75. Resource Coordinators,N., Agarwala,R., Barrett,T., Beck,J., Benson,D.A., Bollin,C., Bolton,E., Bourexis,D., Brister,J.R., Bryant,S.H. *et al.* (2018) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **46**, D8–D13.
76. Pereira,J. and Alva,V. (2021) How do i get the most out of my protein sequence using bioinformatics tools?*Acta Crystallogr. D. Struct. Biol.*, **77**, 1116–1126.
77. Steinegger,M., Meier,M., Mirdita,M., Vöhringer,H., Haunsberger,S.J. and Söding,J. (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinf.*, **20**, 473.
78. Colvin,J.R., Smith,D.B. and Cook,W.H. (1954) The microheterogeneity of proteins. *Chem. Rev.*, **54**, 687–711.
79. Church,D.M., Goodstadt,L., Hillier,L.W., Zody,M.C., Goldstein,S., She,X., Bult,C.J., Agarwala,R., Cherry,J.L., DiCuccio,M. *et al.* (2009) Lineage-Specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.*, **7**, e1000112.
80. Salvador,G.H.M., Fernandes,C.A.H., Magro,A.J., Marchi-Salvador,D.P., Cavalcante,W.L.G., Fernandez,R.M., Gallacci,M., Soares,A.M., Oliveira,C.L.P. and Fontes,M.R.M. (2013) Structural and phylogenetic studies with MjTX-I reveal a multi-oligomeric toxin–a novel feature in lys49-pla2s protein class. *PLoS One*, **8**, e60610.
81. Salvador,G.H.M., Borges,R.J., Lomonte,B., Lewin,M.R. and Fontes,M.R.M. (2021) The synthetic varespladib molecule is a multi-functional inhibitor for PLA2 and PLA2-like ophidic toxins. *Biochim. Biophys. Acta Gen. Subj.*, **1865**, 129913.
82. Sampaio,S.C., Hyslop,S., Fontes,M.R.M., Prado-Franceschi,J., Zambelli,V.O., Magro,A.J., Brigatte,P., Gutierrez,V.P. and Cury,Y. (2010) Crotoxin: novel activities for a classic β-neurotoxin. *Toxicon*, **55**, 1045–1060.
83. Chojnowski,G., Pereira,J. and Lamzin,V.S. (2019) Sequence assignment for low-resolution modelling of protein crystal structures. *Acta Crystallogr D. Struct. Biol.*, **75**, 753–763.
84. Marques,M.A., Purdy,M.D. and Yeager,M. (2019) CryoEM maps are full of potential. *Curr. Opin. Struct. Biol.*, **58**, 214–223.