Little things make big things happen: A summary of micropeptide encoding genes

Jeroen Crappé*, Wim Van Criekinge, Gerben Menschaert*

Lab of Bioinformatics and Computational Genomics (BioBix), Department of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, 9000 Ghent, Belgium

ABSTRACT

Classical bioactive peptides are cleaved from larger precursor proteins and are targeted toward the secretory pathway by means of an N-terminal signaling sequence. In contrast, micropeptides encoded from small open reading frames, lack such signaling sequence and are immediately released in the cytoplasm after translation. Over the past few years many such non-canonical genes (including open reading frames, ORFs smaller than 100 AAs) have been discovered and functionally characterized in different eukaryotic organisms. Furthermore, in silico approaches enabled the prediction of the existence of many more putatively coding small ORFs in the genomes of Saccharomyces cerevisiae, Arabidopsis thaliana, Drosophila melanogaster and Mus musculus. However, questions remain as to what the functional role of this new class of eukaryotic genes might be, and how widespread they are. In the future, approaches integrating in silico, conservation-based prediction and a combination of genomic, proteomic and functional validation methods will prove to be indispensable to answer these open questions.

© 2014 The Authors. Published by Elsevier B.V. on behalf of European Proteomics Association (EuPA). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

1. Introduction

It is a well-known fact that small peptides play important roles in all kinds of biological processes [1]. The largest and most extensively studied class of small peptides comprises classical bioactive peptides. These are enzymatically cleaved from longer protein precursors containing an N-terminal signal sequence, hence directing the translation product toward the secretory pathway (see Fig. 1). Once released from the secretory vesicles, most of these peptides act as ligands of membrane receptors (mostly G protein-coupled receptors) and exert their extra-cellular biological signaling function in a autocrine, paracrine or endocrine way [2]. Examples are neuropeptides, peptide hormones and growth factors [3–6]. Other secreted peptides exercise their function in host defense systems having antimicrobial or toxic properties or show anti-hypertensive, antithrombotic or antioxidative activity [7,8]. Recently, other (non-classical) peptides – encoded by small open reading frames (sORFs) – have been discovered, presumably defining a new eukaryotic gene family [9–16]. These so-called micropeptides are translated from sORFs shorter.
than 100 AAs [14,17]. Sometimes these are also referred to as polycistronic peptides in the case were they are translated from a polycistronic mRNA [13] or as short open reading frame (sORF)-encoded polypeptides (SEPs) [18]. In contrast to other bioactive peptides, micropeptides are not cleaved from a larger precursor protein and lack an N-terminal signaling sequence. As such they are in principle released in the cytoplasm immediately after translation. This review focuses on this new class of peptides (see Fig. 1).

In the past, many molecules have been overlooked because of various biases and/or simplifications introduced in the performed discovery strategy. For example, it was only in 1993 that the first microRNA (lin-4) was discovered in Caenorhabditis elegans [19]. In the 2 subsequent decades more than 2500 microRNAs were identified in human alone [20]. Since micropeptides came into the limelight, ever more research is conducted to this new type of biomolecules, providing increasing evidence that this type of biomolecules is possibly also long overlooked [17]. It was assumed, especially for comprehensive cDNA annotation studies, that protein-coding genes do not code for translation products shorter than 100 AAs [21]. This arbitrarily chosen minimum length reduces the likelihood of false-positive detection by gene-prediction software and genome annotation algorithms, but at the same time vastly underestimates the true number of (atypical) small proteins [22,23]. This generalization is also noticeable in (manually curated) protein databases such as SwissProt-KB, where at the time of writing only 680 (3.4%) out of a total of 20,271 reviewed human proteins have a length shorter than 100 AAs. Although micropeptide research is not yet widespread and much remains to be learned about their abundance, functional activity and localization, a handful of these peptides have recently been functionally annotated in different eukaryotic organisms (see next paragraph for an extensive overview or Table 1 for a brief summary). Though important to an argument of the general conservation and function across all kingdoms of life, sORFs in bacteria and viruses [24–28], will not be covered in this review. The above-mentioned references can serve as a brief overview of putatively coding sORF (pcsORF) detection in lower organisms.

2. Overview of functionally annotated micropeptides

The first eukaryotic micropeptide was only described in 1996. While investigating the function of early nodule 40 (Enod40), formerly annotated as a ncRNA gene in legumes, van de Sande et al. transformed tobacco plants with a soybean GmENOD40-2 construct [29], proving that this construct was active in the non-legume tobacco, modulating the action of auxin. Sequence comparison of the tobacco and legume Enod40 clones revealed a highly conserved sORF coding for a 10 (tobacco) or 12 (soybean) AAs long peptide [29]. Later on, a second overlapping coding sORF of 24 AA was identified in soybean, categorizing Enod40 as a polycistronic mRNA. Enod40 is a well-known factor that functions in root nodule organogenesis in legumes and also displays a high sequence conservation among other plant species including monocots, suggesting a more general biological function [9]. In addition, Enod40 shows a highly conserved secondary topology, giving it the characteristics of a structural RNA [30]. The presence of peptide encoding sORFs and of structured RNA, both playing a role in developmental processes, indicates that Enod40 acts as a bi-functional or dual RNA [31,32]. Since the discovery of this first micropeptide in plants, others have been functionally annotated. In Arabidopsis, the POLARIS (PLS) gene, identified as a promoter trap transgenic line predominantly showing expression in the embryonic basal region, affects root growth and vascular development [10]. Mutation analysis has shown that the 36 AAs peptide encoded by PLS interacts with PIN proteins, forming a network that plays an important role in the hormonal crosstalk between auxin, ethylene and cytokinin

[33,34]. In maize, the recessive mutation of Brick1 (Brk1) leads
to several morphological defects of leaf epithelia [11]. The gene is highly conserved in plants as well as in animals and encodes a 76 AA peptide that lacks any targeting sequence. Research in Arabidopsis has shown that Brk1 is a critical WAVE-complex subunit functioning in a pathway with the ARP2/3 complex [35,36]. **ROTUNDIFOLIA (ROT4)** was identified as an overexpressed novel single exon gene encoding a small 53 AAs peptide in an Arabidopsis mutant with short leaves and floral organs [37]. Phylogenetic analysis in Arabidopsis indicates that ROT4 defines a novel seed-plant specific small peptide gene family, comprising 22 **ROT FOUR LIKE (RTFL)** genes, sharing a conserved 29 AAs region [12]. More recently, two novel maize sORF genes, Zm401p10 and Zm908p11 respectively encoding 89 and 97 AA peptides were identified, playing a required role in pollen development [38–40].

Micropeptide research is not limited to plants; some of the best-studied sORF genes have been identified in the animal kingdom. In silico prediction analysis of cDNAs in Drosophila melanogaster identified several mRNA-like ncRNA candidates putatively encoding ORFs [41,42]. Extensive study on one of these candidates by several groups revealed that the evolutionary conserved tarsal-less (tal) or polished rice (pri) in Drosophila and the orthologous mille-pattes (mlpt) in Tribolium is in fact a polycistronic gene encoding small peptides [13–15]. The tal gene contains a total of 4 sORFs encoding functionally redundant peptides of length 11–32 AAs, playing a role in Drosophila embryogenesis. The absence of a functional tal gene, either by knock-down or via mutation, leads to the absence of trichomes on the body surface, a missing tarsal region, ectopic leg morphogenesis and abnormal abdominal belt and tracheal formation [13,14,43]. On the other hand, the overexpression of tal peptides negatively modulates the Notch signaling pathway [44]. Extensive molecular analysis showed that tal controls epidermal differentiation by modifying the transcription factor **Shavenbaby** (Svb) from a transcriptional activator to a repressor, thus dominantly inhibiting its downstream function in trichome formation [45]. Similarly, RNAi depleted mlpt embryos alter gap gene expression, generally leading to shortened embryos with additional pairs of legs, also missing some posterior abdominal segments [15].

Recently, another member of this set of mRNA like ncRNAs (also comprising the tal gene) led to the identification of two new coding sORFs (28 and 29 AAs long) in the putative noncoding RNA 003 gene (pncr003:2L) [16,41]. Phylogenetic conservation analysis indicated that the pncr003:2L gene shares a common sORF encoding ancestor gene with the human sarcolipin (sln) and its longer parologue phospholamban (pln). In order to reflect their similarity the researchers suggested to rename the pncr003:2L gene and its arthropod homologs to sarcolamban (slc). Visualizing intracellular Ca2+ levels in slc mutants and wild-type controls identified a primary role for slc encoded peptides during the Ca2+ trafficking at the sarcoplasmic reticulum (SER) which is required for heart muscle contraction [16].

### Table 1 - Functionally annotated micropeptides in Eukaryotes.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>sORF length</th>
<th>Species</th>
<th>Proposed function</th>
<th>Conservation</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early nodulin 40(Dnod40)</td>
<td>12, 24</td>
<td>G. max</td>
<td>root nodule organogenesis</td>
<td>Legumes &amp; Monocots</td>
<td>[9,29,30]</td>
</tr>
<tr>
<td>POLARIS (PLS)</td>
<td>36</td>
<td>A. thaliana</td>
<td>hormonal crosstalk during embryogenesis</td>
<td>Plants and Animals</td>
<td>[10,33,34]</td>
</tr>
<tr>
<td>Brick1 (Brk1)</td>
<td>76</td>
<td>Z. mays</td>
<td>morphogenesis of leaf epithelia</td>
<td>Plants (HSPC300 homolog)</td>
<td>[11,35,36,46]</td>
</tr>
<tr>
<td>Rotundifolia (ROT4)</td>
<td>53</td>
<td>A. thaliana</td>
<td>leaf shape morphogenesis</td>
<td>Plants</td>
<td>[12,37]</td>
</tr>
<tr>
<td>Zm401p10</td>
<td>89</td>
<td>Z. mays</td>
<td>tapetum development</td>
<td>Pooceae</td>
<td>[38,39]</td>
</tr>
<tr>
<td>Zm908p11</td>
<td>97</td>
<td>Z. mays</td>
<td>pollen tube growth</td>
<td>Arthropods and Daphnia</td>
<td>[40]</td>
</tr>
<tr>
<td>tarsal-less (tal)</td>
<td>3 × 11, 32</td>
<td>D. melanogaster</td>
<td>leg and actin-based cell</td>
<td>Vertebrates and Arthropods</td>
<td>[13–15,43–45]</td>
</tr>
<tr>
<td>sarcolamban (scl)</td>
<td>28, 29</td>
<td>D. melanogaster</td>
<td>morphogenesis during embryogenesis</td>
<td></td>
<td>[16]</td>
</tr>
</tbody>
</table>

*Number of amino acids in the different translated and functional open reading frames. The organism in which the function has been best studied experimentally.*
one or more intercellular signaling pathways [13,17,47]. More speculative, their small molecular size makes them ideal candidates for cellular translocation as gap peptides, to function as membrane permeable peptides, or to leave the cell via cell-derived (micro) vesicles such as exosomes [48–52].

Conservation of the coding sequence across very large evolutionary distances is another peculiar feature of micropeptides. The recently identified sarcolamban shows conservation of more than 550 million years (see Fig. 2). Analysis of the related peptides indicates a conserved peptide sequence and molecular structure from flies to vertebrates all involved in the regulation of Ca\(^{2+}\) traffic [16]. Tarsal-less is another example of a highly conserved micropeptide. Tal homologs could be identified in other insects and even in Daphnia pulex, a crustacean species. This gene family is at least 440 million years old and shows a varying number of sORFs with an evolutionary trend toward accumulation of more ORFs [14]. A systematic search for new genes in Saccharomyces cerevisiae [53] (see also next paragraph) led to the identification of, among others, smORF2. Functional homologs for this gene with a temperature-sensitive phenotype could be identified in many organisms from yeast to human [54]. To our knowledge, Brk1 is however by far the most conserved sORF encoding gene. Next to being highly conserved throughout the plant kingdom, homologs for the maize brk1 gene have been identified in almost all studied animal eukaryotic genomes, including human, where HSPC300 (mammalian homolog of brk1) also functions in the Scar/WAVE complex [11,36]. It remains to be confirmed that conservation is a typical characteristic of sORF-encoding genes. Most discovered micropeptide genes result from large genetic or computational screenings, focusing on phylogenetic conservation as a proxy to functionality, in this way choosing interesting targets for further (elaborate) downstream in vivo research. This might introduce a bias toward discovery of highly conserved sORFs. On the other hand, a highly conserved state can also point to the role most of these genes play in (embryonic) development, morphogenesis and other very basic and important biological processes. As such, the high sequence and functional conservation of this new type of eukaryotic gene products, might explain many basic but very important functions shared by a plethora of species over different kingdoms.

4. Systematic searches for putatively coding sORFs and micropeptides

sORF encoding genes have, in our opinion, long been overlooked. However, the past decade has seen some important advances in the (genome-wide) identification of pcsORFs. To identify those new and interesting candidates in the vast amount of random sORFs scattered all over the genome, in silico strategies (often making use of expression data) have been devised. S. cerevisiae was the first eukaryotic species to be the subject of such a systematic and elaborate scan. Many yeast sORFs were identified based on comprehensive sequence
database searching. Homology, comparative genomics and expression features from serial analysis of gene expression (SAGE), Northern blotting, RT-PCR and ORF tagging experiments were taken into account [54–61]. Later, Kastenmayer et al. concluded that, based on the above-mentioned independent experimental approaches and computational analyses, at least 299 pcsORFs are present in the S. cerevisiae genome, many of which have potential orthologs in other eukaryotic species. This represents a significant percentage (~5%) of the amount of annotated genes in S. cerevisiae. Of these identified pcsORFs, four were confirmed to produce a translation product and 22 seem to regulate growth [53].

Arabidopsis was the first plant species undergoing a thorough in silico analysis in the search of new sORF encoding genes. Because common gene-finding algorithms have a hard time identifying small protein products and are prone to a high number of false negatives (as already mentioned in the introduction) Hanada et al., developed the Coding Index (CI) measure for pcsORF prediction based on the hexamer composition bias, a general measure to distinguish CDS from non-CDS [62,63]. This CI measure would later form the basis for a specific program package to identify sORFs, named sORFfinder [64]. After performing a six reading frame translation of intergenic sequences of Arabidopsis thaliana, pcsORFs were only assigned as being coding when they demonstrated qualifying CI values, above background tiling array hybridization intensities, evidence of purifying selection based on Ka/Ks values and overlapping ESTs. Using these criteria, 7159 pcsORFs with high coding potential, of which 2376 are subject to purifying selection, were identified in A. thaliana [62,64]. In a recently published follow-up study, elaborating on the function of these new pcsORFs, an array was designed to generate an expression atlas at several developmental stages and under multiple environmental conditions for the 7901 identified pcsORFs [65]. 473 pcsORFs showed a high number of homologs in other plant species and were overexpressed. 49 of those expressed and significantly conserved pcsORFs induced various morphological changes and visible phenotypic effects [65].

Arabidopsis is not the only plant species subject to an integrative procedure to identify pcsORFs at the genome level. After obtaining ~2.6 million expressed sequence tag (EST) reads from a Populus deltoides leaf transcriptome, full-length transcripts from the EST sequences could be reconstructed. Using a computational approach based on coding potential, evolutionary conservation and gene family clustering, and by showing evidence of protein domains, ncRNA motifs, sequence length distribution or mass-spectrometry data, at least 56 pcsORF encoding genes (~200 AAs) new to the Populus genome annotation could be identified [66]. Very recently, work was published exploiting publicly available genome sequences of Phaseolus vulgaris, Medicago truncatula, Gycine max and Lotus japonicus in a search for pcsORFs (30–150 AAs) [67]. A bioinformatics analysis was performed based on evidence of expression (transcription level), presence of known protein regions or domains (translation level) and identification of orthologues genes in the explored genomes. Respectively 6170, 10461, 30521 and 23599 pcsORFs were uncovered within the P. vulgaris, G. max, M. truncatula and L. japonicus genomes. Based on specific EST expression analysis in P. vulgaris, 2336 of the identified pcsORFs showed evidence of gene expression.

In the animal kingdom, the first in silico and systematic search for new pcsORFs was carried out for the model organism D. melanogaster [68]. Starting from putatively non-coding euchromatic DNA, an initial set of 593 586 open reading frames between 30 and 300 basepairs (bps) long could be identified. Using tBlastn, all pcsORFs showing significant similarity with annotated coding sequences or transposons were removed, at the same time only retaining pcsORFs showing significant amino acid sequence similarity with Drosophila pseudoobscura. After realigning extended versions of the conserved pcsORFs with ClustalW, an upper estimate of 4561 pcsORFs were identified in Drosophila. 72% of the in D. pseudoobscura conserved pcsORFs appeared to be true homologs as they were conserved in syntenic regions with regard to the original D. melanogaster pcsORF. Only taking into account syntenic pcsORFs with favorable Ka/Ks values (having a ratio below 0.1), and with transcriptional evidence (based on combining both publicly available RNA-seq and tiling array data), the authors postulate that at least 401 functional sORFs exist in the D. melanogaster genome [68].

Very recently, Crappé et al. combined an in silico approach and experimental evidence by means of ribosome profiling data (see also next paragraph) for a genome-wide search to detect novel pcsORFs in the Mus musculus genome [69]. First, the genome was scanned for sORFs with high coding potential using the sORFfinder package. Secondly, a comprehensive feature matrix with peptide conservation measures, based on UCSC multiple species alignments, was constructed. In a third step, the coding capabilities of these pcsORFs were assessed by means of a machine-learning algorithm. Afterwards, the sORFs with a high coding score were verified for the presence of experimental ribosome profiling signals obtained from mouse Embryonic Stem Cells (mESCs), hinting to sORF translation. Using this combined genome-wide approach dozens of both highly conserved and ribosome-targeted pcsORFs (possibly encoding micropeptides) were identified [69].

5. sORF identification using ribosome profiling

Ribosome profiling is a recently described new strategy to monitor protein synthesis based on deep sequencing of ribosome protected mRNA fragments [70–74]. By exploiting the properties of drugs as harringtonine, puromycin or lacticidomycin, stalling ribosomes at Translation Initiation Sites (TIS), the study of (alternative) (a)TIS with subcodon to single-nucleotide resolution is now possible [28,73,75–79]. In an attempt to provide a genome-wide map of protein synthesis, Ingolia et al. exploited a machine learning algorithm on top of their ribosome profiling data to systematically delineate protein products in mESCs [75]. Special attention was paid to a recently identified and apparently abundant class of RNAs, referred to as long non-coding RNAs (lncRNAs). These lncRNAs were scanned for translated regions, by defining the most highly ribosome-occupied 90 nucleotide window
In this way, Ingolia et al. were able to identify many IncRNA regions, displaying high ribosomal occupancy and containing small open reading frames, classifying them as short polycistronic ribosome-associated coding RNAs (spcRNAs) [75]. Motivated by these findings and while performing a global translation initiation analysis using ribosome profiling (GTI-seq) in HEK293 cells, Lee et al. also specifically characterized the translation of ncRNAs. They were able to identify 228 ncRNAs associated with GTI-seq sequencing reads, often spanning evolutionarily conserved sORFs (median length of 54 nt) and frequently showing alternative initiation at non-AUG start codons [78]. Our own research on mESC ribosome profiling data strengthens the idea that some ncRNAs actually contain putatively coding sORFs. While investigating sORFs within annotated IncRNA regions, we were also able to detect very well-conserved and ribosome targeted pcsORFs [69].

The question if and more specifically to what extent IncRNAs act through their translational sORF products remains up for debate and is one that will not find an answer based on ribosome profiling data alone. For example, the mouse H19 IncRNA transcript functions as a true ncRNA [82,83], even though demonstrating ribosome occupancy [75,84]. This proves that simply ribosome profiling does not suffice as evidence of protein synthesis nor can be proposed as a fool-proof method to distinguish between coding and non-coding transcripts [85,86]. In addition, one has to keep in mind that spurious association of ribosomes could lead to translational noise [87]. The fact that most of the predicted IncRNA transcripts that encode sORFs lack any significant conservation and that IncRNAs are rarely translated in human cell lines seems consistent with these observations [80,81,88,89]. In a follow up study Guttmann et al. developed a metric, the ribosome release score (RRS), enabling sensitive identification of functional protein-coding transcripts based on the termination of translation at the end of the ORF [90,91]. With this metric it is possible to discriminate between protein-coding transcripts and other classes of non-coding transcripts, including IncRNAs. Because the class of IncRNAs closely resembled the ribosome occupancy of other classes of non-coding transcripts with respect to this metric, it was deemed unlikely that IncRNAs, as a class, produce functional products [91]. Future measures will certainly be devised to assess the true coding potential of ribosome profiling occupied mRNA on a case per case basis.

Although most research at the moment points to the true non-coding state of IncRNAs, a subclass could still comprise pcsORFs, considering that some of them have proven to be highly conserved [69]. The absence of detectable peptide products does furthermore not rule out their existence, as the expression of IncRNA encoded sORFs could be very specific in time as well as in space [45,92]. An attractive hypothesis could be that IncRNAs are generally non-coding, but under specific circumstances, enclosed sORFs can be translated (presumably at very low levels), thus rendering these IncRNAs as bifunctional or dual RNAs [31,32,93]. In the end, only scrutinized and functional in vivo analysis will be able to prove if and what IncRNA transcripts give rise to functional small protein products [16,45,94].

6. Detection of micropeptides using mass spectrometry

Few studies exist where mass spectrometry is used for the direct detection of micropeptides, although this still is the gold standard when looking for protein or peptide products. Using a newly developed strategy, combining peptidomics and massive parallel RNA-seq, Slavoff et al. claim the discovery of many previously uncharacterized human sORF-encoded polypeptides (SEPs) in K562 (human leukemia) cells [18]. First, custom databases were constructed containing all possible polypeptides based on the annotated human transcriptome (RefSeq) and an experimental RNA-seq derived K562 transcriptome. Identifying the peptidomics mass spectrometry fragmentation spectra (MS/MS) using these custom polypeptide databases and four previously reported SEPs as positive controls, an extra 86 still uncharacterized SEPs were discovered, bringing the total of unannotated human SEPs to 90 [18,95].

Although this study is one of the early attempts to systematically identify micropeptides by means of mass spectrometry and subsequent peptide-to-spectrum matching strategy, they largely failed to prove the mature forms of micropeptides. Alternative strategies, for which the above approach could serve as a guideline, should lead to the true identification of mature and native forms of endogenous micropeptides as this is still one of the most important aspects of peptidomics.

7. Outlook

Although there are many sophisticated gene prediction programs available, the majority is optimized to predict genes with 100 or more codons, rendering them inappropriate for sORF detection [96–98]. Development of ab initio single-sequence methods (based on codon patterns) and discriminative metrics (pairwise and multi-species alignment-based comparative metrics), suited for the detection of small ORFs, is still in its infancy. On shorter exons, comparative metrics clearly outperform single-sequence based methods, adding discriminatory power as additional species are used. Using hybrid metrics, exploiting the relative independence of their input metrics, further increases performance [94,98,99]. Based on these findings, Crapped and al. combined several metrics, computed from a multi-species alignment and subsequently built a classifier model (using a Support Vector Machine) to classify coding versus non-coding. Although the combination of different metrics generally leads to better performance, it is still dependent on the correctness of the multiple sequence alignment, does not incorporate near-cognate start sites and possibly misses a lot of highly divergent and/or quickly diverging pcsORFs [69]. New and promising metrics with regard to this growing field of sORF detection will certainly emerge. In this respect, PhyloCSF is certainly noteworthy. It is a comparative genomics method that analyzes a multiple sequence alignment using phylogenetic codon models to correctly distinguish between protein-coding and non-coding regions. PhyloCSF clearly outperforms other methods for the analysis of short exons [100].
An integrated approach combining computational and experimental validation stands a better chance to result in meaningful findings than merely performing an in silico prediction [94]. Slavoff et al. compiled a custom mRNA-seq derived polypeptide database to identify mass spectrometry fragmentation spectra and were able to identify 86 uncharacterized SEPs [18]. However, for reasons already mentioned in this paper, the ribosome profiling technique is more suitable than mRNA-seq to delineate the exact ORFs and thus derive putative micropeptide sequences. Menschaert et al. prove that a ribosome profiling (Ribo-seq) derived custom database yields a highly informative search space of translation products for MS/MS-based peptide identification [79]. An automated pipeline converting Ribo-seq information into a custom pcsORF sequence database, by delineating open reading frames from calling the translation start sites and detecting SNP mutations, will prove to be very beneficial in future MS-based studies.

Known micropeptides have a very narrow expression in time as well as in space [45]. These characteristics are probably part of the reason why tarsal-less, one of the best-studied micropeptides to date, has never been identified using mass spectrometry. New and alternative extraction methods should prove to be more effective at extracting cytoplasm bound micropeptides [3]. For example, Schwaad et al. recently reported on an affinity-based approach to enrich and identify cysteine-containing human sORF encoded polypeptides (ccSEPs) from cells. Using this approach they were able to identify 16 novel ccSEPs, derived from uncharacterized sORFs [101]. The development of new mass spectrometry based techniques, such as the reported chemoproteomic approach, will prove indispensable in order to identify and characterize the biological function of micropeptides.

The mere existence of a peptide does not imply that it has a function. Evolutionary conservation is definitely suggestive for functionality, but to pinpoint the actual function, experimental demonstration of a biological effect is required [94]. Approaches used to functionally describe micropeptides such as tarsal-less or sarcolambcan can be seen as a general guide for further in vivo analyses [16,45].

8. Conclusion

Research on short peptides, encoded by small open reading frames, is still in its infancy. Nevertheless, growing evidence points to the existence of these so-called micropeptides, but to what extent and how important this class of new biomolecules is, still needs to be seen. Approaches integrating in silico, conservation-based prediction and a combination of genomic, proteomic and functional validation methods will prove to be indispensable to further explore this micropeptide research field.

Acknowledgments

The financial support of the Institute for the Promotion of Innovation in Flanders (IWT) and the Belgian National Fund for Scientific Research (FWO-Flanders) is gratefully acknowledged. Dr. G. Menschaert is supported by a postdoctoral fellowship of FWO-Flanders. J. Crappe is supported by a fellowship of IWT-Flanders.

References


