Invited Review Open Access

Johannes Koehbach*, Kathryn A.V. Jackson

Unravelling peptidomes by *in silico* mining

Abstract: Peptides of great number and diversity occur in all domains of life and exhibit a range of pharmaceutically relevant bioactivities. The complexity of biological samples including human cells or tissues, plant extracts or animal venom cocktails, often impedes the discovery of novel bioactive peptides using mass spectrometry-based peptidomics analysis. An increasing number of publicly available genome and transcriptome datasets, together with refined bioinformatics analysis, allows for rapid identification of novel peptides which may have been previously unrecognized. Moreover, a combination of information extracted from *in silico* mining approaches together with data derived from mass spectrometry-based studies provides new impetus for future peptidome analyses, including the discovery of novel bioactive peptides that can serve as starting points for drug development.

Keywords: genome, transcriptome, peptidomics, mass-spectrometry, short ORF, peptides

DOI 10.1515/ped-2015-0002
Received December 1, 2014; accepted February 11, 2015

Abbreviations:

MS – mass-spectrometry,
sORF – short open reading frame

1 Naturally occurring bioactive peptides – a pharmacological treasure trove

The variety of natural compounds remains one of the biggest resources for the discovery of novel drug leads [1]. The majority of these substances described to date are low molecular weight compounds generated from secondary metabolism. These compounds typically have preferred druglike properties, in particular a higher stability and oral bioavailability, as compared to larger molecules [2]. However, over the last few decades several ribosomally-synthesized peptides spanning a molecular weight range of 0.5 to 5 kDa have gained particular interest in drug discovery and development approaches, filling a gap between small molecules and large biologicals [3]. Bioactive peptides are a versatile class of biomacromolecules that occur in great number and diversity across organisms, ranging from microbes and plants to invertebrate and vertebrate species, including humans. They have a unique range of functions and are important physiological regulators that can act as hormones by modulating cellular signaling pathways [4], serve as messenger molecules in inter-species communication [5] or display an important part of the defense mechanisms against pathogens and predators [6, 7]. Within humans several peptides have been shown to be useful as disease biomarkers [8, 9].

In the quest for peptide identification, *de novo* characterization and quantification, classical peptidomics approaches are facing serious challenges. The extensive characterization of peptides present in biological samples by means of standard analytical techniques such as liquid chromatography and mass spectrometry (MS) experiments, can be both laborious and inefficient. This is primarily due to the high complexity of biological samples. In addition, samples are often limited in quantity, with bioactive compounds present only in trace amounts. On top of this, indistinct resolution of chromatographic methods [10] further impedes peptide identification. Besides continuous advances in MS-based analytics, the availability of a constantly growing number of publicly accessible genome and transcriptome

***Corresponding author: Johannes Koehbach:** School of Biomedical Sciences, The University of Queensland, 4072 St. Lucia QLD, Australia, E-mail: j.koehbach@uq.edu.au
Kathryn A.V. Jackson: School of Biomedical Sciences, The University of Queensland, 4072 St. Lucia QLD, Australia

© 2015 Johannes Koehbach, Kathryn A.V. Jackson licensee De Gruyter Open.
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License.

datasets, as well as increasingly refined bioinformatics techniques, facilitates a new avenue of peptide discovery that has recently attracted much interest. *In silico* mining has been established not only as an alternative but as a complementary method for novel and potentially bioactive peptide identification within a variety of species [11-14].

Here we review recent achievements for the discovery of bioactive peptides. Particular emphasis will be given to novel bioinformatic approaches that allow the identification of peptides that have previously been overlooked in genome studies. Moreover, the potential of combining data from *in silico* mining and MS-based discovery approaches will be discussed.

2 *In silico* mining revisited – unveiling hidden treasures of genome and transcriptome datasets

Classical peptidomics studies using liquid-chromatography coupled to MS techniques have been successfully used to identify numerous bioactive peptides at the protein level. Nowadays the characterization of peptidomes, rather than one-molecule-per-study approaches has become state-of-the-art [15]. Despite technical advancements, MS-based discovery approaches still have major limitations, in the forms of highly complex mixtures, limited amounts of samples and the low abundance of many peptides. This restricts the rate at which new bioactive molecules can be identified due to the labor-intensive nature of analytical separation from crude natural products [10]. Additionally bioactive compounds might be encoded in pseudo-genes and hence are missed altogether in peptidome analyses [16].

As genome and transcriptome datasets have become more readily available due to advances in next-generation sequencing technology, mining of these data has initiated a new era for the discovery of novel bioactive peptides. Identifying translatable genetic sequences using bioinformatics approaches has become increasingly intertwined with modern natural product discovery [17]. This has been facilitated by progressively more multifarious, yet user-friendly computer programs that are also amenable to non-bioinformaticians [18, 19]. Transcriptomes in particular, since they are typically less expensive and elaborate to generate than genomes, have proven to be a treasure trove of previously unrecognized bioactive peptides. If correctly harnessed through automated bioinformatics approaches, the use of sequence databases promises to significantly expand our current knowledge of natural product peptidomes and make high throughput peptide discovery more feasible [20-23].

To date, the identification of protein products through *in silico* methods has relied heavily on similarity searches for sequence homology to annotated genomes. This allows for the easy detection of sequences of interest and permits the establishment of efficient, repeatable workflows with which to successfully mine large amounts of genetic data and characterize newly identified gene products. A number of publicly accessible online programs are available to mine existing genome/transcriptome data and identify potential hits that have been previously unannotated. An overview of selected and commonly used online tools is given in Table 1. As most workflows are initially based on the identification of homologous sequences, the National Centre for Biotechnology Information’s (NCBI) Basic Local Alignment Search Tool (BLAST)[24] searches are common starting points. The various types of BLAST search, *e.g.* BLASTN or BLASTX, either with or without translation of nucleotides, are used to mine for unannotated putative peptide-encoding transcripts within databases. These databases include Whole Genome Shotgun Assembly (WGS), Transcript Shotgun Assembly (TSA) or Expressed Sequence Tag (EST) databases and are searched using queries with known transcript sequences or protein patterns. If subsequent translation of BLAST search outputs is required, this can be performed with nucleotide translation tools, *e.g.* ExPaSy Translate [25]. Sequence alignment tools, such as Clustal Omega [26], are then able to generate alignments of identified sequences to infer homology. These alignments should ultimately receive manual verification to ensure the quality and accuracy of the hit. The quality of the alignment output is dependent upon the specificity of the query sequence and this must necessarily come as a trade-off. While a non-specific search criteria or query sequence may result in a larger number of hits, these are likely to include a greater number of erroneous results. Greater specificity and constraint of search criteria will reduce the number of hits but with a higher rate of accurate and relevant results. Signal peptides, also frequently used as a search criterion [20], can be identified with the Centre for Biological Sequence Analysis’ SignalP prediction [27]. Further annotation of coding DNA sequences and open reading frames (ORFs) can be accomplished using algorithms such as GeneWise [28]. Successful examples of peptides identified using the basic workflows described above include a suite of neuropeptides in ticks [29], crustaceans [22], and oysters [30] as well as both defensins and neuropeptides in ants [11]. In addition to mining for conserved sequences, specific characteristics of peptides of interest can also be

Table 1. Overview of selected and commonly used publicly available *in silico* mining tools

Name	Description	Source[a]	Reference
Galaxy	Workflow management and bioinformatics toolkit	http://galaxyproject.org/	[18, 47, 48]
Ugene	Bioinformatics toolkit	http://ugene.unipro.ru/	[19]
BLAST	Similarity search	http://blast.ncbi.nlm.nih.gov/Blast.cgi	[24]
GeneWise	Pairwise sequence alignment tool	http://www.ebi.ac.uk/Tools/psa/genewise/	[28]
MEGA	Sequence alignment tool	http://www.megasoftware.net/	[71]
MUSCLE	Multiple sequence alignment tool	http://www.drive5.com/muscle/	[72]
Clustal Omega	Multiple sequence alignment tool	http://www.clustal.org/omega/	[26]
PROSITE	Pattern identification tool	http://prosite.expasy.org/prosite.html	[73]
ExPaSy Translate	Translation of nucleotide into protein sequence	http://web.expasy.org/translate/	[25]
SignalP	Prediction of signal peptide cleavage sites	http://www.cbs.dtu.dk/services/SignalP/	[27]
MEROPS	Peptidase database	http://merops.sanger.ac.uk/	[74]
Pep2Path	Identification of gene clusters for MS-analyzed peptides	http://pep2path.sourceforge.net/	[62]
sORF finder	Identification of sORFs with high coding potential	http://evolver.psc.riken.jp/	[45]
AUGUSTUS	Prediction of genes in eukaryotic genomic sequences	http://bioinf.uni-greifswald.de/augustus/	[75]
PhyloCSF	Distinguishes coding/non-coding RNA	https://github.com/mlin/PhyloCSF/wiki	[76]
DiANNA	Disulfide connectivity predictor	http://clavius.bc.edu/~clotelab/DiANNA/	[77]
NRPquest	Non-ribosomal product identification	http://cyclo.ucsd.edu:4568/nrpquest_full	[61]
GPS	Prediction of kinase-specific phosphorylation sites	http://gps.biocuckoo.org/	[78]
(UniProtKB/Swiss-Prot)/ (UniprotKB/TrEMBL)	Database of manually/ computationally annotated protein sequences	http://www.uniprot.org/	[79]

Footnotes: [a]Links accessed on the 01.12.2014

as used as search criteria, for instance structural motifs such as conserved cysteine frameworks [20, 22]. Using these patterns, disulfide-rich peptides such as conotoxins are easily identified with several hundred sequences identified within single studies [14, 20]. This technique was also successfully used for the identification of putative insulin-like peptides in platyhelminths, which was based on a search for conserved features including a cysteine framework and hydrophobic core, without a reliance on strong sequence homology [31].

While the mining of databases using both of these approaches has yielded a significant amount of annotational data, the continued reliance on homology- and conservation-based searches presents a potential limitation to the identification of truly novel, heterologous peptides that may have unexplored roles or be useful drug leads. Furthermore, it restricts the ability to understand the biosynthetic pathways between resulting peptide products and the gene products that encode them. This is evidenced by the limited ability of current *in silico* approaches to correctly match experimentally-identified peptides that have undergone extensive post-translational modifications with the genetic sequence that encodes them [32, 33].

An expansion of traditional computing workflows is occurring as new methods are explored with which to more fully harness the power of *in silico* data mining. This follows as sequences traditionally ascribed to non-coding regions of the genome have, under closer inspection, been revealed to have protein-encoding potential. While the majority of protein-coding genes are transcribed from long, conserved ORFs which are usually easily identifiable by automated gene annotations [34], attention is now being given to non-traditional reading frames that appear to have the potential to encode functional protein products [13, 35, 36]. These include open reading frames with non-AUG start codons, long intergenic non-coding RNAs, as well as short open reading frames (sORFs) [36, 37]. Programs that are able to explore possible open reading frames and identify these potentially peptide-coding regions could reduce the dependence on homology-based searches for peptide identification. In silico methods are now being adapted to take greater advantage of these opportunities for novel peptide discovery and facilitate large-scale identification of sequences with coding potential in previously overlooked areas of the genome.

Peptides identified from *in silico* data mining have conventionally been detected based on stringent search criteria to reduce false positive predictions, which usually includes a minimum sequence length of >100 amino acids [38]. Due to this typical cut-off length, open reading frames shorter than 100 amino acids have traditionally been disregarded. Recently, however, small open reading frames have been shown to be a potentially rich source of biologically relevant peptides that await annotation [39-41].

Peptides encoded by sORFs have already been found in numerous organisms including yeast [42], invertebrates [39, 43], human cell lines [44], as well as plants [41] suggesting that translation of sORFs seems to be conserved throughout evolution and may occur more often than previously anticipated. This is encouraging for the discovery of a novel pool of bioactive peptides that warrants further exploration. In order to increase the discovery of sORFs through *in silico* data mining, sORF finder [45], a free web-based tool to identify small open reading frames with high coding potential has been released and has been successfully used to detect sORFs in both plant and mammal genomes, representing an important step towards categorizing actively transcribed sORFs [40, 45]. Although sORFs are increasingly being identified in numerous genomes, because of their small size, it can be difficult to discern true, potentially coding reading frames from those that occur purely by chance due to the presence of spurious stop or start codons in a genetic sequence [40]. It therefore remains to be verified how many of these reading frames are actually translatable and whether they encode biologically active peptides. For example, of the 7,442 unannotated sORFs that were identified in *Arabidopsis thaliana* [46], only 155 of these were confirmed to be translated at the peptide level [33].

As the use of *in silico* analysis increases, new programs and algorithms are continuously being created to meet emerging needs of data mining, with a strong history of open-source software (Table 1). In addition to this, the creation of automated workflow programs such as Galaxy allow scientists without programming skills to analyse large amounts of data by assembling stepwise workflows [18, 47, 48]. Due to this continual improvement and increased accessibility of programs, the last decade has seen a massive influx of predicted protein sequences into databases. Many of these have not been demonstrated to have a functional role *in vivo*: relying on further computational analyses, such as structural and statistical methods, to provide evidence for a potentially functional role until one can be experimentally verified [49]. Furthermore, although *in silico* mining is a powerful tool for the discovery of novel gene-encoded and ribosomally-synthesized peptides, it cannot provide critical information regarding the actual presence at peptide level or post-translational modifications that might be required to transform a peptide into its biologically active form [50].

3 The power of two – linking MS data and genome/transcriptome mining

The identification of novel bioactive peptides based on either *in silico* or MS-based methods present their own unique advantages, challenges and limitations (Fig. 1). One of the major challenges and limiting factors for MS-based peptide identification and characterization is the high complexity of sample mixtures coupled with the low abundance of several bioactive compounds. Consequently, purification prior to analysis is a crucial yet perilous step, given the poor stability of many peptidic compounds. This drawback of peptides is also a challenge in their development as potential drug leads. Many potent peptides may simply not be detectable by chemical screens due to their rapid degradation [37]. Due to this, MS analysis may not reflect the true *in vivo* state of a biological sample. Furthermore, the truncated peptide fragments resulting from proteolysis may overpower the signals from remaining endogenous peptides of interest during MS measurement [51]. This holds true for mammalian samples such as blood serum or plasma [10] but applies equally to plant tissues [52] or animal venom cocktails [53].

Furthermore, the lack of publicly available nucleotide datasets for particular organisms may restrict several studies to the use of classical MS-based peptidomics approaches that are dependent on high-resolution MS equipment. The identification of peptides based on MS and tandem MS experiments may then cause additional problems. If no database is available against which MS^n data might be searched, time-consuming manual *de novo* sequencing has to be applied. Additionally, overlapping ion fragmentation patterns that are due to the presence of highly similar peptide sequences require further optimization of sequencing strategies [52], significantly slowing down analyses. Altogether this hampers efficient identification using currently available MS methods and drives the steady development of both the technical equipment as well as methodologies to increase the discovery rate of novel bioactive peptides.

To overcome limitations of MS-based peptidomics, the analysis of peptides at nucleotide level has recently gained increased attention. Such strategies not only reveal peptide sequences but allow the identification of precursor proteins. This may also provide novel insights into their biosynthetic processing and evolution. Additionally, this may also bear important information with regard to the identification of the biological function of novel compounds. To date, next-generation sequencing services, in particular transcriptome analyses, have become reasonably affordable and are likely to soon be part of routine experiments. If utilizing only publicly accessible databases, *in silico* mining is amenable at virtually no cost. However for many organisms, such as plants or invertebrates, reference genomes are not available. This makes *de novo* assembly a less than trivial task [54, 55]. Although peptide sequences derived from such approaches can then yield accurate information about incorporated amino acids, including isobaric residues, they cannot provide any information regarding the mature peptide length or any posttranslational modifications that may be crucial for its bioactivity [50].

This highlights that despite any drawbacks, MS-based peptidomics continues to be an indispensable tool for both the discovery and/or confirmation of peptides present in biological samples. Unannotated reading frames identified by *in silico* data mining must ultimately have their peptide products corroborated by *in vivo* methods, which are dominated by MS approaches [40]. Likewise, peptides that are first isolated by MS can lead to the identification of the encoding sequence. Furthermore, homologous sequences in other species can then be identified, increasing annotational data and further expanding the pool of potential drug leads. This highlights that the combined strength of these approaches hinges on the ability of each method to validate the findings of the other (Fig. 1).

It is therefore evident that a combination of both methods is very likely to accelerate the identification and characterization of novel peptide leads [33]. Indeed, recent studies that made use of a combined analysis using MS-based peptide sequencing together with bioinformatics data mining and molecular approaches

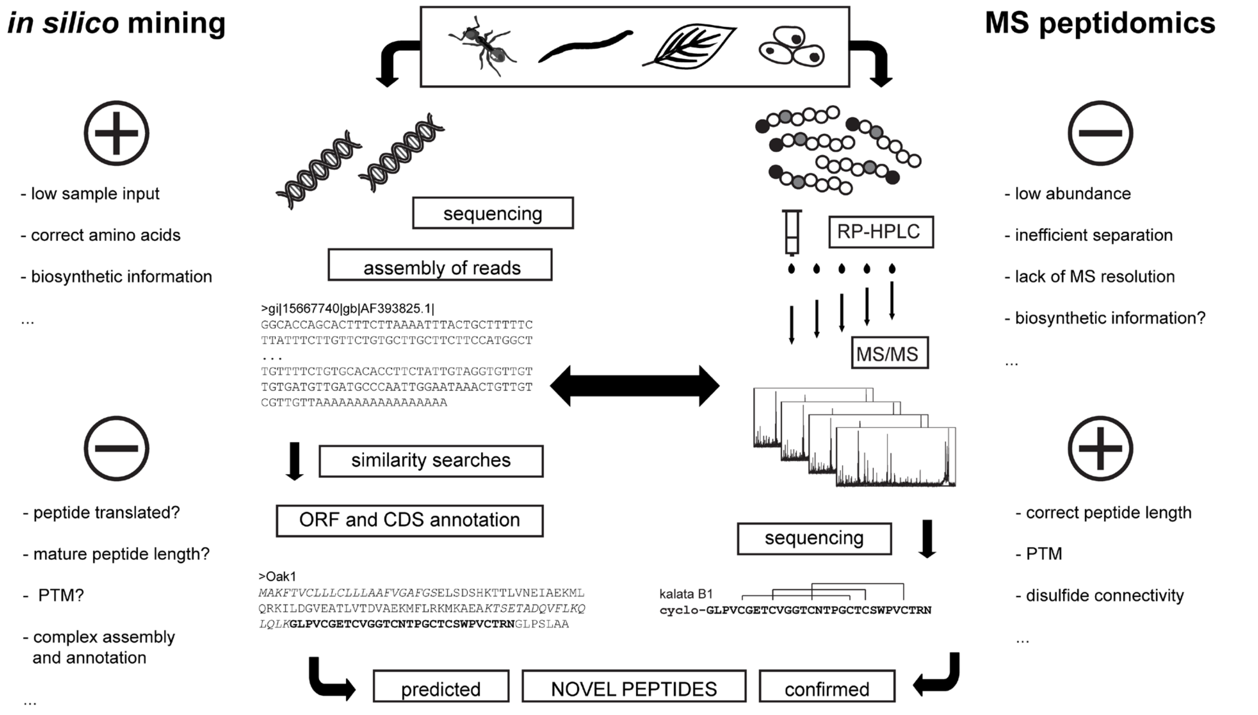
in silico mining
MS peptidomics
- low sample input
- correct amino acids
- biosynthetic information
...
- peptide translated?
- mature peptide length?
- PTM?
- complex assembly and annotation
...
sequencing
assembly of reads
>gi|15667740|gb|AF393825.1|
GGCACCAGCACTTTCTTAAAATTTACTGCTTTTTC
TTATTTCTTGTTCTGTGCTTGCTTCTTCCATGGCT
...
TGTTTTCTGTGCACACCTTCTATTGTAGGTGTTGT
TGTGATGTTGATGCCCAATTGGAATAAACTGTTGT
CGTTGTTAAAAAAAAAAAAAAAAA
similarity searches
ORF and CDS annotation
>Oak1
MAKFTVCLLLCLLLAAFVGAFGSELSDSHKTTLVNEIAEKML
QRKILDGVEATLVTDVAEKMFLRKMKAEAKTSETADQVFLKQ
LQLKGLPVCGETCVGGTCNTPGCTCSWPVCTRNGLPSLAA
RP-HPLC
MS/MS
sequencing
kalata B1
cyclo-GLPVCGETCVGGTCNTPGCTCSWPVCTRN
- low abundance
- inefficient separation
- lack of MS resolution
- biosynthetic information?
...
- correct peptide length
- PTM
- disulfide connectivity
...
predicted
NOVEL PEPTIDES
confirmed

Figure 1: Peptide discovery workflow overview. A comparative outline of *de novo* peptide discovery is shown. Samples from various natural sources can either be used for RNA/DNA extraction (left) or solvent extraction (right). Key steps for each approach are given in textboxes and are discussed within the manuscript. Major benefits and drawbacks are highlighted on the sides. Exemplarily, the coding DNA sequence (CDS) for the precursor containing the cyclotide kalata B1 (Genbank entry: AF393825.1), its translated open reading frame (ORF) sequence Oak1 (UniProt KB entry: P56254) as well as the mature peptide sequence of kalata B1 is shown. The mutual dependence and complementary validation of both discovery pipelines is indicated by a central and double-sided arrow. MS – mass spectrometry. PTM – post-translational modification.

are encouraging examples that a combined approach is a powerful tool for accelerated peptide discovery [12, 37, 41, 56-58]. Linking data from *in silico* studies to MS-based analyses has been shown to allow the connection of peptides to their genetic origin even in an automated fashion [59]. With growing numbers and increasing accuracy of publicly available genome datasets, these studies provide a useful basis that aid in MS-based *de novo* peptide discovery. Importantly, this approach is not only restricted to ribosomally-synthesized peptides [60] but can also be applied to non-ribosomally-synthesized peptides by looking for biosynthetic gene clusters [61, 62].

In the context of linking *in silico* data to MS-based peptidomics, the analysis of a novel class of short polypeptides has recently gained increased attention. Peptides encoded in short open reading frame (see previous section) denote a previously overlooked class of peptides that warrants further characterization. Their presence in biological samples has been shown to be of similar concentrations to other classes of proteins, which is encouraging for future studies regarding the characterization of functional roles of these peptides [37]. The identification of novel sORF encoded peptides can be of interest in itself, due to the discovery or reinforcement of the existence of non-canonical features such as non-AUG start codons [37]. This leaves the question open as to whether these peptides can ultimately be linked to biological functions and if they can then be utilized as novel disease biomarkers, molecular probes or drug leads.

4 Biological functions of novel peptides – challenges and opportunities

Given the advancements in *de novo* peptide identification it is obvious that a wealth of bioactive peptides is awaiting biological characterization and evaluation of any potential as novel drug leads. In this regard there is another significant difference between peptides identified from *in silico* mining versus MS-based approaches. The discovery of potential peptide-encoding sequences within genome or transcriptome datasets does not reveal if the encoded products are actually present as a peptide or if they exert any biological functionality. For example, retrocyclin has been identified as a potent antimicrobial peptide, which, while it is encoded in an expressed pseudogene in human bone marrow, is not actually transcribed [16]. Moreover, the conotoxin Vc1.1, which is a promising peptide to target neuropathic pain, has been found to be active only in its non-native and not post-translationally modified form [56, 63]. Although the identification of peptides via MS-based approaches does not directly provide information regarding bioactive functions, MS discovery often starts with the search for an active compound responsible for a given activity [64]. The identified peptide can then be retrospectively matched to the genetic sequence that encodes it. While this guarantees a functional end-product, the detection of an encoding sequence can be confounded by extensive post-translational modifications, low abundance or sample size, or being of non-ribosomal origin. As exploration of non-canonically-encoded peptides is still in its infancy, it remains to be demonstrated what portion of these newly-discovered sequences actually encode biologically functional peptides. A recent study examining sORFs in *Drosophila* exemplifies how identification of new ORFs does not equate to a similar number of functional peptides. Of nearly 600,000 sORFs that were identified via genome mining, annotational evidence for translation exists for only around 400 of these, and definitive proof of any functionality must still be demonstrated experimentally [65]. Furthermore, of those ORFs that are translated, it must be determined whether their peptide products are both abundant and stable enough to be detected by MS approaches. While MS is subject to detection bias, *e.g.* limited resolution of co-eluting compounds or insufficient ionisation, it is generally biased towards detecting the most abundantly expressed peptides. This may present an advantage in detecting functionality, in that the most highly expressed of these *in silico*-identified peptides are potentially the most likely to be functional [38]. However at the same time it runs the risk of passing over many less abundant but potentially still biologically significant peptides [37].

At a very general level, a peptide can be deemed functional if it is transcribed, translated and biologically relevant [65]. In the absence of direct *in vivo* evidence, as is often the case with newly identified sORFs, relevance can be indicated by genetic conservation across different species [34, 36, 65]. This is due to the consensus that cross-species conservation of an ORF is indicative of a valid protein-coding gene, while non-conserved genes are usually spurious [66]. Intriguing examples of this are neuropeptides such as oxytocin/vasopressin and related nonapeptides. Their existence for several hundred million years, along with their presence within various, evolutionarily distant species [67] indicates an increased probability that novel peptides such as these that are identified by *in silico* mining do indeed exert physiological functions [11, 68].

Moreover, some sORF-encoded peptides may not have a bioactive role in and of themselves: there are speculations that in some cases it may be the action of translation that plays a role [35], or that they may play a regulatory role that does not require active translation of the encoded peptide [36]. However the existence of several human sORF-encoded peptides has been confirmed experimentally by MS-based approaches [37]. Peptides identified using MS were matched to their ORFs using transcriptomic data analysis. Synthetic, isotopically-labelled peptides were then synthesized based on the implicated transcripts and mixed with endogenous peptides. Subsequently MS analysis was applied to verify that the predicted transcripts matched the observed peptide products. To circumvent some of the potential shortcomings associated with MS validation, incorporation of ribosomal profiling may be useful. It is, by its nature, more sensitive than MS analysis and is likely to detect a larger number of new ORFs [38]. The technique is based on deep sequencing of ribosome-protected mRNA fragments and permits a snapshot of active ribosomes and occurring translation events at a point in time [69]. However, ribosomal occupation does not always reflect active translation of the engaged mRNA and contained ORF [70]. It should therefore be used to augment, rather than replace, MS and other *in vivo* analyses of novel peptides.

5 Conclusion

The identification of bioactive peptides using peptidomics analyses has made significant contributions in biomedical research. The comprehensive characterization of peptides from a variety of biological samples has revealed many potent drug leads as well as disease biomarkers. Continuous development in MS technologies, refined molecular biology methods and nucleotide sequencing allows the identification of peptides in high numbers and with unique accuracy. Using complex bioinformatics, the analysis of peptidomes at genome and transcriptome level reveals a novel wealth of peptides of previously unexpected dimension. The combined analysis of peptides at nucleotide as well as peptide level is encouraging not only to facilitate the discovery of novel compounds, but in providing insights into their biosynthetic processing and evolution. Given the current advances in peptide chemistry towards overcoming the issues of instability and low oral bioavailability we remain confident that peptides identified by *in silico* mining can serve as ideal starting points for the development of peptide-based drugs.

Acknowledgements: We would like to thank Richard J. Clark for useful comments on the manuscript. We would like to acknowledge funding by the NHMRC (APP1034642), a UQ Postdoctoral Research Fellowship awarded to JK and an Australian Postgraduate Award scholarship awarded to KAVJ.

Conflict of interest: The authors declare no conflict of interest.

References

[1] Newman D.J., Cragg G.M., Natural products as sources of new drugs over the 30 years from 1981 to 2010, J. Nat. Prod., 2012, 75, 311-335.

[2] Lipinski C.A., Drug-like properties and the causes of poor solubility and poor permeability, J. Pharmacol. Toxicol. Methods, 2000, 44, 235-249.

[3] Craik D.J., Fairlie D.P., Liras S., Price D., The future of peptide-based drugs, Chem. Biol. Drug Des., 2013, 81, 136-147.

[4] Gruber C.W., Muttenthaler M., Freissmuth M., Ligand-based peptide design and combinatorial peptide libraries to target G protein-coupled receptors, Curr. Pharm. Des., 2010, 16, 3071-3088.

[5] Goodson J.L., Nonapeptides and the evolutionary patterning of sociality, Prog. Brain Res., 2008, 170, 3-15.

[6] Brogden K.A., Ackermann M., McCray P.B., Jr., Tack B.F., Antimicrobial peptides in animals and their role in host defences, Int. J. Antimicrob. Agents, 2003, 22, 465-478.

[7] Zasloff M., Antimicrobial peptides of multicellular organisms, Nature, 2002, 415, 389-395.

[8] Schrader M., Selle H., The process chain for peptidomic biomarker discovery, Dis. Markers, 2006, 22, 27-37.

[9] Martelli C., Iavarone F., Vincenzoni F., Cabras T., Manconi B., Desiderio C., Messana I., Castagnola M., Top-down peptidomics of bodily fluids, Peptidomics, 2013, 1, 47-64.

[10] Finoulst I., Pinkse M., Van Dongen W., Verhaert P., Sample preparation techniques for the untargeted LC-MS-based discovery of peptides in complex biological matrices, J. Biomed. Biotechnol., 2011, 2011, 245291.

[11] Gruber C.W., Muttenthaler M., Discovery of defense- and neuropeptides in social ants by genomemining, PLoS ONE, 2012, DOI: 10.1371/journal.pone.0032559.

[12] 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., et al., Cyclotide discovery in Gentianales revisited-identification and characterization of cyclic cystine-knot peptides and their phylogenetic distribution in Rubiaceae plants, Biopolymers, 2013, 100, 438-452.

[13] Frith M.C., Forrest A.R., Nourbakhsh E., Pang K.C., Kai C., Kawai J., Carninci P., Hayashizaki Y., Bailey T.L., Grimmond S.M., The Abundance of Short Proteins in the Mammalian Proteome, PLoS Genet., 2006, DOI: 10.1371/journal.pgen.0020052.

[14] Jin A.H., Dutertre S., Kaas Q., Lavergne V., Kubala P., Lewis R.J., Alewood P.F., Transcriptomic messiness in the venom duct of *Conus miles* contributes to conotoxin diversity, Mol. Cell. Proteomics, 2013, 12, 3824-3833.

[15] Schrader M., Schulz-Knappe P., Fricker L.D., Historical perspective of peptidomics, EuPA Open Proteom, 2014, 3, 171-182.

[16] Cole A.M., Hong T., Boo L.M., Nguyen T., Zhao C., Bristol G., Zack J.A., Waring A.J., Yang O.O., Lehrer R.I., Retrocyclin: a primate peptide that protects cells from infection by T- and M-tropic strains of HIV-1, Proc. Natl. Acad. Sci. U. S. A., 2002, *99*, 1813-1818.

[17] Bachmann B.O., Van Lanen S.G., Baltz R.H., Microbial genome mining for accelerated natural products discovery: is a renaissance in the making?, J. Ind. Microbiol. Biotechnol., 2014, 41, 175- 184.

[18] Goecks J., Nekrutenko A., Taylor J., Team T.G., Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol., 2010, http://genomebiology.com/2010/11/8/R86.

[19] Okonechnikov K., Golosova O., Fursov M., UGENE-team, Unipro UGENE: a unified bioinformatics toolkit, Bioinformatics, 2012, 28, 1166-1167.

[20] Lavergne V., Dutertre S., Jin A.H., Lewis R.J., Taft R.J., Alewood P.F., Systematic interrogation of the *Conus marmoreus* venom duct transcriptome with ConoSorter reveals 158 novel conotoxins and 13 new gene superfamilies, BMC Genomics, 2013, http://www.biomedcentral.com/1471- 2164/14/708.

[21] Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., et al., Phylogenomics resolves the timing and pattern of insect evolution, Science, 2014, 346, 763-767.

[22] Christie A.E., Prediction of the peptidomes of *Tigriopus californicus* and *Lepeophtheirus salmonis* (Copepoda, Crustacea), Gen. Comp. Endocrinol., 2014, 201, 87-106.

[23] Christie A.E., Expansion of the *Litopenaeus vannamei* and *Penaeus monodon* peptidomes using transcriptome shotgun assembly sequence data, Gen. Comp. Endocrinol., 2014, 206, 235-254.

[24] Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J., Basic local alignment search tool, J. Mol. Biol., 1990, 215, 403-410.

[25] Artimo P., Jonnalagedda M., Arnold K., Baratin D., Csardi G., de Castro E., Duvaud S., Flegel V., Fortier A., Gasteiger E., et al., ExPASy: SIB bioinformatics resource portal, Nucleic Acids Res., 2012, 40, W597-W603.

[26] Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., et al., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol., 2011, DOI: 10.1038/msb.2011.75.

[27] Petersen T.N., Brunak S., von Heijne G., Nielsen H., SignalP 4.0: discriminating signal peptides from transmembrane regions, Nat. Meth., 2011, 8, 785-786.

[28] Birney E., Clamp M., Durbin R., GeneWise and Genomewise, Genome Res., 2004, 14, 988-995.

[29] Christie A.E., Neuropeptide discovery in Ixodoidea: An *in silico* investigation using publicly accessible expressed sequence tags, Gen. Comp. Endocrinol., 2008, 157, 174-185.

[30] Stewart M.J., Favrel P., Rotgans B., Wang T., Zhao M., Sohail M., O'Connor W.A., Elizur A., Henry J., Cummins S.F., Neuropeptides encoded by the genomes of the Akoya pearl oyster *Pinctata fucata* and Pacific oyster *Crassostrea gigas*: a bioinformatic and peptidomic survey, BMC Genomics, 2014, http://www.biomedcentral.com/1471-2164/15/840.

[31] Wang S., Luo X., Zhang S., Yin C., Dou Y., Cai X., Identification of putative insulin-like peptides and components of insulin signaling pathways in parasitic platyhelminths by the use of genome-wide screening, FEBS J., 2014, 281, 877-893.

[32] Liu C., Li H., In Silico Prediction of Post-translational Modifications, In: Yu B & Hinchcliffe M. (Eds.), Methods in Molecular Biology, 1st ed., Humana Press, New York, 2011.

[33] Castellana N.E., Payne S.H., Shen Z., Stanke M., Bafna V., Briggs S.P., Discovery and revision of *Arabidopsis* genes by proteogenomics, Proc. Natl. Acad. Sci. U.S.A., 2008, 105, 21034-21038.

[34] Andrews S.J., Rothnagel J.A., Emerging evidence for functional peptides encoded by short open reading frames, Nat. Rev. Genet., 2014, 15, 193-204.

[35] Pauli A., Valen E., Schier A.F., Identifying (non-)coding RNAs and small peptides: Challenges and opportunities, BioEssays, 2014, DOI: 10.1002/bies.201400103.

[36] Bazzini A.A., Johnstone T.G., Christiano R., Mackowiak S.D., Obermayer B., Fleming E.S., Vejnar C.E., Lee M.T., Rajewsky N., Walther T.C., et al., Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation, EMBO J., 2014, 33, 981-993.

[37] Slavoff S.A., Mitchell A.J., Schwaid A.G., Cabili M.N., Ma J., Levin J.Z., Karger A.D., Budnik B.A., Rinn J.L., Saghatelian A., Peptidomic discovery of short open reading frame–encoded peptides in human cells, Nat. Chem. Biol., 2013, *9*, 59-64.

[38] Ma J., Ward C.C., Jungreis I., Slavoff S.A., Schwaid A.G., Neveu J., Budnik B.A., Kellis M., Saghatelian A., Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue, J. Proteome Res., 2014, 13, 1757-1765.

[39] Lu Y., Zhuang Y., Liu J., Mining antimicrobial peptides from small open reading frames in *Ciona intestinalis*, J. Pept. Sci., 2014, 20, 25-29.

[40] Crappe J., Van Criekinge W., Trooskens G., Hayakawa E., Luyten W., Baggerman G., Menschaert G., Combining *in silico* prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs, BMC Genomics, 2013, http://www.biomedcentral.com/1471-2164/14/648.

[41] Yang X., Tschaplinski T.J., Hurst G.B., Jawdy S., Abraham P.E., Lankford P.K., Adams R.M., Shah M.B., Hettich R.L., Lindquist E., et al., Discovery and annotation of small proteins using genomics, proteomics, and computational approaches, Genome Res., 2011, 21, 634-641.

[42] Kastenmayer J.P., Ni L., Chu A., Kitchen L.E., Au W.-C., Yang H., Carter C.D., Wheeler D., Davis R.W., Boeke J.D., et al., Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*, Genome Res., 2006, 16, 365-373.

[43] Galindo M.I., Pueyo J.I., Fouix S., Bishop S.A., Couso J.P., Peptides Encoded by Short ORFs Control Development and Define a New Eukaryotic Gene Family, PLoS Biol., 2007, DOI: 10.1371/journal.pbio.0050106.

[44] Oyama M., Kozuka-Hata H., Suzuki Y., Semba K., Yamamoto T., Sugano S., Diversity of Translation Start Sites May Define Increased Complexity of the Human Short ORFeome, Mol. Cell. Proteomics, 2007, 6, 1000-1006.

[45] Hanada K., Akiyama K., Sakurai T., Toyoda T., Shinozaki K., Shiu S.-H., sORF finder: a program package to identify small open reading frames with high coding potential, Bioinformatics, 2010, 26, 399-400.

[46] Hanada K., Zhang X., Borevitz J.O., Li W.-H., Shiu S.-H., A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are

transcribed and/or under purifying selection, Genome Res., 2007, 17, 632-640.

[47] Blankenberg D., Kuster G.V., Coraor N., Ananda G., Lazarus R., Mangan M., Nekrutenko A., Taylor J., Galaxy: A Web-Based Genome Analysis Tool for Experimentalists, Curr. Protoc. Mol. Biol, 2010, DOI: 10.1002/0471142727.mb1910s89.

[48] Giardine B., Riemer C., Hardison R.C., Burhans R., Elnitski L., Shah P., Zhang Y., Blankenberg D., Albert I., Taylor J., et al., Galaxy: A platform for interactive large-scale genome analysis, Genome Res., 2005, 15, 1451-1455.

[49] Pruess M., Apweiler R., Bioinformatics Resources for In Silico Proteome Analysis, J. Biomed. Biotechnol., 2003, 4, 231-236.

[50] Le T.T., Lehnert S., Colgrave M.L., Neuropeptidomics applied to studies of mammalian reproduction, Peptidomics, 2013, 1, 1-13.

[51] Romanova E.V., Dowd S.E., Sweedler J.V., Quantitation of endogenous peptides using mass spectrometry based methods, Curr. Opin. Chem. Biol., 2013, 17, 801-808.

[52] Hashempour H., Koehbach J., Daly N.L., Ghassempour A., Gruber C.W., Characterizing circular peptides in mixtures: sequence fragment assembly of cyclotides from a violet plant by MALDITOF/TOF mass spectrometry, Amino Acids, 2013, 44, 581-595.

[53] Ueberheide B.M., Fenyö D., Alewood P.F., Chait B.T., Rapid sensitive analysis of cysteine rich peptide venom components, Proc. Natl. Acad. Sci. U. S. A., 2009, 106, 6910-6915.

[54] Góngora-Castillo E., Buell C.R., Bioinformatics challenges in *de novo* transcriptome assembly using short read sequences in the absence of a reference genome sequence, Nat. Prod. Rep., 2013, 30, 490- 500.

[55] Cahais V., Gayral P., Tsagkogeorga G., Melo-Ferreira J., Ballenghien M., Weinert L., Chiari Y., Belkhir K., Ranwez V., Galtier N., Reference-free transcriptome assembly in non-model animals from next-generation sequencing data, Mol. Ecol. Resour., 2012, 12, 834-845.

[56] Jakubowski J.A., Keays D.A., Kelley W.P., Sandall D.W., Bingham J.P., Livett B.G., Gayler K.R., Sweedler J.V., Determining sequences and post-translational modifications of novel conotoxins in *Conus victoriae* using cDNA sequencing and mass spectrometry, J. Mass Spectrom., 2004, 39, 548- 557.

[57] Ma M., Gard A.L., Xiang F., Wang J., Davoodian N., Lenz P.H., Malecha S.R., Christie A.E., Li L., Combining *in silico* transcriptome mining and biological mass spectrometry for neuropeptide discovery in the Pacific white shrimp *Litopenaeus vannamei*, Peptides, 2010, 31, 27-43.

[58] Safavi-Hemami H., Hu H., Gorasia D.G., Bandyopadhyay P.K., Veith P.D., Young N.D., Reynolds E.C., Yandell M., Olivera B.M., Purcell A.W., Combined proteomic and transcriptomic interrogation of the venom gland of *Conus geographus* uncovers novel components and functional compartmentalization, Mol. Cell. Proteomics, 2014, 13, 938-953.

[59] Kersten R.D., Yang Y.L., Xu Y., Cimermancic P., Nam S.J., Fenical W., Fischbach M.A., Moore B.S., Dorrestein P.C., A mass spectrometry-guided genome mining approach for natural product peptidogenomics, Nat. Chem. Biol., 2011, 7, 794-802.

[60] 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., et al., Automated genome mining of ribosomal peptide natural products, ACS Chem. Biol., 2014, 9, 1545-1551.

[61] Mohimani H., Liu W.T., Kersten R.D., Moore B.S., Dorrestein P.C., Pevzner P.A., NRPquest: Coupling Mass Spectrometry and Genome Mining for Nonribosomal Peptide Discovery, J. Nat. Prod., 2014, 77, 1902-1909.

[62] Medema M.H., Paalvast Y., Nguyen D.D., Melnik A., Dorrestein P.C., Takano E., Breitling R., Pep2Path: Automated Mass Spectrometry-Guided Genome Mining of Peptidic Natural Products, PLoS Comput. Biol., 2014, DOI: 10.1371/journal.pcbi.1003822.

[63] Clark R.J., Fischer H., Nevin S.T., Adams D.J., Craik D.J., The synthesis, structural characterization, and receptor specificity of the alpha-conotoxin Vc1.1, J. Biol. Chem., 2006, 281, 23254-23263.

[64] Koehbach J., O'Brien M., Muttenthaler M., Miazzo M., Akcan M., Elliott A.G., Daly N.L., Harvey P.J., Arrowsmith S., Gunasekera S., et al., Oxytocic plant cyclotides as templates for peptide G protein-coupled receptor ligand design, Proc. Natl. Acad. Sci. U. S. A., 2013, 110, 21183-21188.

[65] Ladoukakis E., Pereira V., Magny E., Eyre-Walker A., Couso J.P., Hundreds of putatively functional small open reading frames in *Drosophila*, Genome Biol., 2011, http://genomebiology.com/2011/12/11/R118.

[66] Clamp M., Fry B., Kamal M., Xie X., Cuff J., Lin M.F., Kellis M., Lindblad-Toh K., Lander E.S., Distinguishing protein-coding and noncoding genes in the human genome, Proc. Natl. Acad. Sci. U.S.A., 2007, 104, 19428-19433.

[67] Koehbach J., Stockner T., Bergmayr C., Muttenthaler M., Gruber C.W., Insights into the molecular evolution of oxytocin receptor ligand binding, Biochem. Soc. Trans., 2013, 41, 197-204.

[68] Gruber C.W., Physiology of invertebrate oxytocin and vasopressin neuropeptides, Exp. Physiol., 2014, 99, 55-61.

[69] Ingolia N.T., Ghaemmaghami S., Newman J.R.S., Weissman J.S., Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling, Science, 2009, 324, 218-223.

[70] Guttman M., Russell P., Ingolia N.T., Weissman J.S., Lander E.S., Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins, Cell, 2013, 154, 240-251.

[71] Tamura K., Stecher G., Peterson D., Filipski A., Kumar S., MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0, Mol. Biol. Evol., 2013, 30, 2725-2729.

[72] Edgar R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res., 2004, 32, 1792-1797.

[73] Sigrist C.J.A., de Castro E., Cerutti L., Cuche B.A., Hulo N., Bridge A., Bougueleret L., Xenarios I., New and continuing developments at PROSITE, Nucleic Acids Res., 2013, 41, D344-D347.

[74] Rawlings N.D., Waller M., Barrett A.J., Bateman A., MEROPS: the database of proteolytic enzymes, their substrates and inhibitors, Nucleic Acids Res., 2014, 42, D503-D509.

[75] Stanke M., Keller O., Gunduz I., Hayes A., Waack S., Morgenstern B., AUGUSTUS: *ab initio* prediction of alternative transcripts, Nucleic Acids Res., 2006, 34, W435-W439.

[76] Lin M.F., Jungreis I., Kellis M., PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions, Bioinformatics, 2011, 27, i275-i282.

[77] Ferrè F., Clote P., DiANNA: a web server for disulfide connectivity prediction, Nucleic Acids Res., 2005, 33, W230-W232.

[78] Xue Y., Liu Z., Cao J., Ma Q., Gao X., Wang Q., Jin C., Zhou Y., Wen L., Ren J., GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection, Protein Eng. Des. Sel., 2011, 24, 255-260.

[79] Consortium T.U., Activities at the Universal Protein Resource (UniProt), Nucleic Acids Res., 2014, 42, D191-D198.