ORIGINAL SCIENTIFIC PAPER

Croat. Chem. Acta **2024**, *97*(1), 15–25
Published online: June 1, 2024.
DOI: 10.5562/cca4083

N-Terminal Derivatization of Peptides with 4'-Formylbenzo-18-crown-6-ether for Protein and Species Identification

Luka Ozdanovac,[1] Renata Biba,[1] Marijana Erk,[1] Amela Hozić,[1] Marta Zrno,[2] Mario Cindrić[1,*]

[1] Division of Molecular Medicine, Ruđer Bošković Institute, Zagreb 10000, Croatia
[2] IT Technology, Business Academy Aarhus, Aarhus 8260, Denmark
[*] Corresponding author's e-mail address: mario.cindric@irb.hr

RECEIVED: March 8, 2024 ⭑ REVISED: April 12, 2024 ⭑ ACCEPTED: April 14, 2024

Abstract: Our research objective was to investigate the use of 4'-formylbenzo-18-crown-6-ether (4fb18C6) for the covalent labelling of peptides for unambiguous peptide identification. Specifically, 23 peptides were analysed using a coupled system of liquid chromatography and tandem mass spectrometry with electrospray ionization to test *de novo* sequencing of derivatized and intact peptides. The reaction was optimized for reductive amination to be performed in an aqueous medium at pH 6 using the microwave radiation. After matching the tandem mass spectra of derivatized and intact peptides in the range of ±0.005 Da, the chemical noise was reduced by up to 90 % and six proteins from 17 different species were identified using the BLAST*p* algorithm (NCBI*nr* and UniProtKB/Swiss-Prot databases). Species identification enabled further accelerated database search using the classical method of mass spectra database matching. The presented method enables reliable *de novo* sequencing of peptides and represents a new tool for species identification.

Keywords: *de novo* sequencing, crown ether, *N*-terminal peptide derivatization, noise reduction, database matching, species identification.

INTRODUCTION

RAPID advances in nucleic acid sequencing have revolutionized biological research by increasing throughput and accuracy while reducing overall costs.[1] However, the development of protein sequencing methods has lagged significantly, although the first complete amino acid sequence dates back to the early 1950s when Sanger and his team elucidated the primary structure of insulin.[2] They devoted nearly a decade to deciphering its sequence through a combination of enzymatic digestion, chemical derivatization with 1-fluoro-2,4-dinitrobenzene (FDNB) and paper chromatography. Around the same time, Pehr Edman developed the first sequential method that involved *N*-terminal derivatization with phenyl isothiocyanate (PITC) and yielded phenyltiohidantoin (PTH) amino acids, which were initially analysed by paper chromatography.[3] In 1967 Edman and Begg introduced the protein sequencer, an automated amino acid sequencing instrument that could perform approximately 15 degradation cycles in 24 hours.[4] Later, the overall efficiency of PTH amino acid detection was increased and improved through the use of reversed phase-high performance liquid chromatography (RP-HPLC) and gas phase sequencing, which reduced the duration of the degradation cycle and the amount of peptide required for analysis.[5] Although other derivatization reagents such as dansyl chloride[6] and *o*-phthalaldehyde (OPA)[7] were used to improve quantification limits, the Edman degradation reaction remained the only sequential method for determining the amino acid sequence in peptides and proteins until the 1990s, when the rise of mass spectrometry (MS)-based proteomics began. The advances in MS technology and the advantages of soft ionization techniques, electrospray ionization (ESI)[8] and matrix-assisted laser desorption ionization (MALDI),[9] enabled MS techniques to take a leading role in protein research due to the speed of analysis, sensitivity and quality of amino acid sequence data obtained from complex protein samples.[10]

With the introduction of soft ionization techniques, the development of methods for the analysis of proteins and peptides has focused on improving the quality of the data sets obtained and on algorithms that facilitate the *de novo* identification of peptides.[11] One way to reduce matrix interferences and increase the signal-to-noise ratio is to generate simpler tandem MS spectra by peptide derivatization. Derivatization involves modifying certain amino acid residues or peptide termini (*N*-/*C*-) to improve ionization and fragmentation in tandem mass spectrometry.[12] In other words, derivatization directs fragmentation towards increased formation of b- and y-series ions, allowing for easier interpretation of tandem mass spectra.[13] One of the most commonly used techniques for successful ion activation is collision-induced/activated dissociation (CID/CAD). The activation in the collision cell is triggered by the interaction between analyte ions and an inert gas, such as argon, helium or nitrogen. The obtained tandem mass spectra provide data on both the peptide precursor ion and the resulting fragment ions.[14] The first significant advance in *de novo* amino acid sequencing by derivatization of the *N*-terminus of peptides with sulfonation was achieved by Keough et al. They introduced two different sulfonation reagents for labelling the *N*-terminus of peptides: the cyclic anhydride of 2-sulfobenzoic acid and (chlorosulfonyl)acetyl chloride.[15] The method has been used effectively to identify proteins from peptide mixtures and to uniquely identify proteins isolated after two-dimensional gel electrophoresis protein separation. Since then, the development of derivatization reagents has focused on the creation of synthetic molecules that react selectively, generate simple spectra, and ultimately facilitate the detection and identification of amino acid sequences.[16–19] With this in mind, this study applies a novel methodology using the well-known Schiff base reaction with the aid of microwave irradiation. Specifically, the *N*-terminus of the peptides is covalently labelled with 4'-formylbenzo-18-crown-6-ether (4fb18C6).

Crown ethers (CEs) are macrocyclic polyethers usually containing 3-20 oxygen atoms separated by two or more carbon atoms.[20] The first CEs were synthesized by J. Pedersen in the 1960s when he studied the coordination chemistry and complexation of vanadium and copper.[21] In addition to their ability to form complexes, especially with inorganic and organic cations, as well as anions, CE serve as sensors.[22]

In the context of using CEs as compounds to facilitate the structural analysis of proteins by mass spectrometry, two techniques stand out. The first technique developed by Juliann et al*.* represents an innovative approach known as selective noncovalent adduct protein probing mass spectrometry (SNAPP-MS), a sensitive method for analysing the structural effects of protein–metal interactions.[23] SNAPP-MS is based on the selective binding of a CE to the protonated forms of basic amino acid residues and facilitates the rapid identification and characterization of protein sequence, structure, and conformational changes. The second technique was developed by Craeser *et al*. in which CEs are used as shift reagents in experiments with ion mobility spectrometers. Shift reagents are defined as species that react or complex with an analyte, resulting in the formation of a lower mobility gas phase ion. This process increases drift time and improves separation.[24] Bohrer and Clemmer conducted a study in which they investigated 18-crown-6 ether (18C6) as a shift reagent for multidimensional ion mobility spectrometry-mass spectrometry (IMS/IMS-MS) analyses of tryptic protein digests. This approach improved the IMS-IMS peak capacity and revealed additional peptides compared to IMS-IMS alone.[25] The results of the theoretical and experimental analyses performed by Chen *et al*. demonstrate the formation of stable non-covalent complexes between 18C6 and protonated amines. The complex formation occurs through hydrogen bonds between the hydrogen atom of the protonated amine and the oxygen of the CE.[26,27] In addition, the results of the study performed by Buhl *et al*. showed that 18C6 forms stable complexes with hydronium ions through linear hydrogen bonds involving three hydrogen atoms of the hydronium ion and three oxygen atoms of the CE.[28] Recent studies indicate that CE can abstract a proton or a protonated molecule from protonated peptides after activation by collisions in argon or electron capture/transfer. Furthermore, these studies have shown that CEs are able to change their binding site after electron transfer.[29] Based on these results, we performed a series of CID-MS experiments in positive ion mode with intact and derivatized peptides using ESI to investigate the effects of labelling the *N*-terminus of peptides with 4fb18C6 as a potential reagent for *de novo* sequencing. The derivatization reaction takes place between the formyl group of 4fb18C6 and the amino group of the peptide under slightly acidic conditions (pH 6). This leads to the formation of an imine, which is subsequently reduced in the second step of the reaction with sodium cyanoborohydride, as described in detail in previous research.[19] In addition to selective labelling the *N*-terminus of peptides obtained by trypsin digestion, the aim of this study was to reduce the large amount of data from MS^E analysis without significantly compromising data quality to enable reliable *de novo* sequencing and species identification through a database search.

EXPERIMENTAL

Chemicals and Materials

Ammonium bicarbonate, formic acid, leucine-enkephalin, isopropanol, porcine pancreatic trypsin, potassium hydroxide and trifluoroacetic acid (TFA) were purchased from Sigma-Aldrich (St. Louis, MO, USA). Acetonitrile (ACN), potassium dihydrogen phosphate and tris(2-carboxyethyl) phosphine (TCEP) were purchased from Merck Millipore (Darmstadt, Germany). Sodium cyanoborohydride was purchased from G-Biosciences (St. Louis, MO, USA). 4'-formylbenzo-18-crown-6-ether (4fb18C6) was purchased from Abcr (Karlsruhe, Germany). Bovine serum albumin (BSA), human serum albumin (HSA), and human plasma transferrin were purchased from Sigma-Aldrich (St. Louis, MO, USA). Elongation factor Tu (EF Tu), isoleucine-tRNA ligase (Ile-RS) and valine-tRNA ligase (Val-RS) were synthesized at the Department of Chemistry, Faculty of Science, University of Zagreb. All chemicals used were of analytical grade. All solutions were prepared with ultrapure water (18 MΩ cm), which was generated in-house using a Milli-Q system from Merck Millipore (Darmstadt, Germany). Resin-free AssayMAP cartridges with a volume of 5 µL were purchased from Agilent Technologies (St. Clara, CA, USA), and mixed cation exchange (MCX) stationary phase was purchased from Waters (Milford, MA, USA).

Peptide Preparation and Derivatization Procedure

The proteins (1 mg) were dissolved in 800 µL 50 mmol L^{-1} ammonium bicarbonate (pH 7.8) and reduced with 200 µL 20 mmol L^{-1} TCEP solution for 15 minutes at room temperature. Protein digestion was performed by mixing 10 µL protein solution (1 mg mL^{-1}) and 1 µL trypsin solution (0.2 mg mL^{-1}) and incubating on the thermal shaker for 18 h at 37 °C and 350 rpm. After digestion, the peptides were dried in a vacuum centrifuge to obtain 10 µg of solid peptides per sample. The dried peptides were dissolved in 30 µL of a freshly prepared derivatization solution containing 23.5 mmol L^{-1} 4fb18C6 and 95.5 mmol L^{-1} sodium cyanoborohydride dissolved in 10 mmol L^{-1} potassium dihydrogen phosphate (pH 6.0). The derivatization reaction was carried out in a domestic microwave oven at 180 W for 8 min, in a closed 2 mL plastic Eppendorf tube. The tube was placed on a plastic stand in the centre of the microwave plate and rotated during the reaction to en-sure homogeneous microwave irradiation of the sample.

Positive Pressure-Micro Solid Phase Extraction (PP-µSPE)

Prior to LC-MS/MS analysis, 70 µL of ultrapure water was added to the derivatized peptides and the resulting solution of 0.1 mg mL^{-1} tryptic peptides was purified using the automated clean-up protocol on the PosiTip liquid handling platform (Bene lab, Zagreb, Croatia). Peptides were purified using a 5 µL resin-free cartridge filled with MCX stationary phase. The clean-up protocol started with conditioning of the stationary phase with 100 µL of the priming buffer (100 % methanol) and 100 µL of the equilibration buffer (ultrapure water). The peptide mixture was loaded onto the cartridge and rinsed in three cycles with 100 µL of the equilibration buffer. Finally, the derivatized and purified peptides were eluted with 25 µL of the elution buffer (10 % ACN in 90 % of 1 % ammonium hydroxide). The purification led to a final concentration of the peptide solution of 0.4 mg mL^{-1}.

NanoUPLC-ESI-QTOF Analysis

Peptide samples were separated on a nanoAcquity UPLC system from Waters (Milford, MA, USA) equipped with a nanoAcquity UPLC 2G-V/M Symmetry C18 Trap Column (100 Å, 5 µm, 180 µm · 20 mm) and a nanoAcquity UPLC BEH C18 Analytical Column (130 Å, 1.7 µm, 100 µm · 100 mm). The column temperature was set to 40 °C and the injection volume was 0.2 µL. The mobile phase A was aqueous 0.1 % formic acid and the mobile phase B was 0.1 % formic acid in 95 % acetonitrile. The isocratic addition of solvent A to the trap column was performed at a flow rate of 15 µL min for two minutes. The samples were eluted under gradient elution conditions with a flow rate of 1 µL min^{-1} and a run time of 32 minutes. The following elution gradient was used: 0–3 min, 80 % solvent A; 3–24 min, 45 % solvent A; 24–27 min, 1 % solvent A; 27–29 min, 80 % solvent A; 29–32 min, 80 % solvent A.

The nanoUPLC system was coupled to the Synapt G2-Si nanoESI-QTOF mass spectrometer (Waters, Milford, MA, USA), and instrument parameters were set using MassLynx software v. 4.1. (Waters, Milford, MA, USA). MS and MS^{E} data were acquired in resolution acquisition mode. The parameters were set as follows: nitrogen flow of 1.1 bar at a source temperature of 80 °C, and the capillary voltage was set to 4.2 kV in positive ion mode. The cone voltage was set to 40 V, and the spectrum acquisition time was 1 s for both MS and MS^{E} analysis. The collision energy for MS^{E} analysis was ramped from 20 to 45 eV. Mass accuracy of the raw data was achieved by infusion of 1 ng µL^{-1} leucine-enkephalin dissolved in isopropanol and 0.1 % formic acid at a ratio 1 : 1, at a flow rate of 0.5 µL min^{-1}. MS data were collected in the mass range from 500 to 3000 Da, while the mass range for MS^{E} was from 50 to 4200 Da.

Noise Reduction and Data Matching

After the analysis of intact and derivatized trypsin-digested peptides, the raw data were exported in the form of fragment ions with corresponding *m*/*z* values, intensities and precursor ion masses. The exported data format served as an input file for the peptide *de novo* sequencing tool. To investigate the potential application of 4fb18C6 as a derivatization reagent for *de novo* sequencing, peptides of different sizes and different amino acid compositions were analysed. Prior to analysis with the *de novo* peptide sequencing tool, an overlay of the above data was performed. By overlaying the data of intact and derivatized peptide fragment ions, chemical noise should be reduced. Chemical noise is defined as any signal that does not contribute to an unambiguous interpretation of the peptide sequence by *de novo* sequencing or its assignment by database alignment.[30] In the overlay, corresponding *m/z* values of derivatized and intact fragment ions were assigned with a mass difference of ±0.005 Da for each individual fragment ion. The aim of this approach was to reduce the large amount of data while maintaining data quality, thus simplifying the identification and determination of amino acid sequences from tandem MS spectra.

Data Processing and *de novo* Sequencing

The unreduced data of intact peptides and the reduced data obtained by the overlap described in the previous section, Noise reduction and data matching, were analysed using the DeNovo Explorer tool, which is part of the GPS Explorer TM software package (v. 3.6 Applied Biosystems, Framingham, MA, USA). The parameters of the analysis were set as follows: enzyme: trypsin, number of sequences: 10, mass tolerance: 0.05 Da. The amino acid sequences obtained from the analysis of both datasets were superimposed in order to determine matches in which the same amino acids occur at the corresponding indices. The usual rule for *de novo* sequencing that isoleucine/leucine (isobaric) cannot be distinguished was taken into account [31], and same was considered for glutamine/lysine (nearly isobaric) as DeNovo Explorer software provides for a minimum mass difference of 0.05 Da, which is less than glutamine/lysine mass difference. As trypsin was used for protein digestion, the above rule does not apply when lysine is at the end position of the peptide sequence. After obtaining partial or complete sequence overlaps from both datasets, all possible sequence permutations were generated, followed by a protein BLAST search against the database.

Database Search by Basic Local Alignment Search Tool (BLAST*p*) Algorithm

Basic Local Alignment Search Tool (BLAST) is a sequence similarity search programme that can be used *via* a web interface (https://www.ncbi.nlm.nih.gov/).[32] The algorithm calculates the similarity between protein sequences using a substitution matrix containing the estimated probabilities of replacing one amino acid with another.[31] All permutations obtained were analysed individually for each peptide and protein/species identification was performed by multisequence alignment. Prior to submitting a query sequence request, unidentified amino acids were assigned an X-letter code. When analysing the peptide sequences, two databases were searched depending on the sample: Non-redundant data protein sequences (nr) and UniProtKB/Swiss-Prot (Swissprot). The search included the analysis of the following parameters; percent of query sequence length included in the alignments against the sequence match (Query coverage), the number of hits or alignments expected by chance with the same or better score (E-value) and the percentage of amino acids identical between the aligned query and database sequences (Percent identity).[33]

Database Matching of BLAST*p* Pre-processed Data

After processing the data with the BLAST*p* algorithm, the obtained results which did not allow unambiguous species identification were further analysed with the commercially available ProteinLynx Global Server software (PLGS; v. 3.0.1, Waters, Milford, MA, USA). Derivatized MS^E datasets acquired in positive ion mode served as input files for PLGS software analysis. Workflow parameters included defining trypsin as the digestion enzyme, allowing two missed cleavages per peptide for the initial database search, and defining 4fb18C6 as the variable modifying reagent. The peptide and fragment tolerances were set to automatic. The false discovery rate (FDR) was set to four. The corresponding databank was created based on the results of the identified species obtained from the database search using the BLAST*p* algorithm. Each created database contained the entire proteomes of the identified species downloaded from https://www.uniprot.org/. The processing parameters included a low energy threshold set at 100 counts, an elevated energy threshold set at 10 counts, and an intensity threshold set at 400 counts. The chromatographic peak width and MS resolution were set to automatic.

RESULTS AND DISCUSSION

Crown ethers have been used in various proteomic studies over the last two decades,[23–29] but to our knowledge never for the derivatization of proteins or peptides. Therefore, their analogue 4fb18C6 was used for the first time for the reductive amination of tryptic peptides. This microwave radiation-assisted reaction was developed to investigate the potential of 4fb18C6 as a reagent for the determination of the primary structure of peptides. After chromatographic separation of the derivatized and intact peptides, the mass spectrometric parameters were optimised for simultaneous data-independent analysis (DIA) of labelled and unlabelled ion species using MS^E data acquisition. It is known that non-covalent complex reactions between 18C6 and peptide facilitated peptide sequencing after complex activation by collisions in argon or electron capture/transfer.[29] Since covalently bound 4fb18C6 should produce the same effect after collisions with argon, our first goal was to select a group of peptides to test this hypothesis. A total of 23 peptides from six proteins with different amino acid compositions were selected according to their length and polarity. The length of the selected peptides ranged from 7 to 24 amino acids and the pI from 4.00 to 8.75, as these ranges covered the most of the bottom-up proteomics experiments (Table 1). The peptides derivatized with 4fb18C6 were selected for tandem mass spectrometry based on the obtained mass increment of 324.1573 Da. After tandem MS analysis, the mass increment was obtained only for the b-ion series while the y-ion series remained intact. This was due to the peptide sequencing direction, where the b-ions contain *N*-terminally bound crown ethers and the y-ions do not.[34] This feature of the tandem mass spectra allowed for easy comparison between derivatized and intact peptides (Figure 1).

Noise Reduction and Data Matching

A detailed analysis of the chromatograms based on peak height and area comparisons showed that the reaction conditions were optimized so that 10–20 % of the tryptic peptides were successfully derivatized with 4fb18C6, while 80–90 % of the peptides remained unmodified. The partial and controlled derivatization of up to 30 % of the tryptic digest allowed the direct comparison of intact and derivatized peptides in the same chromatogram instead of comparing two separate chromatograms obtained from two separate injections and two different sample preparations. To determine whether the derivatization reaction occurred on the side chain of lysine, which also contains the amino group, as well as to inspect the occurrence of double derivatization, data obtained on transferrin was subjected to subsequent bioinformatic processing using the PLGS software. Labelling of the *N*-terminus and side chains with 4fb18C6 was set as a variable modification, and the results showed that out of 31 derivatized peptides, five peptides had side chain derivatization and four peptides had double derivatization (Table S1). However, the intensity of the doubly derivatized peptides was within noise range and did not significantly affect the results. Moreover, the total intensity of peptides selectively derivatized at the *N*-terminus was 2.7 times higher than the total intensity of doubly derivatized peptides and peptides with side chain derivatization.

Automated proteomic analysis was based on the sequencing of peptides using their tandem mass spectra, which consisted mainly of b-product ions and their complementary y-product ions. Identification of the correct b- and y-ion pairs was sufficient to deduce the correct amino acid sequence of the peptide. In reality, however, the lack

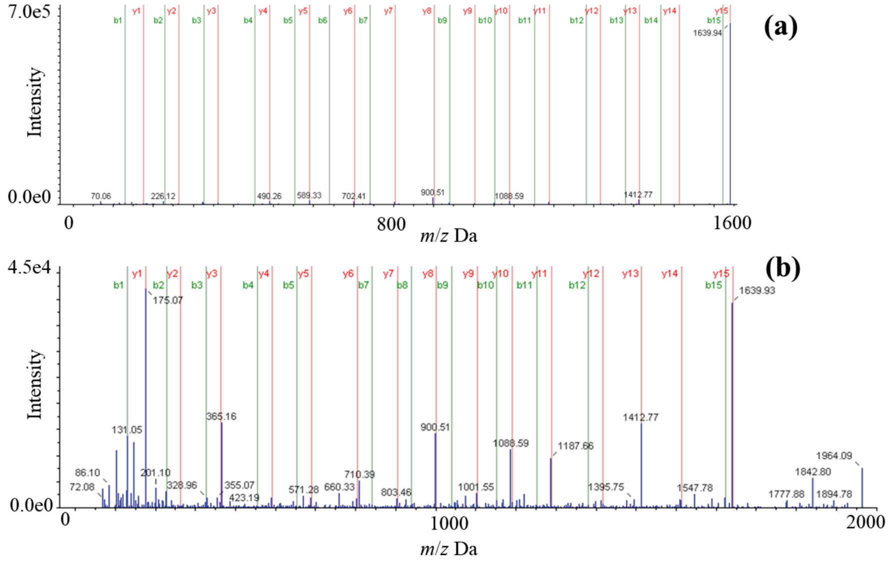

Figure 1. Representative positive tandem mass spectra of human serum albumin (HSA) tryptic peptide KVPQVSTPTLVEVSR: intact peptide ion at *m/z* 1639.9382 (a) and 4fb18C6 (4'-formylbenzo-18-crown-6-ether)–peptide ion at *m/z* 1964.0887 (b). Designated y-ion and b-ion series of peptide fragment ions are presented for both, derivatized and intact ions. Created with BioRender.com.

Table 1. List of selected peptides for tandem MS analysis and their properties (protein of origin, *m/z*, amino acid composition and pI).

Protein	$[M+H]^+$ Da (der)	$[M+H]^+$ Da (intact)	Peptide sequence	Number of amino acids	pI[a]
BSA	1113.6301	789.4714	LVTDLTK	7	5.84
EF Tu	1161.6515	837.4926	EHILLGR	7	6.85
Val-RS	1399.6724	1075.5183	LGNSVDWER	9	4.37
Val-RS	1536.8185	1212.6606	LYKEDLIYR	9	6.07
Transferrin	1302.6453	978.4911	DGAGDVAFVK	10	4.21
EF Tu	1538.7867	1214.6315	FESEVYILSK	10	4.53
HSA	1550.7594	1226.6095	FKDLGEENFK	10	4.68
Ile-RS	1584.7349	1260.5763	EHGSNVWFER	10	5.40
Transferrin	1573.7616	1249.6104	SASDLTWDNLK	11	4.21
BSA	1629.8699	1305.7169	HLVDEPQNLIK	11	5.32
Ile-RS	1759.9598	1435.8032	YVVATELIETVAK	13	4.53
BSA	1891.8939	1567.7407	DAFLGSFLYEYSR	13	4.37
Ile-RS	1738.8632	1414.7085	GVLSHGFALDGEGR	14	5.32
Transferrin	1953.9696	1629.8147	EDPQTFYYAVAVVK	14	4.37
BSA	1991.9664	1667.8094	MPCTEDYLSLILNR	14	4.37
HSA	1964.0887	1639.9382	KVPQVSTPTLVEVSR	15	8.75
EF Tu	2128.0427	1803.8895	GITINTSHVEYDTPTR	16	5.32
EF Tu	2120.1155	1795.9569	TKPHVNVGTIGHVDHGK	17	8.34
Ile-RS	2237.0098	1912.8541	GHMTNEAPGFEGLFYDK	17	4.65
EF Tu	2286.1755	1962.0199	ILELAGFLDSYIPEPER	17	4.00
EF Tu	2441.3184	2117.1638	AIDKPFLLPIEDVFSISGR	19	4.56
Transferrin	2483.1682	2159.0125	IMNGEADAMSLDGGFVYIAGK	21	4.03
Val-RS	2962.5647	2638.4082	SKGNVIDPLDMVDGISLPELLEKR	24	4.44

[a] pI calculation on https://web.expasy.org/compute_pi/.

of complete and regular peptide fragmentation in the available tandem MS data made the data search a complicated task.[35] To make matters worse, it is not always possible to distinguish a b-ion from a y-ion *de novo*.[31] Unresolved ion species and mixed b-ion and y-ion series led to additional nodes (the graph algorithm finds its way from start-to-end through defined amino acid nodes), which complicated the spectral graph and often caused inaccurate results in peptide sequence determination. These complications led to a larger number of possible matches and a higher false discovery rate.[36] Therefore, a critical evaluation of the automated search results is necessary to avoid incorrect assignments.

Statistical data reduction methods included advanced algorithms such as MS-REDUCE, which enabled high-throughput data reduction for mass spectrometry.[37] Additionally, the application of the algorithm for chemical and random additive noise elimination (CRANE), led to a simultaneous increase in the number of identifications and the quantitative accuracy.[38] In our approach (illustrated in Figure 2), however, the two aforementioned data sets were superimposed within a tolerance limit of ± 0.005 Da for the mass difference.

Essentially, two tandem mass spectra of intact and derivatized peptide ions were aligned and ions with an error margin of ± 0.005 Da were selected for further data processing. The ion selection procedure reduced the chemical and non-chemical noise to up to 90 % of all data (Figure 3). Calculations showed that the noise reduction was between 62 % and 90 % without compromising the data quality required for accurate *de novo* sequencing of peptides.

Data Processing and *de novo* Sequencing

Two data sets obtained after tandem mass spectrometry analysis were processed using DeNovo Explorer tool. The

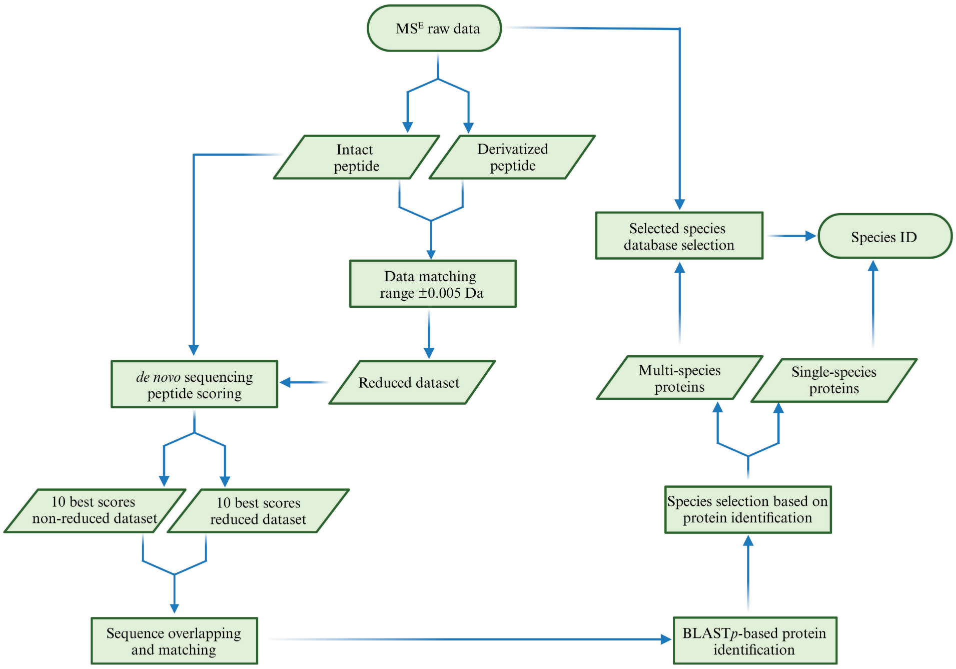
MS^E raw data
Intact peptide
Derivatized peptide
Data matching range ±0.005 Da
Reduced dataset
de novo sequencing peptide scoring
10 best scores non-reduced dataset
10 best scores reduced dataset
Sequence overlapping and matching
BLAST*p*-based protein identification
Species selection based on protein identification
Multi-species proteins
Single-species proteins
Selected species database selection
Species ID

Figure 2. Overview of the bioinformatics workflow. Each field represents a component corresponding to a set of tasks that provide a specific, well-defined functionality. The workflow starts with data-independent acquisition (MS^E is a data independent acquisition technique) of mass spectra, followed by data reduction, *de novo* sequencing of reduced and unreduced data sets, sequence overlap and additional alignment of the selected sequences. Selected peptide sequences are matched against UniProt or NCBI databases (BLAST*p*-based protein identification). The alignment of the peptide sequences with the mentioned databases could define one or more identified species. For more than one identified species, additional database confirmation is required, reducing the number of selected species in the database search by BLAST*p*-based protein identification. Created with BioRender.com.

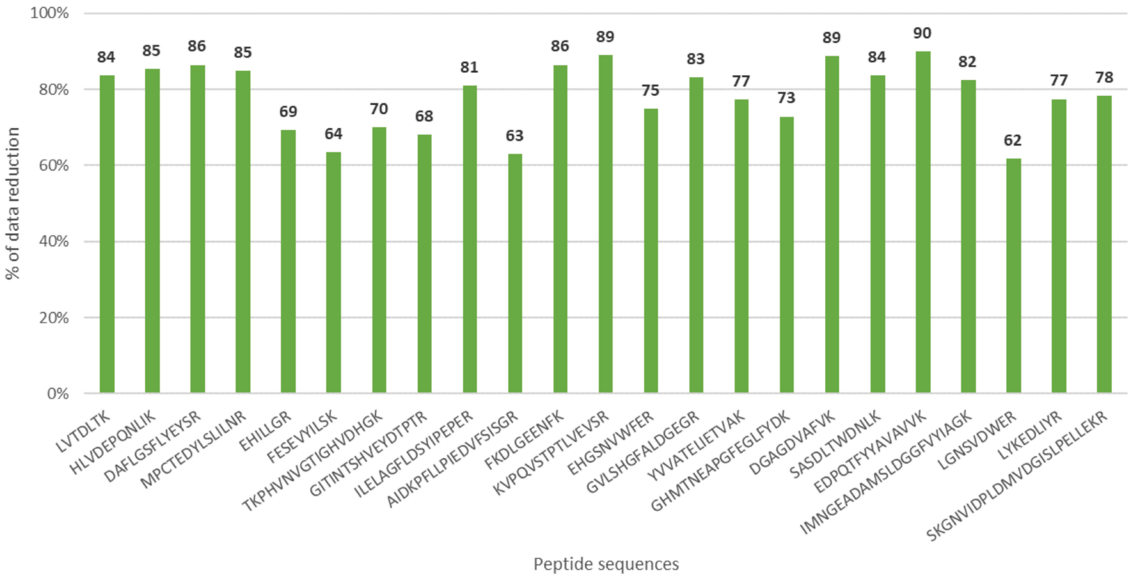

Figure 3. The data reduction was calculated for 23 peptides after matching the DIA tandem mass spectra of the derivatized and intact peptide fragment ions with a tolerance of ±0.005 Da. The procedure removes most of the chemical and non-chemical noise and facilitates the *de novo* sequencing determination of peptides. Created with BioRender.com.

DOI: 10.5562/cca4083

first data set contained fragment ions obtained after analysis of the intact peptide, while the second data set, called the reduced data set, was generated after matching the ions of the intact and derivatized peptide. The processing software assigned a score to each peptide, which was calculated by the graph algorithm after evaluating the spectra list of the measured masses. The score depends on the mass accuracy, the intensity and the probability that the score was achieved by chance. Ten peptides with the best scores from both lists were compared, and the final output contained the amino acid sequence with the most matching amino acids (Table S2-S24). The isobaric amino acids leucine and isoleucine as well as the nearly isobaric amino acids glutamine and lysine (0.036 Da) were considered for further data processing. More specifically, all possible letter combinations Q, K, I and L were exchanged in the selected peptide, so that at the end of the procedure several peptides were generated and used for the NCBI-BLAST*p* database search (Table S2-S24). Non-matching amino acids were labelled with the letter X, where the symbol "X" was used for a position where any amino acid is accepted in the database search (Table 2).

Database Search by Basic Local Alignment Search Tool (BLAST*p*) Algorithm

The amino acid letter chains listed in Table 2 with gaps (marked with "X") and swapped letter combinations Q, K, I and L were applied to the BLAST*p* algorithm for the database search without specifying the organism. The assigned

Table 2. List of identified proteins and species obtained by applying peptides and the BLAST*p* search algorithm against the NCBI*nr* and UniProtKB/Swiss-Prot (SwissProt) databases. Results of the database search includes E-value, database correct sequence, Percent identity and identified species

Protein	*de novo* query[(a)]	E-value	Database correct sequence	Percent identity / %	Identified species
Albumin	MXXTEDYLSLILNR	$5\cdot10^{-6}$	MPCTEDYLSLILNR	85	*Bos taurus*
	LVTDLTK	$4\cdot10^{1}$	LVTDLTK	100	
	HLVDEPQNLIK	$4\cdot10^{-5}$	HLVDEPQNLIK	100	
	DAFLGSFLY	$1\cdot10^{-2}$	DAFLGSFLY	100	
Elongation factor Tu	EHILLGR	$1\cdot10^{1}$	EHILLGR	100	*Escherichia coli* *Shigella flexneri*
	FESEVYIL	$5\cdot10^{-2}$	FESEVYIL	100	
	PHVNVXXIGHVD	$2\cdot10^{-3}$	PHVNVGTIGHVD	83	
	TINTSHVEYDTPTR	$3\cdot10^{-8}$	TINTSHVEYDTPTR	100	
	VFSISGR	$4\cdot10^{1}$	VFSISGR	100	
	ILELAGFXXSYIPE	$2\cdot10^{-5}$	ILELAGFLDSYIPE	86	
Albumin	DLGEE	$2\cdot10^{2}$	DLGEE	100	*Pongo abelii* *Homo sapiens* *Macaca. fasciculari* *Felis catus* *Bos taurus* *Macaca mulatta*
	KVPQVXXPTLVEVSR	$6\cdot10^{-6}$	KVPQVSTPTLVEVSR	87	
Isoleucine-tRNA ligase	EHGSNVWFER	$1\cdot10^{-1}$	EHGSNVWFER	100	*Priestia. megaterium*
	LSHGFALD	$4\cdot10^{2}$	LSHGFALD	100	
	YVVATELIETVAK	$3\cdot10^{-3}$	YVVATELIETVAK	100	
	MTNEAPGFEGLFYDK	$1\cdot10^{-6}$	MTNEAPGFEGLFYDK	100	
Valine-tRNA ligase	LGNSVDWER	$4\cdot10^{-3}$	LGNSVDWER	100	*Escherichia coli* *Salmonella enterica* *Shigella dysenteriae* *Shigella boydii* *Shigella flexneri*
	LYKEDLIYR	$2\cdot10^{-3}$	LYKEDLIYR	100	
	SKXNXXXPLDXXDGISLPELLEKR	$3\cdot10^{-10}$	SKGNVIDPLDMVDGISLPELLEKR	75	
Transferrin	DGAGDVAFVK	$6\cdot10^{-3}$	DGAGDVAFVK	100	*Homo sapiens* *Pan troglodytes*
	SASDLTWDNLK	$8\cdot10^{-5}$	SASDLTWDNLK	100	
	EDPQTFYYAVAVVK	$3\cdot10^{-8}$	EDPQTFYYAVAVVK	100	
	ADAMSXXGGFVYIAGK	$6\cdot10^{-7}$	ADAMSLDGGFVYIAGK	88	

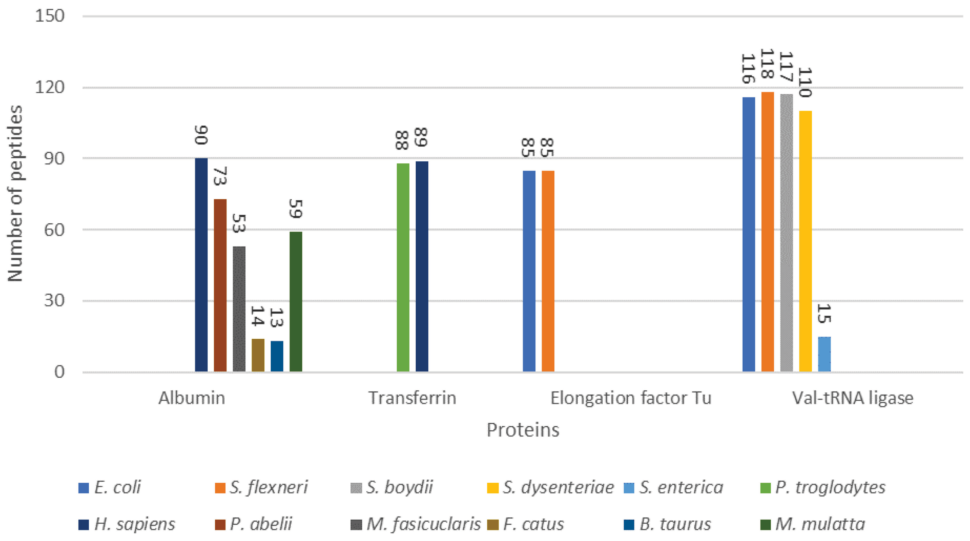

Figure 4. The number of identified peptides obtained through PLGS analysis for proteins categorized by different species based on results obtained from BLAST*p* search against the NCBI database. Created with BioRender.com.

peptide sequences were used for the database search in the NCBI*nr* and UniProtKB/Swiss-Prot (SwissProt) databases. Based on the database search of 23 peptides, six proteins in 17 different species were found. The bovine serum albumin and isoleucine-tRNA ligase proteins provided unambiguous results on protein origin (*Bos taurus* and *Priestia megaterium* species, respectively), while the remaining peptides (elongation factor Tu, albumin, valine-tRNA ligase and transferrin) indicated protein origin but not species. However, the BLAST*p*[39] results restricted further processing of the search by standard database matching with a limited number of species (up to a maximum of six species). The peptide score criteria were evaluated using the percentage coverage of the query (75–100 %) and the E-value, which indicates the number of expected hits of similar quality (score) that could be found purely by chance ($4\cdot10^{2}$–$3\cdot10^{-10}$). The length of the query sequence included in the alignments against the sequence match (Query coverage) was 100 % for all 23 peptides.

It should be emphasised that the exclusion criterion for correct protein and species identification included the results with the highest query coverages and the lowest E-values.

Database Matching of BLAST*p* pre-Processed Data

The results of the NCBI*nr* and UniProtKB/Swiss-Prot (SwissProt) databases using the BLAST*p* algorithm provided unambiguous species identification for two of the six samples analysed. For four other proteins, the BLAST*p* results helped to narrow down the number of species to only two to six species, as described for *de novo* sequencing. In addition, the BLAST*p de novo* database search was able to reduce the number of database matches and significantly shorten database processing when a standard database alignment with a limited number of species was used (Table 2). The standard mass spectrometry database match, which is explained in detail in the experimental section, clearly yielded human albumin as the resulting protein. The database search for transferrin, elongation factor Tu and Val-tRNA ligase, on the other hand, yielded no clear results. In the case of transferrin and elongation factor Tu, however, the database search restricted the origin of the species to only two closely related species, *Homo sapiens* and *Pan troglodytes*, as well as *Escherichia coli* and *Shigella flexneri*. The Val t-RNA ligase protein offered three potential species: *E. coli*, *Shigella boydii* and *S. flexneri*; two of which, *Salmonella enterica* and *Shigella dysenteriae*, were excluded (Figure 4). A longer acquisition time in mass spectrometry, more protein material and a larger number of designated peptides would contribute to unambiguous results, but at this stage of the development of the mass spectrometry method and algorithm, the results presented showed the clear path for further development of proteomics-based biotyping.

CONCLUSION

In this study, a novel method for labelling the *N*-terminus of peptides with 4'-formylbenzo-18-crown-6-ether (4fb18C6) was successfully applied. Since 18-crown-6 ether has a high affinity for the hydronium ion and protonation is essential for electrospray peptide ionization, the 4fb18C6 aldehyde was chosen as the derivatization reagent. The controlled derivatization conditions, in which 10–20 % of the peptides were derivatized while 80–90 % remain unmodified, allowed direct comparison of peptides in the same chromatogram within the same analysis.

Twenty-three peptides were analysed using a coupled system of liquid chromatography and tandem mass spectrometry with positive electrospray ionization. The tandem mass spectra of derivatized and intact peptides were aligned with an accuracy of ± 0.005 Da. After alignment of the data, chemical noise was reduced by up to 90 % and six proteins from 17 different species were identified using the BLAST*p* algorithm (NCBI*nr* and UniProtKB/Swiss-Prot databases). In three out of six cases, the species could be unambiguously identified, while in the remaining cases, *de novo* sequencing and database alignment narrowed down the species identification to two to four potential results. The development of derivatization reagents and advanced algorithms that enable automated determination of amino acid sequences from complex peptide and protein samples is an essential tool in *de novo* peptide sequencing. The importance of this research is demonstrated by the selective labelling of the *N*-terminus of peptides and the reduction of large data sets while preserving their quality through automated systems that enable reliable *de novo* sequencing and species identification *via* a database search.

Acknowledgment. This research was financially supported by the European Regional Development Fund "Qua/Qua Protein" project (grant number: KK.01.1.1.07.0023).

Supplementary Information. Supporting information to the paper is attached to the electronic version of the article at: https://doi.org/10.5562/cca4083.

PDF files with attached documents are best viewed with Adobe Acrobat Reader which is free and can be downloaded from Adobe's web site.

REFERENCES

[1] B. M. Floyd, E. M. Marcotte, *Annu. Rev. Biophys.* **2022**, *51*, 181–200. https://doi.org/10.1146/annurev-biophys-102121-103615

[2] A. O. W. Stretton, *Genetics* **2002**, *162*, 527–532. https://doi.org/10.1093/genetics/162.2.527

[3] P. Edman, E. Högfeldt, L. G. Sillén, P.-O. Kinell, *Acta Chem. Scand.* **1950**, *4*, 283–293. https://doi.org/10.3891/acta.chem.scand.04-0283

[4] P. Edman, G. Begg, *Eur. J. Biochem.* **1967**, *1*, 80–91. https://doi.org/10.1111/j.1432-1033.1967.tb00047.x

[5] R. M. Hewick, M. W. Hunkapiller, L. E. Hood, W. J. Dreyer, *J. Biol. Chem.* **1981**, *256*, 7990–7997. https://doi.org/10.1016/S0021-9258(18)43377-7

[6] J. M. Walker in *Methods in Molecular Biology Proteins Vol 1.* (Ed.: J. M. Walker) Humana Press, Totowa, New Jersey, **1984**, pp. 203–212.

[7] B. N. Jones, J. P. Gilligan, *J. Chromatogr. A* **1983**, *266*, 471–482. https://doi.org/10.1016/S0021-9673(01)90918-5

[8] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, C. M. Whitehouse, *Science*, **1989**, *246*, 64–71. https://doi.org/10.1126/science.2675315

[9] F. Hillenkamp, M. Karas, R. C. Beavis, B. T. Chait, *Anal. Chem.* **1991**, *63*, 1193–1203. https://doi.org/10.1021/ac00024a716

[10] M. M. Vecchi, Y. Xiao, D. Wen, *Anal. Chem.* **2019**, *91*, 13591–13600. https://doi.org/10.1021/acs.analchem.9b02754

[11] C. C. A. Ng, Y. Zhou, Z.-P. Yao, *Anal. Chim. Acta* **2023**, *1268*, 341330. https://doi.org/10.1016/j.aca.2023.341330

[12] V. V. Ilyushenkova, M. E. Zimens, N. Y. Polovkov, A. P. Topolyan, R. S. Borisov, V. G. Zaikin, *Talanta* **2023**, *253*, 123922. https://doi.org/10.1016/j.talanta.2022.123922

[13] B. Kuchibhotla, S. R. Kola, J. V. Medicherla, S. V. Cherukuvada, V. M. Dhople, M. R. Nalam, *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 1216–1226. https://doi.org/10.1007/s13361-017-1606-2

[14] J. Mitchell Wells, S. A. McLuckey, *Methods Enzymol.* **2005**, *402*, 148–185. https://doi.org/10.1016/S0076-6879(05)02005-7

[15] T. Keough, R. S. Youngquist, M. P. Lacey, *Proc. Natl. Acad. Sci.* **1999**, *96*, 7131–7136. https://doi.org/10.1073/pnas.96.13.7131

[16] T. Keough, R. S. Youngquist, M. P. Lacey, *Anal. Chem.* **2003**, *75*, 156–165. https://doi.org/10.1021/ac031274i

[17] A. Butorac, M. S. Mekić, A. Hozić, J. Diminić, D. Gamberger, M. Nišavić, M. Cindrić, *Rapid Commun. Mass Spectrom.* **2016**, *30*, 1687–1694. https://doi.org/10.1002/rcm.7594

[18] K. M. Hassell, J. R. Stutzman, S. A. McLuckey, *Anal. Chem*. **2010**, *82*, 1594–1597. https://doi.org/10.1021/ac902732v

[19] L. Ozdanovac, L. Dončević, A. Hozić, R. Biba, E. Svetličić, A. Janeš, M. Cindrić, *Rapid Commun, Mass Spectrom.* **2023**, *37*, e9534. https://doi.org/10.1002/rcm.9534

[20] M. Kralj, L. Tušek-Božić, L. Frkanec, *ChemMedChem*, **2008**, *3*, 1478–1492. https://doi.org/10.1002/cmdc.200800118

[21] C. J. Pedersen, *Science* **1988**, *241*, 536–540. https://doi.org/10.1126/science.241.4865.536

[22] G. W. Gokel, W. M. Leevy, M. E. Weber, *Chem. Rev.* **2004**, *104*, 2723–2750. https://doi.org/10.1021/cr020080k

[23] T. Ly, R. R. Julian, *J. Am. Soc. Mass Spectrom*. **2008**, *19*, 1663–1672. https://doi.org/10.1016/j.jasms.2008.07.006

[24] M. D. Howdle, C. Eckers, A. M.-F. Laures, C. S. Creaser, *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1–9. https://doi.org/10.1016/j.jasms.2008.10.002

[25] B. C. Bohrer, D. E. Clemmer, *Anal. Chem.* **2011**, *83*, 5377–5385. https://doi.org/10.1021/ac200892r

[26] Y. Chen, M. T. Rodgers, *Anal. Chem.* **2012**, *84*, 7570–7577. https://doi.org/10.1021/ac301804j

[27] Y. Chen, M. T. Rodgers, *Anal. Chem.* **2012**, *134*, 2313–2324. https://doi.org/10.1021/ja2102345

[28] M. Bühl, R. Ludwig, R. Schurhammer, G. Wipff, *J. Phys. Chem. A* **2004**, *108*, 11463–11468. https://doi.org/10.1021/jp045879+

[29] M. Abdelmouleh, M. Lalande, E. Nicol, G. Frison, G. van der Rest, J.-C. Poully, *ChemPhysChem* **2021**, *22*, 1243−1250. https://doi.org/10.1002/cphc.202100075

[30] R. G. Cooks, K. L. Busch, *J. Chem. Educ.* **1982**, *59*, 926. https://doi.org/10.1021/ed059p926

[31] I. O'Bryon, S. C. Jenson, E. D. Merkley, *Protein Sci.* **2020**, *29*, 1864–1878. https://doi.org/10.1002/pro.3919

[32] S. McGinnis, T. L. Madden, *Nucleic Acids Res.* **2004**, *32*, 20–25. https://doi.org/10.1093/nar/gkh435

[33] P. D. Newell, A. D. Fricker, C. A. Roco, P. Chandrangsu, S. M. Merkel, *J. Microbiol. Biol. Educ.* **2013,** *14*, 238–243. https://doi.org/10.1128/jmbe.v14i2.637

[34] J. S. Brodbelt, *Anal. Chem.* **2016**, *88*, 30–51. https://doi.org/10.1021/acs.analchem.5b04563

[35] D. Tabb, D. Friedman, A. J. Ham, *Nat*. *Protoc*. **2006**, *1*, 2213–2222. https://doi.org/10.1038/nprot.2006.330

[36] R. Wu, X. Zhang, R. Wang, H. Wang, *Appl. Sci.* **2023**, *13*, 4604. https://doi.org/10.3390/app13074604

[37] M. G. Awan, F. Saeed, *Bioinformatics* **2016**, *32*, 1518–1526. https://doi.org/10.1093/bioinformatics/btw023

[38] A. J. Seneviratne, S. Peters, D. Clarke, M. Dausmann, M. Hecker, B. Tully, P. G. Hains, Q. Zhong, *Bioinformatics* **2021**, *37*, 4719–4726. https://doi.org/10.1093/bioinformatics/btab563

[39] G. Hu, L. Kurgan, *Curr. Protoc. Protein Sci.* **2019**, *95*, e71. https://doi.org/10.1002/cpps.71