Small Proteins, Big Data

George W. Preston

King's College London, 150 Stamford Street, London, SE1 9NH; e-mail: george.preston@kcl.ac.uk.

The idea that a protein consisting of relatively few amino acids (e.g., one hundred or fewer) could be biologically important is not, of course, a new one. One well-known example is ubiquitin (76 amino acid residues), a protein with various regulatory roles in eukaryotes (Ozkaynak *et al*., 1987). Another, even smaller example is sarcolipin (Fig. 1), a 31-amino-acid protein found in the membrane of the sarcoplasmic reticulum in skeletal muscle (Delcourt *et al*., 2018; Mascioni *et al*., 2002). Recently, however, it has been realised that the methods by which proteins are commonly discovered do not necessarily give small proteins a chance to be detected. Now, by integrating analyses done at both the nucleic acid and protein levels, investigators are able to probe the 'small' fraction of the proteome more comprehensively. In this brief survey, I will discuss what defines a small protein and how small proteins are analysed.

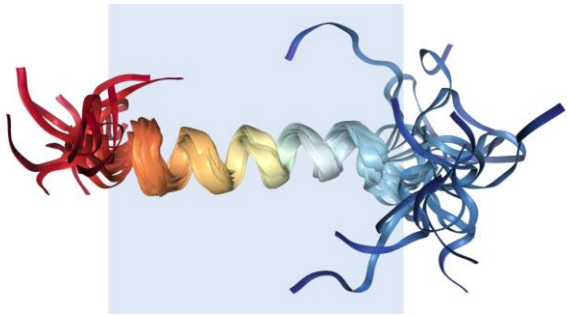

Fig. 1. Models of the three-dimensional structure of human sarcolipin based on data from nuclear magnetic resonance spectroscopy (Mascioni *et al*., 2002). Image of Protein Data Bank entry 1JDM (Mascioni *et al*., 2002) created with NGL Viewer (Rose & Hildebrand, 2015) and embellished in Inkscape (version 0.92). The shaded block represents the approximate extent of a lipid environment.

DNA Makes RNA Makes (Small) Protein

In eukaryotes and prokaryotes alike, protein synthesis involves the translation of messenger ribonucleic acid (mRNA). mRNA contains at least one *open reading frame*—a run of nucleotides, beginning with a start codon and ending with a stop codon, that is translated into a polypeptide by the ribosome. Analysis of the cellular mRNA pool is one way of enumerating small proteins, and the detection of open reading frames is central to this approach. Without evidence at the mRNA level, there would be no way of knowing that an observed small protein was not a proteolytic fragment of a larger protein. Indeed, ubiquitin—although inarguably small—does not have its own open reading frame (it is, in yeast at least, a proteolytic fragment of one of four larger proteins; Ozkaynak *et al*., 1987). On this basis, ubiquitin would be ignored by methods that detect small proteins *via* their open reading frames. On the other hand, it is possible for a small protein to go undetected at the mRNA level. This could happen if its open reading frame co-existed, in the same mRNA, with that of a larger protein (Delcourt *et al*., 2018).

How Small is Small?

"Small proteins", write VanOrsdel *et al*. in a recent special issue of *Proteomics*, are proteins "containing 75 or fewer amino acids and encoded in a short open reading frame" (VanOrsdel *et al*., 2018). To me, these sounded similar to the "microproteins" reported by He *et al*. in a recent article in *Journal of Proteome Science* (He *et al*., 2018). Microproteins, it turns out, can be a little larger ("peptides composed of 100 amino acids ... or fewer"), and some authors use the term to refer to proteins whose function is analogous to that of microRNA; but the point is that the cut-off length for a small protein is somewhat arbitrary. This is not to say, however, that there isn't a good reason for having a cut-off. From the bioinformatician's perspective, there is a very real possibility of a short open reading frame occurring in mRNA purely by chance, and it therefore makes sense to control for this. The use of *proteogenomics* approaches, which involve the cross-referencing of protein and nucleic acid sequences, can help to sort the true proteins from the noise (see case study).

PROTEIN OR PEPTIDE?

The answer to this question is largely academic, particularly in cases where only the sequence of amino acids is known. The definitions of 'protein' and 'peptide' are actually quite loose (IUPAC-IUB Joint Commission on Biochemical Nomenclature, 1984), but 'peptide' tends to be used for the shorter, less-structured polymers, and 'protein' for longer, more folded ones. To give a concrete example, consider a procedure that we routinely perform in our laboratory—the enzymatic digestion of albumin with trypsin. Surely no one would dispute that albumin (or trypsin, for that matter) is a protein, and that the products of the digestion are peptides. It should not be assumed, however, that a short polymer should lack higher-order structure or, conversely, that a long polymer should be highly ordered. Globular structure has been detected in polymers as short as ten amino acid residues (Honda *et al.*, 2008), whilst some large proteins have been observed to be substantially unfolded in the native state (Fink, 2005).

CASE STUDY: SMALL PROTEINS IN YEAST

He *et al.* used the proteogenomics approach to look for small proteins in baker's yeast (*Saccharomyces cerevisiae*) (He *et al.*, 2018). In doing so, the authors first had to address a problem that is encountered frequently when investigating a specific subset of proteins—the need for enrichment. For He *et al.*, enrichment entailed the removal of proteins heavier than about 30 kDa (notice how, for practical purposes, the cut-off is now a *mass* rather than a length). After lysing cells and extracting their contents, four different methods of enrichment were trialled. Two of these involved separating the proteins using gel electrophoresis, whilst the other two used molecular-weight cut-off filtration. Whichever method was used, the enriched small proteins were eventually reduced, *S*-alkylated and digested before being analysed—as peptides—by ultraperformance liquid chromatography with on-line mass spectrometry. The most interesting of the detected peptides (the identity of which was confirmed using a synthetic standard) led the authors to a novel gene. A bioinformatic analysis indicated that this gene, which the authors named 'YKL104W-A', encodes a small protein of 84 amino acid residues.

CONCLUSION

It seems likely that small proteins are currently underrepresented in the scientific literature and in databases. This is in spite of their potentially important biological functions. Technical and conceptual advances appear to be remedying the situation and, in the present era of omics methods and 'big data', it is possible that many more small proteins will be discovered.

REFERENCES

1. Delcourt V, Staskevicius A, Salzet M, Fournier I, Roucou X (2018) Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA. *Proteomics* 10, article number 1700058. PMID: 28627015
2. Fink AL (2005) Natively unfolded proteins. *Curr. Opin. Struct. Biol.* 15, 35-41. PMID: 15718131
3. He C, Jia C, Zhang Y, Xu P (2018) Enrichment-based proteogenomics identifies microproteins, missing proteins, and novel smORFs in *Saccharomyces cerevisiae*. *J. Proteome. Res.* 17, 2335-2344. PMID: 29897761
4. Honda S, Akiba T, Kato YS, Sawada Y, Sekijima M, Ishimura M, Ooishi A, Watanabe H, Odahara T, Harata K (2008) Crystal structure of a ten-amino acid protein. *J. Am. Chem. Soc.* 130, 15327-15331. PMID: 18950166
5. IUPAC-IUB Joint Commission on Biochemical Nomenclature (1984) Nomenclature and symbolism for amino acids and peptides. *Biochem. J.* 219, 345-373. PMID: 6743224
6. Mascioni A, Karim C, Barany G, Thomas DD, Veglia G (2002) Structure and orientation of sarcolipin in lipid environments. *Biochemistry* 41, 475-482. PMID: 11781085
7. Ozkaynak E, Finley D, Solomon MJ, Varshavsky A (1987) The yeast ubiquitin genes: a family of natural gene fusions. *EMBO J.* 6, 1429-1439. PMID: 3038523
8. Rose AS, Hildebrand PW (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.* 43, W576-W579. PMID: 25925569
9. VanOrsdel CE, Kelly JP, Burke BN, Lein CD, Oufiero CE, Sanchez JF, Wimmers LE, Hearn DJ, Abuikhdair FJ, Barnhart KR, Duley ML, Ernst SEG, Kenerson BA, Serafin AJ, Hemm MR (2018) Identifying new small proteins in *Escherichia coli*. *Proteomics* 10, article number 201700064. PMID: 29645342