discover proteins that interact with DNA and RNA in different cellular contexts. This could provide an integrated description of how DNA-, RNA-, and protein-protein interactions govern cell physiology.

In a recent study, Leuenberger *et al.* used proteome-wide heat denaturation to measure protein stability from cell lysates using a different method called limited proteolysis–coupled mass spectrometry (*11*). They found that half of the detected proteins that were computationally predicted to lack stable tertiary structures (that is, intrinsically disordered) exhibited a two-state denaturation profile, which is indicative of a stable structure. This seeming contradiction may now be interpreted in light of Tan *et al.*'s findings. Because intrinsically disordered proteins (IDPs) interact with other structural partners, this may result in a melting curve similar to that seen for structured proteins.

Because of their lack of stable tertiary structure and their promiscuous interactions, IDPs are referred to as the dark proteome (*12*, *13*). Techniques such as TPCA

"TPCA [thermal proximity coaggregation] can be performed on intact cells... allowing proteome-wide detection of interactions."

could provide much-needed insights into protein-protein interactions involving IDPs in a cellular context, and on a proteome-wide scale. This would be especially useful considering the role of IDPs in modulating protein interaction networks. By offering the possibility to decipher and interpret the dynamic interactome, techniques such as TPCA may be the key to determining how cellular function emerges from dynamic changes in protein interaction networks. ■

REFERENCES AND NOTES

1. E. L. Huttlin *et al.*, *Cell* **162**, 425 (2015).
2. T. Rolland *et al.*, *Cell* **159**, 1212 (2014).
3. N. Sahni *et al.*, *Cell* **161**, 647 (2015).
4. X. Wang *et al.*, *Nat. Biotechnol.* **30**, 159 (2012).
5. C. S. H. Tan *et al.*, *Science* **359**, 1170 (2018).
6. K. Ghosh, K. Dill, *Biophys. J.* **99**, 3996 (2010).
7. R. Jafari *et al.*, *Nat. Protoc.* **9**, 2100 (2014).
8. D. Martinez Molina *et al.*, *Science* **341**, 84 (2013).
9. M. M. Savitski *et al.*, *Science* **346**, 1255784 (2014).
10. K. V. Huber *et al.*, *Nat. Methods* **12**, 1055 (2015).
11. P. Leuenberger *et al.*, *Science* **355**, eaai7825 (2017).
12. A. Bhowmick *et al.*, *J. Am. Chem. Soc.* **138**, 9730 (2016).
13. https://darkproteome.wordpress.com/about/what-is-the-dark-proteome/

ACKNOWLEDGMENTS

We thank G. Slodkowicz and M. M. Solano for reading the manuscript and the Medical Research Council (MC_U105185859) and European Research Council (ERC-COG-2015-682414; IDR-Seq) for financial support.

10.1126/science.aat0576

PROTEOMICS

Proteoforms as the next proteomics currency

Identifying precise molecular forms of proteins can improve our understanding of function

By **Lloyd M. Smith**[1] *and* **Neil L. Kelleher**[2]

Proteoforms—the different forms of proteins produced from the genome with a variety of sequence variations, splice isoforms, and myriad posttranslational modifications (*1*)—are critical elements in all biological systems (see the figure, left). Yang *et al.* (*2*) recently showed that the functions of proteins produced from splice variants from a given gene—different proteoforms—can be as different as those for proteins encoded by entirely different genes. Li *et al.* (*3*) showed that splice variants play a central role in modulating complex traits. However, the standard paradigm of proteomic analysis, the "bottom-up" strategy pioneered by Eng and Yates some 20 years ago (*4*), does not directly identify proteoforms. We argue that proteomic analysis needs to provide the identities and abundances of the proteoforms themselves, rather than just their peptide surrogates. Developing new proteome-wide strategies to accomplish this goal presents a formidable but not insurmountable technological challenge that will benefit the biomedical community.

The function of proteins can be strongly modulated by posttranslational modifications (PTMs) such as phosphorylation (consider kinase cascades), acetylation, methylation (consider histones), and many more of the >400 known PTMs in biology. These sources of variation combine to create a complex and largely uncharted world of natural proteins. Knowledge of the identities and quantities of these proteoforms present in dynamic biological systems is indispensable to development of a complete picture of functional regulation at the protein level.

Conventional proteomics digests protein mixtures into peptides, some of which are identified by tandem mass spectrometry (MS). Each identified peptide acts as a surrogate for the presence of the protein molecule from which it is derived. This strategy provides invaluable information on protein expression in complex systems. However, as many different gene products, isoforms, and proteoforms can contain the same peptide, direct information about the proteoforms present is lost (see the figure, bottom). This issue is the proteomic analog of the problem of "phasing" in genomics (*5*)—determining whether multiple alleles are present on the same segment of DNA. The step of digestion into peptides is essential to the success and robustness of the bottom-up strategy, as well-behaved peptides are more amenable to liquid chromatographic separation and MS analysis than are intact proteins. However, only inferences can be made as to the actual proteoform or proteoforms from which the identified peptide was derived (*6*).

An alternative approach is "top-down" proteomics, in which whole proteins are analyzed directly using tandem MS methods (see the figure, top left). Although great strides have recently been made in the top-down analysis of high-mass proteins (*7*) and complex proteomic samples (*8*), limitations remain to be addressed in the degree of sequence coverage and the ability to analyze low-abundance species. A complementary approach reported the proteome-wide identification of proteoforms in yeast, based primarily upon a high-accuracy determination of their intact mass, aided by a corollary measurement of the number of lysine residues in the molecule (*9*). Comparison of the measured masses and lysine counts with a theoretical database of possible yeast proteoforms yielded proteoform identifications. Further comparisons of all experimental masses with one another revealed related proteoforms differing by common PTMs, yielding more identifications. These pairwise relations (experimental:theoretical and experimental:experimental) were assembled into "families" of related proteoforms (see the figure, top center).

Such "proteoform families" offer a new and more detailed way of viewing the proteome (see the figure, top right). To extend the strategy to mammalian genomes, RNA sequencing can be used to construct sample-specific proteoform databases that capture the genetic variation and extent of splicing

[1]*Department of Chemistry, University of Wisconsin, 1101 University Avenue, Madison, WI 53706-1396, USA.* [2]*Departments of Chemistry and Molecular Biosciences and Feinberg School of Medicine, Northwestern University, Evanston, IL 60208, USA. Email: smith@chem.wisc.edu*

patterns in the sample (*10, 11*). Integrating such proteogenomic data with synergistic information obtained from bottom-up (for PTM identification and localization), top-down (for protein identification and PTM localization), and intact mass measurements (for proteoform identification) can provide the comprehensive analysis needed to broadly identify and quantify proteoforms in complex samples.

The question of how many proteoforms exist in nature quickly arises in this discussion (*12*). This question may prove impossible to answer fully, as errors in transcription and translation can produce numerous low-abundance proteoforms, perhaps as few as only a single molecule per cell, or even a single molecule in a large population of cells. We currently can only detect proteoforms present at concentrations above the instrumental detection limits of existing mass spectrometers, although the advent of single-molecule nanopore or other strategies for proteoform identification may change that landscape in the future.

However, the number and variety of proteoforms expressed in biological systems appear to be far below the calculated combinatorial possibilities (*12*). Garcia and coworkers have pioneered MS methods for histone proteoform analysis, finding much smaller numbers of histone proteoform variants than the maximal number of combinatorial possibilities would suggest (*13*). Similarly, in a deep study of histone H4 proteoforms by Coon and co-workers, only 74 were identified (*14*). This stands in striking contrast to the ~3 million possibilities that are theoretically possible from the combinatorial explosion of known site-specific modifications (*14*). This difference may simply indicate that many or most proteoforms are not detectable with current technology, and that we are only able to see at present the few of those that are most abundant. Alternatively, nature may only make and use a small subset of the proteoforms that are theoretically possible, as deduced from the combinatorial possibilities offered by considering all of the various possible PTM combinations. Understanding which of these explanations is correct, or perhaps a blend of both, will require improved technologies that can reveal proteoforms at ever lower abundance.

Proteoform analyses will become increasingly straightforward as information is accrued and archived on the proteoforms that actually exist in nature and can be observed. Establishing a comprehensive atlas of identified proteoforms for human and other species has begun, and over time this atlas will begin to yield transformative insights into the levels and roles of proteoform complexity present in biological systems. As proteoforms are tightly linked to the functioning of cells and tissues that underlie complex phenotypes, their identification and quantification will provide critical insights into the fundamental workings of biological systems (see the figure, top right). Proteoforms should also help identify key diagnostic markers and therapeutic targets and thereby provide greater statistical power for deciphering human disease phenotypes. ■

REFERENCES AND NOTES

1. L. M. Smith *et al.*, *Nat. Methods* **10**, 186 (2013).
2. X. Yang *et al.*, *Cell* **164**, 805 (2016).
3. Y. I. Li *et al.*, *Science* **352**, 600 (2016).
4. J. K. Eng, A. L. McCormack, J. R. Yates, *J. Am. Soc. Mass Spectrom.* **5**, 976 (1994).
5. S. R. Browning, B. L. Browning, *Nat. Rev. Genet.* **12**, 703 (2011).
6. A. I. Nesvizhskii, R. Aebersold, *Mol. Cell. Proteomics* **4**, 1419 (2005).
7. X. Han, M. Jin, K. Breuker, F. W. McLafferty, *Science* **314**, 109 (2006).
8. J. C. Tran *et al.*, *Nature* **480**, 254 (2011).
9. M. R. Shortreed *et al.*, *J. Proteome Res.* **15**, 1213 (2016).
10. X. Wang *et al.*, *J. Proteome Res.* **11**, 1009 (2012).
11. V. C. Evans *et al.*, *Nat. Methods.* **9**, 1207 (2012).
12. R. Aebersold *et al.*, *Nat. Chem. Biol.* **14**, 206 (2018).
13. Z.-F. Yuan, A. M. Arnaudo, B. A. Garcia, *Annu. Rev. Analyt. Chem.* **7**, 113 (2014).
14. D. Phanstiel *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4093 (2008).

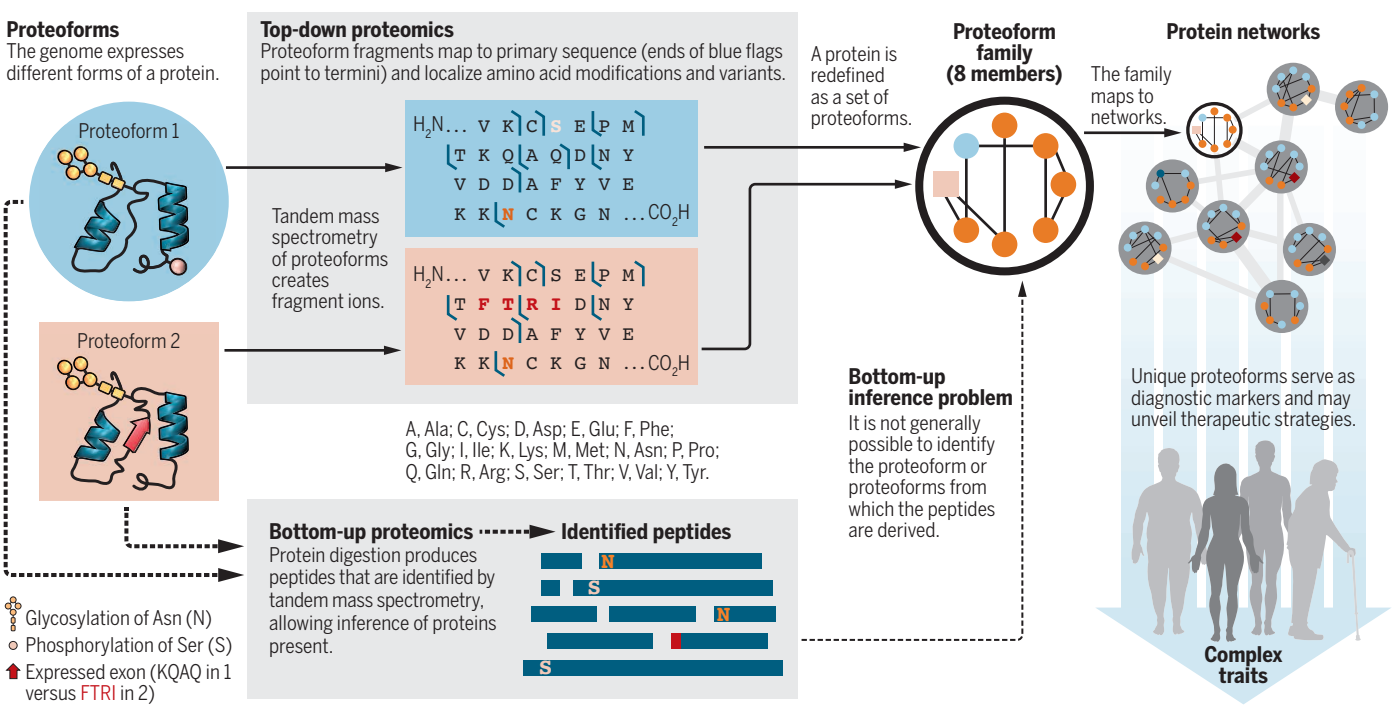
ACKNOWLEDGMENTS

We thank J. Loo, J. Chamot-Rooke, L. Pasa-Tolic, Y. Ge, and Y. Tsybin for their comments and suggestions and D. Walt for pointing out the potential effects of errors in transcription and translation. The Proteoform Atlas is supported by a grant from the Paul G. Allen Family Foundation (http://repository.topdown-proteomics.org; Award 11715). We thank the National Institute of General Medical Sciences for their support under grants 1R01GM114292 (L.M.S.) and P41 GM108569 (N.L.K.). The authors are members of the Consortium for Top Down Proteomics.

10.1126/science.aat1884

Identifying proteoforms within their families and protein networks

Proteoforms underlie complex traits and molecular mechanisms in biology. Top-down (whole protein) and bottom-up (peptide) proteomics methods are compared.



GRAPHIC: ADAPTED BY A. KITTERMAN/*SCIENCE* FROM M. MULLOWNEY AND S. PATRIE/NORTHWESTERN UNIVERSITY

Proteoforms as the next proteomics currency

Lloyd M. Smith and Neil L. Kelleher

Science **359** (6380), 1106-1107.
DOI: 10.1126/science.aat1884