Unveiling the diversity of cyclotides by combining peptidome and transcriptome analysis

Johannes Koehbach[*] and Richard J. Clark

School of Biomedical Sciences, The University of Queensland, 4072 St. Lucia QLD, Australia

[*] *Corresponding author:*
Johannes Koehbach
email: (j.koehbach@uq.edu.au)
phone: +61-3365-2299

Keywords: circular peptides, *in silico* mining, peptidomics, mass-spectrometry

Abbreviations: RiPP, ribosomally synthesized and post-translationally modified peptides; CCK, cyclic cystine knot; MS, mass-spectrometry; LC, liquid chromatography; RP-HPLC, reversed phase high-performance liquid chromatography

Abstract

Circular peptides have attracted much interest in recent drug development efforts, particularly due to their increased stability over linear counterparts. The family of plant cyclotides represents one of the largest classes of naturally-occurring backbone-cyclized peptides displaying exceptional sequence variability and plasticity around three knotted disulfide bonds. Accordingly a multitude of pharmaceutically as well as agrochemically relevant bioactivities have been ascribed to them. Their abundance across various species within flowering plants is highlighted by estimated numbers of up to 150 000 different sequences present in single plant families and over 160 at the species level. However this vast diversity impedes thorough sequence characterization by standard analytical methods using mass spectrometry and thus limits access to a wealth of potentially bioactive compounds that may represent novel lead molecules. Recently the ribosomal origin of cyclotides has been exploited as an alternative way to discover novel sequences. The analysis at nucleotide level allows not only the identification of peptides but also their parent precursor proteins. This combined approach opens access to the discovery of sequences that can provide novel structural templates for a variety of pharmaceutical as well as agrochemical applications. Here we review recent literature related to the discovery of cyclotides. Challenges and opportunities using classical mass spectrometry workflows and novel approaches such as *in silico* mining will be discussed.

Naturally occurring circular peptides – endless discovery

Natural product discovery has a long and successful history and has led to the development of several pharmaceuticals.[1] Most of these molecules are small molecular weight compounds that typically exhibit highly preferred properties such as oral bioactivity and good stability. However over the recent decades peptides have attracted much interest in drug discovery approaches. Peptides have evolved to become versatile macromolecules with a range of functions and can act as hormones, signalling molecules or defence agents.[2,3] From a pharmacological point-of-view they display an evolutionary advantage over solely rationally designed compounds and thus provide unique starting points for the development of novel peptide-based drugs. Their potential has been widely appreciated and they are thought to be able to fill the niche between small molecules (<500 Da) and larger biologicals (>5000 Da).[4] Of particular interest are cyclic peptides as such compounds display increased stability as compared to linear counterparts, one of the major challenges for turning peptides into therapeutics.[5] Interestingly there is a plethora of naturally-occurring circular peptides that have been discovered in all kingdoms of life, ranging from bacteria to plants to mammals.[6,7] Some have already proven to be valuable drugs with the immunosuppressant cyclosporine A being one of the best known examples. Cyclosporine A is a fungal peptide of non-ribosomal origin whereas many peptides of natural origin are ribosomally-synthesized and post-translationally modified (RiPP).[8] This includes for example bacteriocins (e.g. enterocin AS-48 or subtilosin A), mammalian θ-defensins (e.g. RTD-1) and the family of plant cyclotides, which is by far the largest group of circular RiPPs known to date. In addition to their head-to-tail cyclic nature, cyclotides contain six conserved cysteine residues that form three interlocked disulfide bonds, and this combination is referred to as the cyclic cystine knot (CCK) motif.[9] Although the CCK motif confers all cyclotides with a similar three-dimensional fold, there is unique variability and plasticity in the inter-cysteine loop sequences.[10] It is hence not surprising that there is estimated to be several tens of thousands of different cyclotide sequences in existence although only a modest number of around 300 sequences have been published hitherto and deposited in CyBase (www.cybase.org.au), the database of cyclic peptides.[11,12] Several bioactivities have been ascribed to the cyclotides including uterotonic,[13] insecticidal,[14] anticancer[15] or immunosuppressive[16]. Thus, it is apparent that efficient discovery methodologies are crucial in order to harness the mostly untouched potential of this class of naturally-occurring peptides.

Classical chemical analysis approaches for the *de novo* characterization of peptides face serious challenges. Often an extensive characterization of peptides present in biological samples by means of standard analytical techniques such as liquid chromatography and mass-spectrometry

(MS) experiments is both laborious and inefficient. This is mostly due to the high complexity of extracts paired with limited quantities of biological samples and low abundance of active compounds. Although single peptides may be present at a scale of g/kg plant weight, cyclotide concentrations can span several orders of magnitude and some peptide may only be present in trace amounts in the low milli- or even microgram range. Moreover, lack of resolution using chromatographic methods further impedes novel peptide identification. However, in the case of ribosomally-synthesized peptides their identification can also be performed at the nucleotide level. The availability of a continuously growing number of publicly accessible genome and transcriptome datasets, as well as refined bioinformatics methodologies, facilitate a new avenue of peptide discovery that has recently attracted much interest. *In silico* mining has thus been successfully used not only as an alternative but complementary method for novel and potentially bioactive peptide identification within a variety of species.[17-20]

In this review we discuss achievements for the discovery of bioactive circular plant peptides and the current challenges and limitations of MS-based and nucleotide mining workflows. The potential of combining data from *in silico* mining and MS-based discovery approaches to unveil the diversity of cyclotides will be discussed.

Along the circle - MS-based discovery of cyclic peptide sequences

The identification and primary characterization of novel circular peptides even using state-of-the art analytical methods is not a trivial task. Not only are the highly complex samples often present in limited amounts, the nature of a cyclic backbone further poses challenges for MS experiments. For tandem MS sequencing a linear peptide chain is required to retain charges and hence circular peptides have to be derivatized prior to analysis and can be easily overlooked in proteomics analyses. In particular during the early stages of cyclotide research their unusual properties confounded researchers and although the first peptide from the African plant *Oldenlandia affinis*, kalata B1, was discovered and partially characterized in the early 70's[21,22] it took until 1995 until its primary and tertiary structures were fully elucidated.[23] At that time challenging sequence characterizations were similarly experienced for other cyclotides such as cyclopsychotride A[24], violapeptide 1[25] and the circulins A and B.[26] Techniques such as acid hydrolysis prior to Edman degradation were needed suggesting a blocked N-terminus of the analysed peptides. Further distinct mass shifts upon reduction of the peptide and alkylation of cysteine residues as well as a subsequent endoproteinase Glu-C digest that yielded a linear peptide amenable to sequencing provided first evidence of the cyclic nature. MS was initially only used for the detection of the peptide mass but tandem MS sequencing has now become the method of choice for *de novo* characterization. This led to the development of protocols such as inter-cysteine loop sequencing following aminoethylation of cysteine residues[27] or partial acid hydrolysis prior to liquid-chromatography (LC)-MS analysis.[28] As an alternative to LC-MS,[29-33] matrix-assisted laser desorption ionization (MALDI) MS has been successfully used in bottom-up *de novo* characterization studies.[18,34-39] MALDI MS has also been shown to be useful for quantification studies[40] and less frequently used in top-down sequencing approaches.[41] Other methods include the combination of nano-LC with fourier-transform MS allowing high mass accuracy analysis without the need for enzymes.[42] The advances in MS go hand in hand with the increasing numbers of cyclotide discovered. Studies performed in the 1990s typically reported only a few peptides per study but it was quickly realised that cyclotides occur in a library-like manner with many different variants present in a single plant species.[43,44] This provided incentive to develop efficient MS methods for thorough analyses to unveil the full peptide cocktail in cyclotide-producing plant samples. Additionally, as cyclotides seem to be widely distributed among flowering plants there was also a need for a robust and reliable screening procedure that allows the rapid and accurate analysis of large sample numbers with regard to the presence or absence of cyclotides.

In 2008 Gruber *et al*. presented a screening methodology that involves the analysis of plant extracts by high-performance (HP)LC and MS.[45] This widely used and optimised workflow for MS-based peptide discovery is outlined in Figure 1A and makes use of the typical features of cyclotides, i.e. their hydrophobic properties on reversed phase (RP)-HPLC, the typical mass ranging between 2500-4000 Da and the six conserved cysteine residues. It involves the solvent extraction of fresh or dried plant samples followed by HPLC and MS analysis. If late eluting peaks are observed that contain compounds in the expected mass range, reduction and alkylation is carried out to confirm the presence of the six conserved cysteine residues. Samples are then subject to further purification and/or enzymatic digestion and tandem MS experiments. Subsequently, the distinctive fragmentation pattern between linear versus cyclic and knotted peptides were reported as a valuable tool and additional identification criterion.[37] It has also been acknowledged that at least a partial sequence of two or more adjacent loops should be obtained prior to characterizing a plant as a cyclotide-containing species.[18] This was deemed necessary as other disulfide-rich peptides could fulfil all of above mentioned criteria but represent other peptide classes such as for example thionins[46] or knottins.[47] Often the 'ring-opening', i.e. the linearization of the peptide that is required for tandem MS analysis of cyclotides is used as a further identification criterion. Almost all cyclotides contain a conserved glutamic acid residue that upon treatment with endoproteinase Glu-C results in a single cleavage and the addition of a water molecule resulting in a distinctive mass shift. It has to be noted that this is not applicable for peptides containing multiple or no Glu residues and linear cyclotides, also referred to as acyclotides[38] or uncyclotides.[48] Accordingly a combination of several identification criteria and importantly sequence information is required for an unambiguous identification of cyclotides.

Although this optimised screening methodology (Figure 1A) allows the analysis of large sample numbers, the last step, i.e. tandem MS sequencing typically is the most challenging and rate limiting one. This is mostly due to the high complexity of samples together with the low abundance of several compounds and an overall limited sample amount. Due to the high similarity of peptides that may only differ in single amino acids separation via RP-HPLC can be very inefficient and hence make the use of mixtures inevitable. Within such unfractionated samples peptides of the same mass and/or *m/z* ratio may co-elute and hence confound interpretation of spectra. To overcome such limitations a combination of different single and double protease digests can prove useful and may allow sequence characterization without the need for laborious separation.[34] Another approach to accelerate the time-consuming manual *de novo* discovery is to implement automated database search workflows of MS data as described by Colgrave *et al*.[49] Briefly, a custom-made database containing replications of all known cyclic peptide sequences allows the calculation of all possible proteolytic fragments by search engines and then tandem MS data are searched against this database. Overall MS sequencing approaches typically yield only about 10-20 sequences[34,48,49] although recent work using LC-MS estimates around 70 unique cyclotide masses per species.[16] This clearly indicates the need for continuous improvements to optimise discovery workflows.

There have been advances in MS analysis, including hard- and software as well as workflow strategies, but MS-based cyclotide analysis still remains challenging, especially if focused on thorough sequence characterization and when unfractionated or limited samples are used. Newer strategies, in particular the above mentioned automated database analysis of tandem MS data[49] are helping to overcome these challenges and to avoid laborious and often inefficient separation work. Recent work demonstrated the high-resolution performance of today’s mass spectrometer by identifying 126 putative cyclotide masses within a crude peptide extract using LC-MS deconvolution. After a simple round of preparative RP-HPLC tandem MS evidence was obtained for a total of 82 different peptides.[50] Interestingly numerous linear variants as well as peptides carrying post-translational modifications such as aspartate methylation, glutamate ethylation, tryptophan oxidation, as well as deamidation of asparagine and glutamine were identified. It has to be noted that the optimised database used for the automated search included nucleotide-derived sequences as will be described later. Recently the detection of PTMs using MS technologies revealed for the first time the presence of glycosylated forms of cyclotides[51] within species of the violet family and thus further expanding the pool of cyclotides to be discovered. Although MS-based analysis is indispensable for cyclotide characterization it faces limitations with regard to fully unveiling their sequence diversity among flowering plants. This becomes apparent with estimated numbers of different sequences ranging between 9000 and 150 000 for individual plant families[45,52] and up to a theoretical library of 600 million sequences.[51] Thus it is evident that discovery solely based on MS is not able to fully harness the potential of this vast sequence variability. Fortunately cyclotides are RiPP and thus transcribed and translated gene products. This allows their identification not only at peptide but also nucleotide level presenting alternate pathways for peptide discovery.

Advantages of nucleotide analysis and *in silico* discovery

Cyclotides arise from larger precursor proteins and cDNA sequencing approaches have led to the discovery of precursor sequences in *Oldenlandia affinis*,[53] and several species of the family of Violaceae[54-56] as well as Fabaceae[35] and Poaceae.[57] These studies not only discovered novel peptides but also provided a better understanding of the underlying gene architecture and biosynthetic processing. Most of the discovered peptides were eventually confirmed using MS methods. Besides cDNA library approaches the use of *in silico* studies provide an alternative for peptide discovery. In the past such studies were restricted to a low number of available nucleotide datasets, whereas now there is a large amount of accessible data, which has created the new challenge of how to best filter this information. The last decade has seen a rapid development of next-generation sequencing services, and in particular transcriptome analyses, which has opened new avenues for peptide discoveries. RNA sequencing has become reasonably affordable and is on the way to becoming part of routine experiments. Sequencing and subsequent assembly of the raw data still requires advancement of bioinformatics methods as for many organisms, such as plants or invertebrates, reference genomes are not available. This makes *de novo* assembly a less than trivial task.[58,59] However continuous development of automated workflow programs, user friendly software and webportals are being created to meet the emerging needs of data mining and allow also non-bioinformaticions to access or self-process these data.[60] In this regard it is worth noting that *de novo* assembly requires high performance computing facilities and thus may incur additional costs. However, if restricted to publicly accessible databases *in silico* mining is amenable at virtually no cost. The workflow for discoveries based on nucleotide sequencing is shown in Figure 1B.

One of the first studies making use of such *in silico* mining techniques for the discovery of cyclotide-like sequences was in crop plants.[61] Homology searches using the freely available BLAST tool[62] were performed using a custom-made query set of known cyclotide sequences. This approach led to the identification of so-called cyclotide-like sequences in a number of monocot plants. Although none of the *in silico* hits were confirmed at the peptide level, the findings were intriguing. Several hits from *Oryza sativa* (rice), *Triticum aestivum* (wheat) or *Zea mays* (maize) showed a similar precursor structure with regard to the putative peptide domain as those described from *Oldenlandia affinis*, however a significantly shortened or unusual C-terminal tail was observed. This led to speculations that the evolution of cyclotides among flowering plants started prior to the divergence between monocotyl and dicotyl plants. However, large-scale screening studies investigating the distribution of cyclotides suggest multiple independent gain-of-function mutations as the evolutionary explanation for cyclotide occurrence.[18,45]

Beside such evolutionary insights, nucleotide analysis also provides information regarding pathways underlying the *in planta* biosynthesis of analysed peptides. Embedded in larger precursor molecules the cyclotide domain must undergo several steps including the excision, cyclization and disulfide bond formation to yield the mature peptide. In particular conserved residues within the N- and C-terminal flanking regions have been identified and play a major role in the cyclization via an asparaginyl-endopeptidase mediated mechanism.[63-65]

More recently BLAST searches against the EST database at NCBI revealed cyclotides within the Solanaceae plant family.[38] Screening of EST data also has proven useful for describing the abundance of cyclotides in *Oldenlandia affinis* with the identification of 31 precursor sequences as well as enzymes involved in their biosynthesis.[66] An even greater number of novel peptide sequences have recently been reported for two violet species. This includes 53 novel cyclotides from *Viola baoshanensis* that were characterized from both sequencing a cDNA library and transcriptome analysis.[55,67] A thorough analysis of the cyclotide containing plant *Viola tricolor* led to the identification of 98 precursor sequences that encode for 108 different peptides underpinning the high identification capacity of *in silico* mining.[50] This includes both peptides that exhibit low similarities but also examples differing in single amino acid residues. Subtle differences, i.e. isobaric residues such as leucine/isoleucine and amino acid isoforms of aspartic acid/aspartate or glutamic acid/glutamine are readily found using automated searches of nucleotide datasets. Such BLAST searches are commonly used for the identification of homologuous peptides and proteins.

In order to specifically harness the sequence information of cyclotides hidden in publicly available datasets customized scripts such as CyPerl or CyExcel[68] have been developed and tailored approaches using regular expression searches have been applied.[69]

Lastly another advantage of *in silico* discovery is the low amount of sample that is needed to perform such analyses. As discussed, the presence of several tens of peptides with highly similar sequences is a major bottleneck in MS-based approaches especially when sample amounts are limited. The amounts of RNA required for next generation sequencing can be obtained from as little as single leaf samples enabling thorough analyses of complex peptide cocktails without the need for large amounts of plant material.

Combined -omics approaches allow seamless discovery

As described above, nucleotide analysis is a powerful tool for the discovery of novel peptides. However there are fundamental limitations if one solely relies on genome or transcriptome derived data as is the case for studies restricted to peptidome analysis. Thus it is evident that a combined analysis that makes use of the advantages of either approach (Table 1) is likely to accelerate the discovery of novel peptides. The complementarity of *in silico* and MS-based approaches is illustrated in Figure 2.

Firstly, one cannot be sure if an *in silico* discovered peptide is present within a plant until there is MS evidence and secondly the cyclic nature and presence of other posttranslational modification needs to be confirmed using above described methods such as by analysis of tandem MS fragmentation pattern and distinct mass shifts upon enzymatic digestion. The identification of precursor sequences that lack a C-terminal tail within the monocot plant *Panicum laxum*[57] is a good example highlighting these complementary advantages of *in silico* and MS discovery. Whereas the lack of or an unusual C-terminal tail sequence may suggest the linear nature or an alternative cyclization mechanism only MS data can provide evidence. Similarly a peptide from *Viola tricolor* that has been found to be linear could be matched to a precursor sequence lacking a C-terminal tail sequence.[50] It is worth mentioning here that recently bioinformatics tools have been reported that can predict cyclic peptides with high confidence.[70] The characterization of peptide sequences that can be retrospectively matched to precursor sequences not only provides the peptide's correct length, but also reveals processing sites and thus insights into biosynthetic mechanisms. In turn the correct assignment of amino acids, in particular isobaric residues leucine/isoleucine or isoforms of amino acids such as asparagine/aspartic acid or glutamine/glutamic acid obtained from *in silico* data can help to dissect ambiguities from MS analysis and therefore confirmatory experiments such as amino acid analysis would not be required. This can prove difficult, since as pointed out earlier the separation and purification can be challenging if not unfeasible. To unveil the diversity and increase the number of peptides identified per plant a recent study by Hellinger *et al*.[50] highlighted the usefulness of automated database searches from transcriptome derived sequences for an efficient analysis of tandem MS data. The use of a carefully and continuously updated customised cyclotide database has led to the identification of more than 80 cyclotide sequences within HPLC fractions of a single violet plant. This was achieved by initially mining the transcriptome of *Viola tricolor* that allowed the identification of 55 novel full-length cyclotide sequences that were added to an existing database to yield a total of 367 sequences against which tandem MS data were searched. More than 30 peptides initially identified at transcript level could be confirmed at the peptide level. Overall this study combined LC-MS deconvolution, tandem MS sequencing and transcriptome analysis and identified a total of more than 160 cyclotides in a single plant species. Other automated approaches include for example an algorithm termed ‘Cycloquest’ that has been developed by Mohimani *et al*.[71] It aims to link genomic information to MS data and works similarly to Sequest or Mascot but is specifically tailored for the identification of cyclic peptides. More recently such algorithms have been optimized for the analysis of peptides that produce poor MS fragmentation patterns and have a high number of post-translational modifications such as lanthipeptides.[72]

It is evident that combining nucleotide analysis and proteomic data provide a ‘seamless’ discovery of novel peptides due to their complementary nature of advantages and drawbacks. Continuous development of computational methods to predict peptides from transcriptomes provides the essential basis for the identification of peptides in biological samples using MS. This streamlined workflow allows a rapid and accurate analysis of complex samples for not only cyclotides but also other classes of naturally-occurring and ribosomally-synthesised peptides.

Endless peptides and never-ending questions

There is an increase in the speed and number of peptides discovered using the above described combined -omics approaches. The challenge remains on how best to harness this sequence diversity for practical applications. The wealth of compounds awaiting characterization not only provides a broad range of starting points for bioactivity testing, but also poses a number of questions and difficulties.

Notably there is often a certain level of mismatch observed between sequences obtained from transcriptome and proteome data.[50,73] Firstly, putative mature cyclotide domains within nucleotide datasets are always assigned based on homology with already known sequences. In the case of novel or unusual N- and C-terminal flanking residues this can lead to the incorrect assignment of shortened or elongated sequences for which no MS evidence can be obtained. Highly post-translationally modified peptides may also prove difficult to match to parent precursor sequences. Importantly it has to be mentioned that the presence of peptides may differ within different samples of the same species. This can be due to tissue-specific expression, seasonal changes in peptide expression or exposure to different environmental conditions that might affect expression patterns both at transcript and peptide levels.[44,73-75] Hence samples used for combined proteome/nucleotide approaches should ideally originate from the same plant and the same tissue to reduce mismatching, although even this cannot rule out discrepancies between precursor and mature peptide sequences.[73] Other possibilities that can cause a mismatch are sequencing or assembling errors for nucleotide derived data or the creation of peptide artefacts due to harsh experimental handling of samples. Further the choice of solvent used for the extraction may also bias the obtained amount as well as the peptide profile. Although a mixture of methanol and DCM is commonly used for the extraction of cyclotides this solvent did not give comparable yields for CCK peptides from *Momordica cochinchinensis* seeds as were obtained with an acetonitrile/water mixture.[76] Therefore a careful selection of plant samples and extraction methods is vital to guarantee that samples subject to MS analysis are representative of the plant's peptide content.

Another important question is how the identification of such a wealth of novel peptides can efficiently be used for drug development. For screening approaches that are not based on any specific bioactivity the question remains as to what extent large-scale discovery approaches are useful in speeding up the development of novel peptide drugs. Extensive analyses of an ever-growing number of sequences may help to identify residues that play important roles in maintaining structure as well as having possible implications for biological activity or amenability to chemical modifications. For example it is evident that loop 6 within cyclotides shows a particularly high sequence variation and incorporates the native cyclization site (Figure 1). It is therefore not surprising that most studies that exploit the amenability of cyclotides to peptide grafting are in this region.[77] Moreover, together with data from resolved NMR structures QSAR models for example can provide useful information as has been recently shown for linking physicochemical properties of cyclotides to their cytotoxic and anthelmintic activities.[78] Such computational methods may prove particularly useful in assessing a novel peptide's probability to exhibit an interesting bioactivity profile prior to testing it in *in vitro* assays.

Despite the increased number of peptides discovered several questions remain unanswered. Why do so many, but not all plants produce cyclotides, and why so many? Their occurrence in a variety of species across different plant families and both in monocotyl and dicotyl plants suggest they evolved as multiple independent gain-of-function events. However as to now this benefit for cyclotide-producing plants is still not fully understood. Their reported insecticidal properties suggest a role within the plant defence machinery[53,79,80] but knowledge regarding underlaying mechanisms-of-action and specific targets remain limitied. They have been found to be able to interact with membranes[81,82] and only recently it has been shown that native cyclotides are able to modulate G-protein coupled receptor activity allowing speculations regarding a possible role as signalling molecules.[83]

Conclusion

The discovery and application of circular peptides are currently widely appreciated in biomedical research. Cyclotides have proven to be a uniquely versatile scaffold for the development of peptide drug leads. The comprehensive identification and characterization of cyclotides as well as their precursor proteins from a variety of plant species has revealed many potent drug leads and provides valuable information about their biosynthesis, distribution and evolution. Continuous development in MS technologies together with refined molecular biology methods and next generation sequencing allows the identification of novel peptides both in high numbers and with unique accuracy. The analysis of peptidomes at genome or transcriptome level reveals a novel wealth of peptides of previously unexpected dimensions. This provides researchers with countless possibilities and hence unveiling the diversity of cyclotides remains of great interest to fully exploit the endless potential of these peptides for a variety of pharmaceutical as well as agrochemical applications.

Acknowledgements

We would like to acknowledge funding by a UQ Postdoctoral Research Fellowship awarded to JK and an ARC Future Fellowship (FT100100476) to RJC.

References

1. Newman, D. J.; Cragg, G. M. J Nat Prod 2012, 75, 311-335.

2. Gruber, C. W.; Muttenthaler, M.; Freissmuth, M. Curr Pharm Des 2010, 16, 3071-3088.

3. Brogden, K. A.; Ackermann, M.; McCray, P. B., Jr.; Tack, B. F. Int J Antimicrob Agents 2003, 22, 465-478.

4. Craik, D. J.; Fairlie, D. P.; Liras, S.; Price, D. Chem Biol Drug Des 2013, 81, 136-147.

5. Otvos, L.; Wade, J. D. Frontiers in Chemistry 2014, 2, 62.

6. Cascales, L.; Craik, D. J. Org Biomol Chem 2010, 8, 5035-5047.

7. Adamska, A.; Janecka, A. Curr Med Chem 2015, 22, 352-359.

8. Arnison, P. G.; Bibb, M. J.; Bierbaum, G.; Bowers, A. A.; Bugni, T. S.; Bulaj, G.; Camarero, J. A.; Campopiano, D. J.; Challis, G. L.; Clardy, J.; Cotter, P. D.; Craik, D. J.; Dawson, M.; Dittmann, E.; Donadio, S.; Dorrestein, P. C.; Entian, K. D.; Fischbach, M. A.; Garavelli, J. S.; Göransson, U.; Gruber, C. W.; Haft, D. H.; Hemscheidt, T. K.; Hertweck, C.; Hill, C.; Horswill, A. R.; Jaspars, M.; Kelly, W. L.; Klinman, J. P.; Kuipers, O. P.; Link, A. J.; Liu, W.; Marahiel, M. A.; Mitchell, D. A.; Moll, G. N.; Moore, B. S.; Muller, R.; Nair, S. K.; Nes, I. F.; Norris, G. E.; Olivera, B. M.; Onaka, H.; Patchett, M. L.; Piel, J.; Reaney, M. J.; Rebuffat, S.; Ross, R. P.; Sahl, H. G.; Schmidt, E. W.; Selsted, M. E.; Severinov, K.; Shen, B.; Sivonen, K.; Smith, L.; Stein, T.; Sussmuth, R. D.; Tagg, J. R.; Tang, G. L.; Truman, A. W.; Vederas, J. C.; Walsh, C. T.; Walton, J. D.; Wenzel, S. C.; Willey, J. M.; van der Donk, W. A. Nat Prod Rep 2013, 30, 108-160.

9. Craik, D. J.; Daly, N. L.; Bond, T.; Waine, C. J Mol Biol 1999, 294, 1327-1336.

10. Clark, R. J.; Daly, N. L.; Craik, D. J. Biochem J 2006, 394, 85-93.

11. Mulvenna, J. P.; Wang, C.; Craik, D. J. Nucleic Acids Res 2006, 34, D192-194.

12. Kaas, Q.; Craik, D. J. Biopolymers 2010, 94, 584-591.

13. Gran, L. Lloydia 1973, 36, 174-178.

14. Pinto, M. F.; Fensterseifer, I. C.; Migliolo, L.; Sousa, D. A.; de Capdville, G.; Arboleda-Valencia, J. W.; Colgrave, M. L.; Craik, D. J.; Magalhaes, B. S.; Dias, S. C.; Franco, O. L. J Biol Chem 2012, 287, 134-147.

15. Gerlach, S. L.; Rathinakumar, R.; Chakravarty, G.; Göransson, U.; Wimley, W. C.; Darwin, S. P.; Mondal, D. Biopolymers 2010, 94, 617-625.

16. Gründemann, C.; Koehbach, J.; Huber, R.; Gruber, C. W. J Nat Prod 2012, 75, 167-174.

17. Gruber, C. W.; Muttenthaler, M. PLoS ONE 2012, 7, e32559.

18. Koehbach, J.; Attah, A. F.; Berger, A.; Hellinger, R.; Kutchan, T. M.; Carpenter, E. J.; Rolf, M.; Sonibare, M. A.; Moody, J. O.; Wong, G. K.; Dessein, S.; Greger, H.; Gruber, C. W. Biopolymers (Pept Sci) 2013, 100, 438-452.

19. Jin, A. H.; Dutertre, S.; Kaas, Q.; Lavergne, V.; Kubala, P.; Lewis, R. J.; Alewood, P. F. Mol Cell Proteomics 2013, 12, 3824-3833.

20. Velasquez, J. E.; van der Donk, W. A. Curr Opin Chem Biol 2011, 15, 11-21.

21. Gran, L. Acta Pharmacol Toxicol (Copenh) 1973, 33, 400-408.

22. Sletten, K.; Gran, L. Meddelelser fra Norsk Farmaceutisk Selskap 1970, 35, 69-82.

23. Saether, O.; Craik, D. J.; Campbell, I. D.; Sletten, K.; Juul, J.; Norman, D. G. Biochemistry 1995, 34, 4147-4158.

24. Witherup, K. M.; Bogusky, M. J.; Anderson, P. S.; Ramjit, H.; Ransom, R. W.; Wood, T.; Sardana, M. J Nat Prod 1994, 57, 1619-1625.

25. Schöpke, T.; Agha, M. I. H.; Kraft, R.; Otto, A.; Hiller, K. Scientia Pharmaceutica 1993, 61, 145-153.

26. Gustafson, K. R.; Sowder, R. C.; Henderson, L. E.; Parsons, I. C.; Kashman, Y.; Cardellina, J. H.; McMahon, J. B.; Buckheit, R. W.; Pannell, L. K.; Boyd, M. R. J Am Chem Soc 1994, 116, 9337-9338.

27. Göransson, U.; Broussalis, A. M.; Claeson, P. Anal Biochem 2003, 318, 107-117.

28. Sze, S. K.; Wang, W.; Meng, W.; Yuan, R.; Guo, T.; Zhu, Y.; Tam, J. P. Anal Chem 2009, 81, 1079-1088.

29. Plan, M. R.; Göransson, U.; Clark, R. J.; Daly, N. L.; Colgrave, M. L.; Craik, D. J. Chembiochem 2007, 8, 1001-1011.

30. Ireland, D. C.; Colgrave, M. L.; Craik, D. J. Biochem J 2006, 400, 1-12.

31. Broussalis, A. M.; Göransson, U.; Coussio, J. D.; Ferraro, G.; Martino, V.; Claeson, P. Phytochemistry 2001, 58, 47-51.

32. Gerlach, S. L.; Burman, R.; Bohlin, L.; Mondal, D.; Göransson, U. J Nat Prod 2010, 73, 1207-1213.

33. Wang, C. K.; Colgrave, M. L.; Gustafson, K. R.; Ireland, D. C.; Göransson, U.; Craik, D. J. J Nat Prod 2008, 71, 47-52.

34. Hashempour, H.; Koehbach, J.; Daly, N. L.; Ghassempour, A.; Gruber, C. W. Amino Acids 2013, 44, 581-595.

35. Poth, A. G.; Colgrave, M. L.; Lyons, R. E.; Daly, N. L.; Craik, D. J. Proc Natl Acad Sci U S A 2011, 108, 10127-10132.

36. Nguyen, G. K.; Zhang, S.; Nguyen, N. T.; Nguyen, P. Q.; Chiu, M. S.; Hardjojo, A.; Tam, J. P. J Biol Chem 2011, 286, 24275-24287.

37. Poth, A. G.; Colgrave, M. L.; Philip, R.; Kerenga, B.; Daly, N. L.; Anderson, M. A.; Craik, D. J. ACS Chem Biol 2011, 6, 345-355.

38. Poth, A. G.; Mylne, J. S.; Grassl, J.; Lyons, R. E.; Millar, A. H.; Colgrave, M. L.; Craik, D. J. J Biol Chem 2012, 287, 27033-27046.

39. He, W.; Chan, L. Y.; Zeng, G.; Daly, N. L.; Craik, D. J.; Tan, N. Peptides 2011, 32, 1719-1723.

40. Colgrave, M. L.; Jones, A.; Craik, D. J. J Chromatogr A 2005, 1091, 187-193.

41. Nguyen, G. K.; Zhang, S.; Wang, W.; Wong, C. T.; Nguyen, N. T.; Tam, J. P. J Biol Chem 2011, 286, 44833-44844.

42. Hashempour, H.; Ghassempour, A.; Daly, N. L.; Spengler, B.; Rompp, A. Protein Pept Lett 2011, 18, 747-752.

43. Göransson, U.; Broussalis, A. M.; Claeson, P. Anal Biochem 2003, 318, 107-117.

44. Trabi, M.; Svangard, E.; Herrmann, A.; Göransson, U.; Claeson, P.; Craik, D. J.; Bohlin, L. J Nat Prod 2004, 67, 806-810.

45. Gruber, C. W.; Elliott, A. G.; Ireland, D. C.; Delprete, P. G.; Dessein, S.; Göransson, U.; Trabi, M.; Wang, C. K.; Kinghorn, A. B.; Robbrecht, E.; Craik, D. J. Plant Cell 2008, 20, 2471-2483.

46. Stec, B. Cell Mol Life Sci 2006, 63, 1370-1385.

47. Chiche, L.; Heitz, A.; Gelly, J. C.; Gracy, J.; Chau, P. T.; Ha, P. T.; Hernandez, J. F.; Le-Nguyen, D. Curr Protein Pept Sci 2004, 5, 341-349.

48. Nguyen, G. K.; Lim, W. H.; Nguyen, P. Q.; Tam, J. P. J Biol Chem 2012, 287, 17598-17607.

49. Colgrave, M. L.; Poth, A. G.; Kaas, Q.; Craik, D. J. Biopolymers 2010, 94, 592-601.

50. Hellinger, R.; Koehbach, J.; Soltis, D. E.; Carpenter, E. J.; Wong, G. K.; Gruber, C. W. J Proteome Res 2015, 14, 4851-4862.

51. Burman, R.; Yeshak, M. Y.; Larsson, S.; Craik, D. J.; Rosengren, K. J.; Göransson, U. Front Plant Sci 2015, 6, 855.

52. Simonsen, S. M.; Sando, L.; Ireland, D. C.; Colgrave, M. L.; Bharathi, R.; Göransson, U.; Craik, D. J. Plant Cell 2005, 17, 3176-3189.

53. Jennings, C.; West, J.; Waine, C.; Craik, D.; Anderson, M. Proc Natl Acad Sci U S A 2001, 98, 10614-10619.

54. Burman, R.; Gruber, C. W.; Rizzardi, K.; Herrmann, A.; Craik, D. J.; Gupta, M. P.; Göransson, U. Phytochemistry 2010, 71, 13-20.

55. Zhang, J.; Liao, B.; Craik, D. J.; Li, J. T.; Hu, M.; Shu, W. S. Gene 2009, 431, 23-32.

56. Herrmann, A.; Burman, R.; Mylne, J. S.; Karlsson, G.; Gullbo, J.; Craik, D. J.; Clark, R. J.; Göransson, U. Phytochemistry 2008, 69, 939-952.

57. Nguyen, G. K.; Lian, Y.; Pang, E. W.; Nguyen, P. Q.; Tran, T. D.; Tam, J. P. J Biol Chem 2013, 288, 3370-3380.

58. Góngora-Castillo, E.; Buell, C. R. Nat Prod Rep 2013, 30, 490-500.

59. Cahais, V.; Gayral, P.; Tsagkogeorga, G.; Melo-Ferreira, J.; Ballenghien, M.; Weinert, L.; Chiari, Y.; Belkhir, K.; Ranwez, V.; Galtier, N. Mol Ecol Resour 2012, 12, 834-845.

60. Koehbach, J.; Jackson, K. A. V. Peptidomics 2015, 2, 17-25.

61. Mulvenna, J. P.; Mylne, J. S.; Bharathi, R.; Burton, R. A.; Shirley, N. J.; Fincher, G. B.; Anderson, M. A.; Craik, D. J. Plant Cell 2006, 18, 2134-2144.

62. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. J Mol Biol 1990, 215, 403-410.

63. Nguyen, G. K.; Wang, S.; Qiu, Y.; Hemu, X.; Lian, Y.; Tam, J. P. Nat Chem Biol 2014, 10, 732-738.

64. Harris, K. S.; Durek, T.; Kaas, Q.; Poth, A. G.; Gilding, E. K.; Conlan, B. F.; Saska, I.; Daly, N. L.; van der Weerden, N. L.; Craik, D. J.; Anderson, M. A. Nat Commun 2015, 6, 10199.

65. Conlan, B. F.; Colgrave, M. L.; Gillon, A. D.; Guarino, R.; Craik, D. J.; Anderson, M. A. J Biol Chem 2012, 287, 28037-28046.

66. Qin, Q.; McCallum, E. J.; Kaas, Q.; Suda, J.; Saska, I.; Craik, D. J.; Mylne, J. S. BMC Genomics 2010, 11, 111.

67. Zhang, J.; Li, J.; Huang, Z.; Yang, B.; Zhang, X.; Li, D.; Craik, D. J.; Baker, A. J.; Shu, W.; Liao, B. J Plant Physiol 2015, 178, 17-26.

68. Zhang, J.; Hua, Z.; Huang, Z.; Chen, Q.; Long, Q.; Craik, D. J.; Baker, A. J.; Shu, W.; Liao, B. Planta 2015, 241, 929-940.

69. Porto, W. F.; Miranda, V. J.; Pinto, M. F.; Dohms, S. M.; Franco, O. L. Biopolymers 2016, 106, 109-118.

70. Kedarisetti, P.; Mizianty, M. J.; Kaas, Q.; Craik, D. J.; Kurgan, L. Biochim Biophys Acta 2014, 1844, 181-190.

71. Mohimani, H.; Liu, W. T.; Mylne, J. S.; Poth, A. G.; Colgrave, M. L.; Tran, D.; Selsted, M. E.; Dorrestein, P. C.; Pevzner, P. A. J Proteome Res 2011, 10, 4505-4512.

72. Mohimani, H.; Kersten, R. D.; Liu, W. T.; Wang, M.; Purvine, S. O.; Wu, S.; Brewer, H. M.; Pasa-Tolic, L.; Bandeira, N.; Moore, B. S.; Pevzner, P. A.; Dorrestein, P. C. ACS Chem Biol 2014, 9, 1545-1551.

73. Slazak, B.; Jacobsson, E.; Kuta, E.; Göransson, U. Phytochemistry 2015, 117, 527-536.

74. Seydel, P.; Gruber, C. W.; Craik, D. J.; Dornenburg, H. Appl Microbiol Biotechnol 2007, 77, 275-284.

75. Trabi, M.; Craik, D. J. Plant Cell 2004, 16, 2204-2216.

76. Mahatmanto, T.; Poth, A. G.; Mylne, J. S.; Craik, D. J. Fitoterapia 2014, 95, 22-33.

77. Poth, A. G.; Chan, L. Y.; Craik, D. J. Biopolymers (Pept Sci) 2013, 100, 480-491.

78. Park, S.; Stromstedt, A. A.; Göransson, U. PLoS ONE 2014, 9, e91430.

79. Jennings, C. V.; Rosengren, K. J.; Daly, N. L.; Plan, M.; Stevens, J.; Scanlon, M. J.; Waine, C.; Norman, D. G.; Anderson, M. A.; Craik, D. J. Biochemistry 2005, 44, 851-860.

80. Gruber, C. W.; Cemazar, M.; Anderson, M. A.; Craik, D. J. Toxicon 2007, 49, 561-575.

81. Wang, C. K.; Wacklin, H. P.; Craik, D. J. J Biol Chem 2012, 287, 43884-43898.

82. Henriques, S. T.; Huang, Y. H.; Castanho, M. A.; Bagatolli, L. A.; Sonza, S.; Tachedjian, G.; Daly, N. L.; Craik, D. J. J Biol Chem 2012, 287, 33629-33643.

83. Koehbach, J.; O'Brien, M.; Muttenthaler, M.; Miazzo, M.; Akcan, M.; Elliott, A. G.; Daly, N. L.; Harvey, P. J.; Arrowsmith, S.; Gunasekera, S.; Smith, T. J.; Wray, S.; Göransson, U.; Dawson, P. E.; Craik, D. J.; Freissmuth, M.; Gruber, C. W. Proc Natl Acad Sci U S A 2013, 110, 21183-21188.

Table 1. Characteristics and obtained information of *in silico* and MS-based discovery approaches.

	In silico mining (genome/transcriptome data)	Mass-spectrometry analysis (MS and MS^n)
Peptide expression	No	Yes
Correct amino acids	Yes	No
Mature peptide length/mass	No	Yes
Cyclic or linear peptide	No[a]	Yes
Posttranslational modifications	No	Yes
Biosynthetic origin	Yes	No
Sample input	n.a.[b] - Low	Low-High[c]
Identification capacity (peptides/study)	High	Low-Medium
Costs	n.a. - High[d]	Medium-High[e]
Special equipment needed	No	Yes

Footnotes: [a]recent bioinformatics analysis suggests the prediction of cyclic peptides from nucleotide data,[70] [b]n.a. not applicable if performed using publicly available datasets only, [c]identification is possible from single leaf samples however thorough analysis and full sequence characterization often requires substantial amounts of plant material, [d]sequencing of large sample numbers result in reasonably high project costa, also access to high performing computing facilities may cause additional costs, [e]costs can vary and depend on sample number and access to MS facilities

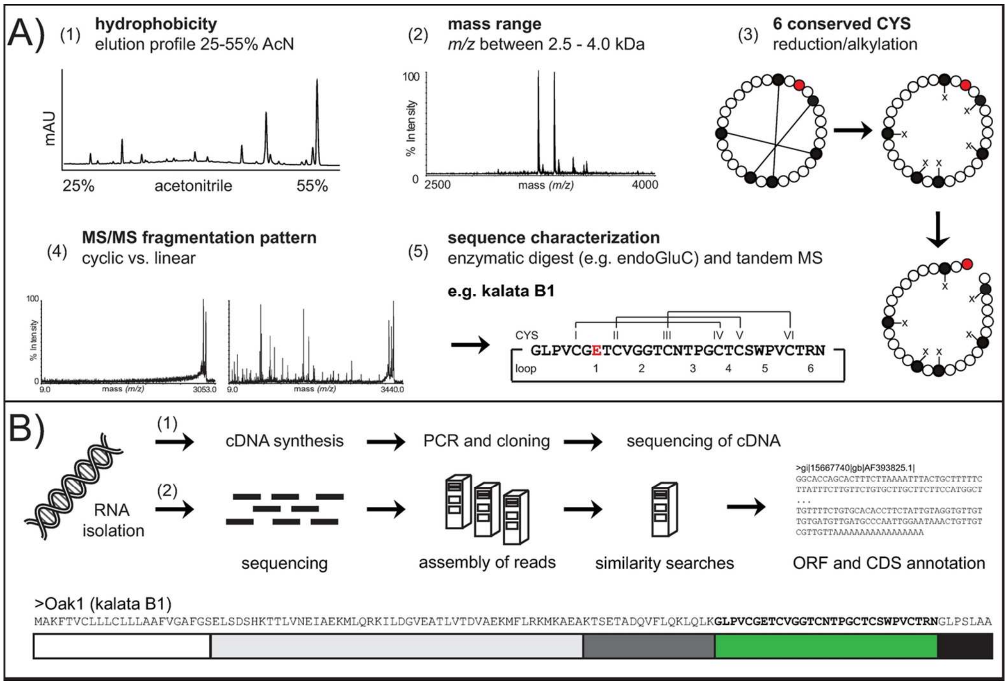
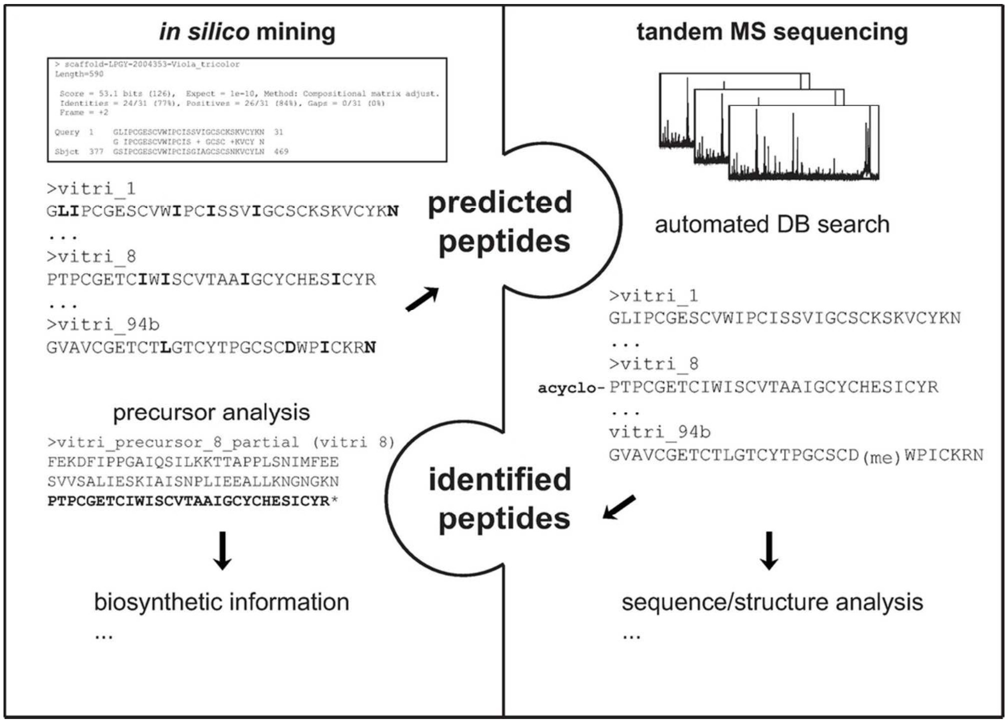

A)
(1) hydrophobicity
elution profile 25-55% AcN
mAU
25%
acetonitrile
55%
(2) mass range
m/z between 2.5 - 4.0 kDa
100
% Intensity
0
2500
mass (m/z)
4000
(3) 6 conserved CYS
reduction/alkylation
(4) MS/MS fragmentation pattern
cyclic vs. linear
100
% Intensity
0
9.0
mass (m/z)
3053.0
9.0
mass (m/z)
3440.0
(5) sequence characterization
enzymatic digest (e.g. endoGluC) and tandem MS
e.g. kalata B1
CYS I II III IV V VI
GLPVCGETCVGGTCNTPGCTCSWPVCTRN
loop 1 2 3 4 5 6
B)
(1)
cDNA synthesis
PCR and cloning
sequencing of cDNA
RNA
isolation
(2)
sequencing
assembly of reads
similarity searches
ORF and CDS annotation
>gi|15667740|gb|AF393825.1|
GGCACCAGCACTTTCTTAAAATTTACTGCTTTTTC
TTATTTCTTGTTCTGTGCTTGCTTCTTCCATGGCT
...
TGTTTTCTGTGCACACCTTCTATTGTAGGTGTTGT
TGTGATGTTGATGCCCAATTGGAATAAACTGTTGT
CGTTGTTAAAAAAAAAAAAAAAAA
>Oak1 (kalata B1)
MAKFTVCLLLCLLLAAFVGAFGSELSDSHKTTLVNEIAEKMLQRKILDGVEATLVTDVAEKMFLRKMKAEAKTSETADQVFLQKLQLKGLPVCGETCVGGTCNTPGCTCSWPVCTRNGLPSLAA

Figure 1. Cyclotide discovery. The key steps of the two different discovery pipelines as discussed in the manuscript. (A) Classical peptidomics analysis makes use of the typical physico-chemical properties of cyclotides such as hydrophobicity (1), mass (2) and cysteine content (3). Conserved cysteine residues (CYS) are shown in black and the glutamic acid is highlighted in red. The difference in MS/MS fragmentation pattern (4) as well as the sequence of the prototypic cyclotide kalata B1 including disulfide connectivities and cyclic backbone is shown (5). The conserved glutamic acid (red) is highlighted. Cysteine residues (CYS) are numbered with roman numerals, and intercysteine loop numbers are indicated below the sequence. (B) Nucleotide discovery from gene/transcript to precursor sequence. (1) cDNA based approaches include PCR and cloning and sequencing of the cDNA clones. (2) Transcriptome analysis requires sequencing and assembly prior to in silico mining. The prototypic precursor sequence of Oak1 (encoding for kalata B1) is shown with its typical elements that include an ER signal sequence (white), a N-pro region (light grey) a N-terminal repeat domain (dark grey), the mature cyclotide (green) and a C-terminal tail sequence (black). ORF open reading frame, CDS coding DNA sequence.
127x86mm (300 x 300 DPI)



```
> scaffold-LPGY-2004353-Viola_tricolor
Length=590

 Score = 53.1 bits (126),  Expect = 1e-10, Method: Compositional matrix adjust.
 Identities = 24/31 (77%), Positives = 26/31 (84%), Gaps = 0/31 (0%)
 Frame = +2

Query  1    GLIPCGESCVWIPCISSVIGCSCKSKVCYKN  31
            G IPCGESCVWIPCIS + GCSC +KVCY N
Sbjct  377  GSIPCGESCVWIPCISGIAGCSCSNKVCYLN  469
```

```
>vitri_1
GLIPCGESCVWIPCISSVIGCSCKSKVCYKN
...
>vitri_8
PTPCGETCIWISCVTAAIGCYCHESICYR
...
>vitri_94b
GVAVCGETCTLGTCYTPGCSCDWPICKRN
```

```
>vitri_precursor_8_partial (vitri 8)
FEKDFIPPGAIQSILKKTTAPPLSNIMFEE
SVVSALIESKIAISNPLIEEALLKNGNGKN
PTPCGETCIWISCVTAAIGCYCHESICYR*
```

```
>vitri_1
GLIPCGESCVWIPCISSVIGCSCKSKVCYKN
...
>vitri_8
acyclo-PTPCGETCIWISCVTAAIGCYCHESICYR
...
vitri_94b
GVAVCGETCTLGTCYTPGCSCD(me)WPICKRN
```

Figure 2. Combined -omics workflow. Streamlined discovery of novel cyclotides makes use of the in silico identification of predicted peptide sequences (left) that are used for the generation of a database (DB) against which tandem MS sequence data (right) can be searched. Identified peptides can be retrospectively matched to their precursor sequences allowing insights into their biosynthetic origin.
93x66mm (300 x 300 DPI)